Advances in Theoretical Economics

Volume 1, *Issue* 1 2001 *Article* 3

Reinforcement Learning in Repeated Interaction Games

Jonathan Bendor
Stanford University

Dilip Mookherjee
Boston University

Debraj Ray
New York University

Advances in Theoretical Economics is one of *The B.E. Journals in Theoretical Economics*, produced by bepress.com.

Reinforcement Learning in Repeated Interaction Games

Abstract

We study long run implications of reinforcement learning when two players repeatedly interact with one another over multiple rounds to play a finite action game. Within each round, the players play the game many successive times with a fixed set of aspirations used to evaluate payoff experiences as successes or failures. The probability weight on successful actions is increased, while failures result in players trying alternative actions in subsequent rounds. The learning rule is supplemented by small amounts of inertia and random perturbations to the states of players. Aspirations are adjusted across successive rounds on the basis of the discrepancy between the average payoff and aspirations in the most recently concluded round. We define and characterize pure steady states of this model, and establish convergence to these under appropriate conditions. Pure steady states are shown to be individually rational, and are either Pareto-efficient or a protected Nash equilibrium of the stage game. Conversely, any Pareto-efficient and strictly individually rational action pair, or any strict protected Nash equilibrium, constitutes a pure steady state, to which the process converges from non-negligible sets of initial aspirations. Applications to games of coordination, cooperation, oligopoly, and electoral competition are discussed.

1 Introduction

In complex environments, expected payoff maximization does not often seem plausible as a *description* of how players actually make decisions. This notion supposes that every player understands the environment well enough to precisely estimate payoff functions, formulate beliefs concerning the actions of others, and subsequently compute the solution to an optimization problem. Each of these activities requires expensive resources, with respect to gathering and processing of information.[1] Very often, a simple enumeration of the list of all available feasible actions is too demanding. Indeed, the decision of how much resources to devote to information gathering and processing is itself a higher-order decision problem, leading quickly to infinite regress. It is not at all obvious even how to formulate a theory of rational behavior under such circumstances (Lipman (1991)).

These concerns create a space for behavioral models that are cognitively less demanding and more plausible descriptions of real decision-making processes. One approach is to posit a notion of a "satisfactory payoff" for an agent, and then to assume that the agent tends to repeat "satisfactory" actions, and explores alternatives to "unsatisfactory" actions. This view originated in the behavioral psychology literature as stimulus-response models.[2] Similar models have been studied as parables of automata learning in the computer science and electrical engineering literature.[3] Amongst economists, early pioneers of adaptive "satisficing" models include Simon (1955, 1957, 1959), Cross (1973) and Nelson and Winter (1982). More recently, Gilboa and Schmeidler (1995) have developed an axiomatic basis for such an approach, while their theoretical implications have been explored by a number of authors.[4] Experimental support in favor of the reinforcement learning hypothesis *vis-a-vis* the traditional rational play hypothesis and belief learning has been extensively discussed in more recent literature on experimental games.[5]

However, the implications of reinforcement learning *in a strategic context* have not received much attention, except for specific classes of games and special families of learning

[1] For instance, Cournot duopolists need to know the demand function for their product, which requires them to devote significant expenditures to marketing research. They need to combine this with knowledge of their own cost functions, and of beliefs concerning the output of their competitor, then solve for a profit-maximizing output (presumably by using suitable algorithms to solve corresponding programming problems). Formulating this decision problem as a Bayesian game of incomplete information further increases the resources required to formulate and solve the resulting optimization problems.

[2] See Estes (1954), Bush, Mosteller and Thompson (1954), Bush and Mosteller (1955), Luce (1959, Chapter 4) and Suppes and Atkinson (1960).

[3] See Lakshmivarahann (1981), Narendra and Mars (1983), Narendra and Thathachar (1989), and Papavassilapoulos (1989).

[4] See Arthur (1993), Bendor, Mookherjee and Ray (1992,1995), Börgers and Sarin (1997, 2000), Dixon (2000), Gilboa and Schmeidler (1995), Karandikar, Mookherjee, Ray and Vega-Redondo (1998), Kim (1995a), Pazgal (1997) and Palomino and Vega-Redondo (1999).

[5] See Selten and Stoecker (1986), Selten (1991), Mookherjee and Sopher (1994, 1997), Roth and Erev (1995), Kim (1995b), Erev and Roth (1995, 1998) and Camerer and Ho (1999).

rules.[6] The purpose of this paper is to provide a general theory of reinforcement learning when two players repeatedly interact with one another to play an arbitrary finite action game, using a minimal set of assumptions about the nature of reinforcement process. This helps identify some key properties of reinforcement learning models that are both quite general and distinctive, relative to alternative models of learning and evolution in games.

We incorporate only two fundamental assumptions concerning the nature of reinforcements, which are common to most formulations in existing literature, and receive considerable support in the psychological literature.[7] The first is *positive reinforcement* (PR), which says that an action generating a satisfactory payoff experience tends to be selected with a (weakly) higher probability in the next play. The second assumption is *negative reinforcement* (NR), which states that after an unsatisfactory payoff experience, players will attempt all other actions with positive probability. A few additional but mild restrictions are imposed: the reinforcement rules are modified by small degrees of *inertia*, wherein — absent any other information — players increase probability weight on the most recently selected action, and by random *perturbations*, in which players might develop slightly different behavioral propensities with small probability. This last notion is akin to — but not entirely congruent with — the idea of experimentation with different actions. Its role is to prevent players from getting locked into low payoff actions owing to historically low aspirations and lack of experimentation with alternative actions.

The notion of "satisfactory" inherently requires a player to be endowed with some *aspiration* level that is used as a reference point or threshold. How are such aspirations formed? It is plausible that while aspirations shape behavior in the short to intermediate run, they themselves adapt in the long-run to past payoff experiences. In this paper we study a two-way sequenced dynamic between aspirations and behavior. Specifically, we examine a long-lived relationship between two players, the duration of which is divided into what we call *rounds*. Within any given round the two players play the game successively a large number of times with fixed (though possibly player-specific) aspirations. *Across* rounds, aspirations are adjusted on the basis of the discrepancy between average payoffs and aspirations in the previous round. This formulation involves first evaluating the limiting average outcome within any given round, and then taking limits across rounds in order to identify long-run aspirations and induced behavior. The main advantage of this formulation is that it limits the dimensionality of the state space at each stage of the dynamic, thereby allowing us to provide a general theory applicable to arbitrary finite action games. In contrast, a model in which aspirations and players' behavioral propensities simultaneously evolve, as in the model of Karandikar *et al* (1998),

[6] Related literature is discussed more thoroughly in Section 9 and in Bendor, Mookherjee and Ray (2000).

[7] See Erev and Roth (1998) for relevant citations to this literature.

involves a higher dimensional dynamic which can be analysed only for a special class of 2 by 2 games.

While we do not neglect the dynamics, our focus is on the steady states of this process. We devote particular attention to *pure* steady states, in which agents have deterministic aspirations and select particular actions with probability one. We justify this focus by providing some results concerning convergence to such states. Specifically, we demonstrate global convergence (to pure steady states) in symmetric games where players have symmetric aspirations, and report some partial convergence results in the more general case (e.g., if initial aspirations of both players lie in suitable intermediate ranges).

Most of the paper is devoted thereafter to characterizing such pure steady states. It is shown that they correspond to an intuitive notion of stability of distributions over the (behavior) states of players. This notion is called a *pure stable outcome (pso)*, which requires *consistency* with the underlying aspirations (i.e., the payoffs must equal the aspirations), and a certain form of *stability* with respect to random perturbations of the states of players. The main convenience of this stability criterion is that it can be checked given *only* the payoff matrix of the game, i.e., without an explicit analysis of the entire underlying dynamic.

We then establish a number of key properties of pso's. They are *individually rational*, in the sense that players attain at least (pure strategy) maxmin payoffs. Moreover, they are *either* Pareto-efficient *or* a "protected" Nash equilibrium. The latter is a Nash equilibrium with the additional "saddle point" property that unilateral deviations cannot hurt the opponent, nor generate a Pareto improvement. An example of this is mutual defection in the Prisoners' Dilemma, or a pure strategy equilibrium of a constant sum game.

A converse to this result can also be established: *any* Pareto-efficient and strictly individually rational action pair is a pso, and so is any protected strict Nash equilibrium. In particular the former result indicates that *convergence to non-Nash outcomes is possible under reinforcement learning in repeated interaction settings.* For instance, cooperation is possible in the Prisoners' Dilemma.

One interpretation of this result is that in the one-player problem induced by the strategy of the other player, convergence to strongly dominated actions can occur. In contrast, in a "genuine" single-person decision making environment with deterministic payoffs, our assumptions on the learning process guarantee convergence to the optimal decision.[8] Therefore convergence to dominated actions is due to the interaction between the learning dynamics of the two players, rather than their inability to solve simple single person decision problems. In particular, an action that is strongly dominated in

[8]This may not be true of one-person decision problems with random payoffs. See Section 10 for further discussion of this point.

a single person setting, is no longer so in a game setting owing to the feedback effects induced by the learning process of other players. These results are distinctive to models of reinforcement learning in repeated interaction settings, in contrast to models of "rational" learning, "best response" learning, or evolutionary games.[9]

Other results include a generic existence theorem for pso's, and some properties of mixed stable distributions. These general characterization results yield sharp predictions for a number of games of coordination, cooperation and competition.

The paper is organized as follows. Section 2 describes the basic model of how players adjust their states within a given round, with given aspirations. Section 3 discusses the dynamics of aspirations across successive rounds. Section 4 introduces steady states and the notion of stable distributions, while Section 5 provides results concerning convergence to pure stable outcomes. Section 6 provides a characterization of such outcomes, while Section 7 contains additional remarks on pure and mixed stable states. Section 8 applies our results to specific games. Section 9 discusses related literature, and Section 10 suggests possible extensions of our model. Finally, Section 11 concludes. Technical proofs are relegated to the Appendix.

2 Reinforcement Behavior with Given Aspirations

Two players named A and B possess finite action sets $\mathcal{A}$ and $\mathcal{B}$, with (pure) actions $a \in \mathcal{A}$, $b \in \mathcal{B}$. Let $\mathcal{C} \equiv \mathcal{A} \times \mathcal{B}$; a *pure action pair* is then $c = (a, b) \in \mathcal{C}$. Player A has a *payoff function* $f : \mathcal{C} \to I\!R$ and B has a payoff function $g : \mathcal{C} \to I\!R$. We shall refer to the vector-valued function $h \equiv (f, g)$ as the payoff function of the game.

Players inherit aspirations $(F, G) \in I\!R^2$ in any given round. Within each round they play a large number of times $t = 1, 2, \ldots$ successively, with fixed aspirations. Let H denote the pair (F, G). For the rest of this section, we study the dynamics of play within a given round, with fixed aspirations H. The next Section will then turn to the aspiration dynamics across rounds.

The *state* of a player at any given play of the game is represented by a probability vector, whose components are probability weights assigned to different actions. This represents the player's psychological inclination to select amongst them, based on past experience.[10] The *state of the game* at the beginning of any play is represented by

[9] Our analysis also shows that similar results obtained in specific settings and with particular forms of learning rules in repeated interaction settings (e.g., Bendor, Mookherjee and Ray (1992, 1995), Kim (1995a), Pazgal (1997), Karandikar et al (1998) and Dixon (2000)) actually do generalize substantially.

[10] The theory will also apply to alternative formulations of the state variable, e.g., in terms of a vector of 'scores' assigned to different actions that summarize the degree of success achieved by them in the past, which determine choice probabilities (as in the model of Luce (1959)). This version of the model is described further below. For ease of exposition we adopt the formulation of choice probabilities as the state variable throughout this paper.

$\gamma \equiv (\alpha, \beta)$ in the set $\Gamma \equiv \Delta(\mathcal{A}) \times \Delta(\mathcal{B})$. Γ, endowed with its Borel sets, will represent the state space for the given pair of players within a given round.

We now describe how a player's state is updated from one play to the next. It is defined primarily by a process of *reinforcement*, modified slightly by *inertia*, and perturbed occasionally by random *trembles*.

2.1 Reinforcement

In the definitions that follow, we describe the reinforcement process for player A, with the understanding that B modifies her state via an analogous process. For player A, let α be his current psychological state, a the action currently chosen (according to the probability $\alpha(a)$), f the payoff currently received, and F the fixed aspiration level. A *reinforcement rule* R^A maps this into a new state $\tilde{\alpha}$ at the next play. We assume that R^A is continuous, maps totally mixed states to totally mixed states, and satisfies the following two restrictions:

Positive Reinforcement (PR) If $f \geq F$, then $\tilde{\alpha}(a) \geq \alpha(a)$.

Negative Reinforcement (NR) If $f < F$, then $\tilde{\alpha}(a') > 0$ for all $a' \neq a$.

These restrictions are weak, requiring that satisfactory payoff experiences do not cause probability weights to decline (PR), while unsatisfactory experiences cause other actions to be tried (NR). A more symmetric formulation might be that a failure results in the player reduces the probability weight on the chosen action, and simultaneously increases the weight on alternative actions.[11] Our NR formulation is clearly weaker than such a condition. Indeed, it has bite only for current states which are not totally mixed, and serves to rule out the possibility that a player converges to a pure action despite being perpetually disappointed with it.

Examples of rules satisfying these conditions include the Bush-Mosteller learning model (e.g., Bush, Mosteller and Thompson (1954), Bush and Mosteller (1955)) concerning stimulus response of subjects in experiments where outcomes are classified into *success* and *failure*, i.e., the payoff function f is dichotomous and maps into $\{0, 1\}$. The notion of an aspiration level is then implicit in the definition of payoff experiences as satisfactory ($f = 1$) or unsatisfactory ($f = 0$). In such contexts the model prescribes linear adjustment of probability weights following any given choice of actions:

$$\tilde{\alpha} = [1 - \phi]\alpha + \phi\lambda \tag{1}$$

where $\phi = \phi(a, f)$ is an adjustment parameter lying between 0 and 1, and $\lambda = \lambda(a, f)$ is a probability vector in $\Delta(\mathcal{A})$. In particular, if $f = 1$ then $\lambda(a, f)$ could be the vector

[11] This presumes that there is no 'similarity' relation among actions, which might cause the weight on actions similar to a recently unsuccessful choice to also be reduced.

putting weight 1 on action a: then action a is positively reinforced, and the player responds by moving linearly in the direction of the pure strategy δ_a concentrated on a. On the other hand if a failure is realized ($f = 0$), then $\lambda(a, f)$ could be a vector which puts zero weight on action a and positive weight on all the other actions, so the player reduces the weight on action a and increases it on all the other actions. If θ lies strictly between 0 and 1, then (PR) and (NR) are satisfied. In the case where the payoff function is not dichotomous, the Bush-Mosteller rule can be generalized as follows: if $F \in I\!R$ represents the aspiration of the player,

$$\tilde{\alpha} = [1 - \phi]\alpha + \phi\lambda \tag{2}$$

where $\phi = \phi(a, f, F)$ and $\lambda = \lambda(a, f, F)$. In particular $\lambda(a, f, F) = \delta_a$ if the player is satisfied ($f \geq F$), and assigns positive probability to any $a' \neq a$ otherwise. A particular case of this is studied by Börgers and Sarin (2000) in a context with a single player and two actions.[12]

Our approach can also be extended to the formulation of Luce (1959) based on assignment of scores to different actions, where choice probabilities depend on relative scores, and scores are updated adaptively based on experience. A version of this approach has been tested experimentally by Erev and Roth (1995, 1998). The extension would require scores to be used as the state variable, rather than the choice probabilities.[13]

The reinforcement learning rules described above satisfy a *diminishing distance* (DD) property of the induced Markov process, studied extensively by Norman (1972). Informally, this property is an extension of the contraction mapping notion to the stochastic case. For simplicity of exposition, we apply a restricted version of this property to our reinforcement rule; all the results we need can be obtained from the weaker specification as well.[14]

Definition R^A satisfies the *strong diminishing distance (SDD)* property if there exists $r \in (0, 1)$ such that for any current experience (a, f, F) and any two states α, α' for player A that map respectively into $\tilde{\alpha} \equiv R^A(\alpha, a, f, F)$ and $\tilde{\alpha}' \equiv R^A(\alpha', a, f, F)$ at the following play,

$$||\tilde{\alpha} - \tilde{\alpha}'|| \leq r||\alpha - \alpha'||. \tag{3}$$

Norman (1972, Chapters 1-3) establishes that the DD property (and *a fortiori*, the strong version given here) implies that the properties of the induced Markov process over a (compact) state space are entirely analagous to those of finite Markov chains. The SDD property will be useful at some parts of the analysis below. But as will become

[12] Their paper also allows aspirations to evolve simultaneously with choice probabilities.

[13] Details of such an extension are available at http://www.econ.nyu.edu/user/debraj/Papers/bmrLR.pdf.

[14] The DD property is weaker by requiring that the *expected* distance in a finite number of steps (not necessarily one step) is contracting.

evident, it is less fundamental for our purposes than the reinforcement properties (PR) and (NR) which we assume throughout the rest of the paper.

2.2 Inertia

We combine reinforcement with some inertia. Recall that δ_a denotes the probability vector concentrated on action a. We shall refer to this as the *pure strategy state (pss) a* for player A. We assume that the new state of player A puts some positive weight on the most recently selected action, with the remainder selected according to the reinforcement rule R^A:

INERTIA (I) There is $\epsilon \in (0,1)$ such that player A's psychological state next period, α', can be written as

$$\alpha' = L^A(\alpha, a, f, F) \equiv \epsilon\delta_a + (1-\epsilon)R^A(\alpha, a, f, F) \tag{4}$$

Property (I) is similar to analogous assumptions in Binmore-Samuelson (1997) and Karandikar *et al* (1998), and can be motivated from more primitive considerations such as switching costs. It implies that a positive reinforcement will be followed by an increase in the probability weight on the most recently selected action at some minimal geometric rate (unless it was already at one already). In the absence of this assumption, such a property can be directly imposed by strengthening (PR) in a manner which also happens to be satisfied by many common learning rules. The inertia assumption simplifies the analysis to some extent, so we shall also impose it throughout the rest of our analysis.

2.3 Trembles

Finally, we suppose that at each play, it is possible (with small probability η) that a player may simply gravitate to a new psychological state — typically within some neighborhood of the old state — instead of following the reinforcement rule. This ensures that players will occasionally experiment with different actions, preventing them from getting stuck with actions generating lower payoffs than others, with correspondingly low aspirations that cause them to be satisfied with such low payoffs.

To formalize this, suppose that for small values of η, the new state is generated according to a density $e_A(.|\alpha)$ whose support includes an open neighborhood $N(\alpha)$ of α, and assume that the density e_A is continuous in α. With remaining ("large") probability $1-\eta$, the state of player A is updated according to the rule L^A discussed above.

For simplicity we assume that the probability η is the same for both players, and that the perturbations are independent across players and successive plays. For $i = A, B$, denote by E^i the rule that results when L^i (already described above) is combined with these perturbations.

2.4 Limiting Outcomes Within a Round: No Perturbations

The given aspirations (F, G) and the updating rules for the two players generate a Markov process on the state space Γ. When the tremble probability $\eta = 0$ we shall refer to the corresponding transition kernel $\mathcal{P}$ as the *unperturbed process*. Specifically, for given current state $\gamma \in \Gamma$, $\mathcal{P}(\gamma, S)$ is the probability assigned to Borel set S at the next round, where for any given S, $\mathcal{P}(., S)$ is a measurable function. From any initial state γ, this induces a sequence of probability measures at successive plays $\mu_t = \mathcal{P}^t(\gamma, .)$, where $\mathcal{P}^t$ denotes the t-step transition kernel corresponding to $\mathcal{P}$.

A measure μ over the state space Γ is said to be *invariant* for $\mathcal{P}$ if $\mu.\mathcal{P} = \mu$.[15] An invariant measure μ for $\mathcal{P}$ is said to be a *long-run distribution* if from *any* $\gamma \in \text{supp } \mu$, the sequence of probability measures $\mathcal{P}^t(\gamma, .)$ converges weakly to μ. A long-run distribution thus has the property that all the states in its support 'communicate' with one another. Its support is stochastically closed (i.e., $\gamma \in \text{supp } \mu$ implies that with probability one it the state will stay in supp μ forever thereafter) and does not contain any proper subset which is stochastically closed.

The *empirical distribution* ν_n over the first n plays $t = 1, 2, \ldots, n$ is defined as the measure over the state space Γ by the proportion of visits

$$\nu_n(S) = \frac{1}{n} \sum_{t=1}^{n} I_S(\gamma_t)$$

to any Borel set S, where I_S denotes the indicator function of the set S.

Definition The unperturbed process $\mathcal{P}$ is said to be *weakly ergodic* if (i) it has a finite number of *long run distributions*; and (ii) with probability one the empirical distribution ν_n converges weakly to some long-run distribution as $n \to \infty$.

If the unperturbed process is weakly ergodic, the asymptotic empirical distribution over the state if any given round is well defined and given by one of its long-run distributions. The following result can be obtained following a straightforward application of Theorems 4.3 and 4.4 in Norman (1972, Chapter 3).[16]

[15] For a transition kernel $\mathcal{P}$ on Γ and a measure μ on Γ, we define the measure $\mu.\mathcal{P}$ by $\mu\mathcal{P}(S) = \int_\Gamma \mathcal{P}(\gamma, S)\mu(d\gamma)$ for any Borel set S.

[16] Indeed, this result follows only from the SDD property and does not require either PR or NR. The proof involves defining the event space by whether or not a given player experiences inertia, and the action pair actually chosen. Then if the reinforcement rules of each player satisfy the SDD property, the Markov process over the state space can be verified to have the DD property and is hence compact. Since the state space is compact, Theorems 4.3 and 4.4 in Norman (1972) can be applied to yield the result. In the case of the Luce-Erev-Roth model, the SDD property is satisfied when the state variable is taken to be the vector of scores. Hence the score dynamic is weakly ergodic, in turn implying weak ergodicity of the induced choice probabilities.

PROPOSITION **1** *Suppose that the reinforcement rules R^A and R^B satisfy the SDD property. Then for any pair of aspirations, the unperturbed process is weakly ergodic.*

Actually, the structure of the model can be exploited to a considerable extent to yield sharper results, provided we place some conditions on the set of aspirations. To this end, say that an action pair c is *mutually satisfactory* (MS) relative to aspiration pair $H = (F, G)$ if $\{f(c), g(c)\} \geq H$. Now define an aspiration pair H to be *intermediate* if it is (strictly) individually rational:

$$H \gg \underline{H}, \tag{5}$$

where $\underline{H}$ denotes the pair of (pure strategy) maxmin payoffs, and if there is some action pair that is MS relative to H. Also, say that an aspiration pair H is *low* if

$$H \ll \underline{H}. \tag{6}$$

PROPOSITION **2** *Let H be an aspiration pair that is either intermediate or low, and suppose that (PR) and (NR) are satisfied. Then from any initial state, the unperturbed process $\mathcal{P}$ converges almost surely to some pure strategy state c which is MS relative to H. Each such MS pair exhibits a positive probability of being reached in this way if the initial state is totally mixed.*

In particular, $\mathcal{P}$ is weakly ergodic, and the set of corresponding long-run distributions of $\mathcal{P}$ is the set of degenerate distributions δ_c concentrated on pure strategy states c that are MS relative to H.

This proposition is adopted from our earlier work (Bendor, Mookherjee and Ray (1992, 1995)); see also Proposition 2 and Remark 3 in Börgers and Sarin (1997). The outline of the underlying argument is the following. First, it can be shown that the reinforcement and inertia assumptions imply that starting from an arbitrary initial state, an MS action pair will be played within the next two plays with probability bounded away from zero. For if an MS action pair is not played immediately, some player must be dissatisfied and subsequently must try other actions, causing an MS action pair to possibly be played within the next two plays. And once an MS action pair is played, both players will be positively reinforced. Combined with inertia, the probability weight on these actions will increase at least at a geometric rate. This ensures that an infinite run on this action pair has a probability bounded away from zero, and so must eventually happen almost surely. Such an infinite run would cause the probability weights on these actions to converge to one.

It should be noted that a diminishing distance property is *not* needed to obtain Proposition 2. Indeed, Propositions 1 and 2 may be viewed as embodying different approaches to establishing weak ergodicity. One uses stronger assumptions to yield

weak ergodicity for *any* pair of aspirations, while the other exploits the positioning of aspirations to obtain more structured results. In any case, the weak ergodicity of the unperturbed process will be necessary in our discussion of the aspiration dynamic in the next Section. So for the rest of the paper we shall assume that (SDD), (PR) and (NR) are simultaneously satisfied.[17]

Note that the unperturbed process may have multiple long-run distributions. For example, in the context of Proposition 2, if there are two MS action pairs, then the corresponding degenerate distributions concentrated on either pair constitute long-run distributions. The unperturbed learning process *cannot* be ergodic in that case: the long run empirical distribution (while well-defined) will depend on the initial state and the actual historical patterns of play. In other words it is inherently unpredictable and accordingly must be treated as a random variable.

2.5 Limiting Outcomes Within a Round: Small Trembles

The multiplicity of long-run distributions provokes the following concern: while a particular distribution may receive significant probability under the unperturbed process (depending on initial states), the "overall" probability of reaching such distributions may be low if states are not *robust*, i.e., immune to trembles.

As an instance of this phenomenon, consider the following example of a Prisoners' Dilemma:

	C	D
C	(2,2)	(0,3)
D	(3,0)	(1,1)

Suppose that aspirations are at $(0.5, 0.5)$. Proposition 2 implies that the unperturbed process has exactly two limits, respectively concentrated on the pure action pairs (C, C) and (D, D) that are mutually satisfactory relative to the given aspirations. Moreover, each limit has a positive probability of being reached. Now consider what happens if we introduce trembles. This permits "transitions" to occur (with low probability depending on the infrequency of the tremble) between the two limits. However, there is an asymmetry in these transitions. Consider an individual tremble from (C, C): the "trembler" benefits by shifting weight to D. Because the "non-trembler" loses (relative to her aspirations), she shifts weight to D as well. Then (D, D) is played with positive probability, from which the untrembled process can converge to the other pure limit

[17] If initial aspirations are intermediate, then the SDD assumption is actually unnecessary; the latter assumption is required only to ensure that the aspiration dynamic is globally well-defined.

(D, D). Hence a single tremble suffices to move from the the pure (C, C) limit to the pure (D, D) limit.

But the reverse is *not* true. Starting from the (D, D) limit, a single tremble causes the "trembler" to try C, which causes her payoff to fall below aspirations. Hence the "trembler" tends to revert back to D. In the meantime, the deviation benefits the "non-trembler", who thus continues to stick to the pure strategy concentrated on D. Thus the pure (D, D) limit exhibits a stability to one-person trembles that the pure (C, C) limit does not. If trembles are infrequent and independent across players, two-person trembles will occur with vanishingly small probability (relative to one-person trembles) and so may be ignored. Then with probability close to one, the process will spend almost all the time in the long-run near the (D, D) limit, despite the fact that the (C, C) limit can be reached with positive probability in the absence of trembles. In such cases it is appropriate to ignore the pure (C, C) limit owing to its lack of robustness to small trembles.

We now present the formal argument. First note that a positive tremble probability ensures that the resulting process must be ergodic:

PROPOSITION **3** *Fix $\eta > 0$ and some pair of aspirations. Then the perturbed Markov process $\mathcal{P}_\eta$ is strongly ergodic, i.e., there exists a measure μ_η such that the sequence of probability measures $\mathcal{P}_\eta^t(\gamma, .)$ from any initial state γ converges strongly to μ_η.*

The next step is to take the tremble probability to zero, and focus on the "limit" of the corresponding sequence of ergodic distributions μ_η. The word "limit" is in quotes because there is no guarantee of convergence.[18] So we employ an analogue of trembling-hand perfection: admit possibly multivalued predictions of long-run outcomes for any given round; indeed, all those which are robust to *some* sequence of vanishing tremble probabilities. This is explained in more detail below.

Given any initial state $\gamma \in \Gamma$, define the long-run average transition $\mathcal{R}(\gamma, .) \equiv \lim_n \frac{1}{n} \sum_{t=1}^n \mathcal{P}^t(\gamma, .)$, if this limit measure is well-defined (in the weak convergence topology). Next, define $\mathcal{Q}(\gamma, .)$ to be the one-step transition when exactly one player $i = A, B$ is chosen randomly to experience a tremble, while the other player employs the unperturbed update. In other words, with probability one half it is generated by the composition of the tremble for A and the learning rule L^B, and with probability one half by the reverse combination.

PROPOSITION **4** *(i) Given any sequence of positive tremble probabilities converging to zero, the corresponding sequence of ergodic distributions has a (weakly) convergent sub-*

[18] The limit of the sequence of ergodic distributions may depend on the precise sequence along which the tremble probability goes to zero. These problems are analogous to those arising in the analysis of stability of Nash equilibria.

sequence. (ii) Suppose μ^* is a limit point of such a sequence. Then μ^* is an invariant measure for the transition $\mathcal{Q}.\mathcal{R}$ (as well as for $\mathcal{P}$), provided $\mathcal{R}$ is well-defined.

The first part of this Proposition implies that there always exists a long-run outcome of the untrembled process which is robust with respect to some sequence of vanishing trembles. It is possible, however, that there is more than one long-run outcome with this property. The second part of this proposition (which is based on Theorem 2 in Karandikar *et al* (1998)) describes a property satisfied by each of these 'robust' long run outcomes (provided $\mathcal{R}$ is well-defined): any such distribution is also robust with respect to a *single* perturbation of a single (randomly chosen) player, followed by the indefinite operation of the unperturbed process thereafter. This property will play a key role in the discussion of stability in Section 4 below.

Given the possibility of multiple robust long-run outcomes, there is no basis to select any of these over the rest. Hence we must entertain the possibility that any one of these could arise, depending on initial conditions and the history of play. Specifically, consider *any* invariant measure μ^* which is stable with respect to some sequence of vanishing trembles. Clearly μ^* can be expressed as a convex combination of a finite number of long-run distributions of the untrembled process (since it is itself invariant for the untrembled process). Hence given the set of long-run distributions $\mu_1, \ldots, \mu_K$ of $\mathcal{P}$, there exist weights $\beta_i \geq 0$ such that

$$\mu^* = \sum_{i=1}^{K} \beta_i \mu_i \tag{7}$$

Given aspirations $H = (F, G)$, we can then define

$$\mathcal{D}(H) \equiv \{\mu_i | \beta_i > 0 \text{ for some ergodic limit } \mu^*\};$$

the set of long-run distributions of $\mathcal{P}$ that receive positive weight in some stable limit μ^*. These are the long-run distributions that are robust with respect to some sequuence of vanishing trembles; any one of them could arise in the play of any pair of players with aspirations H. The trembles merely serve to eliminate 'non-robust' long-run outcomes, i.e., which receive zero weight in *every* possible stable invariant measure of the untrembled process. This is entirely analogous to the approach that is now standard in the literature, e.g., Kandori, Mailath and Rob (1993) and Young (1993).

3 Aspiration Dynamics Across Rounds

Thus far we have identified a mechanism that selects a particular class of robust long run distributions, $\mathcal{D}(H)$, beginning from any aspiration vector H. These long run distributions are associated with corresponding average payoffs for each player. This suggests

an updating rule for aspirations across rounds — simply take a weighted average of aspirations H_T and the average payoff vector (Π_T) earned in the most recently concluded round:

$$H_{T+1} = \tau H_T + (1-\tau)\Pi_T \tag{8}$$

where τ is an adjustment parameter lying between 0 and 1. This presumes that players play infinitely many times within any round, a formulation we adopt in the interest of simplicity. It can be thought of as an approximation to the context where players play a large but finite number of times within any round. This is partially justified by the results of Norman (1972) concerning the geometric rate of convergence of diminishing distance Markov processes to the corresponding long run distributions (Meyn and Tweedie (1993) also provide similar results for the perturbed model). The assumption is analogous to corresponding assumptions of infinite number of random matches between successive stages of the finite player model of Kandori, Mailath and Rob (1993).[19]

The next question is: which distribution do we use to compute average payoffs Π_T? One tempting route is to use (one of) the trembled limit(s) μ^* described in the previous section. But this is conceptually problematic when μ^* represents a mixture of more than one robust long run distribution of the unperturbed process. For as we discussed in the previous section, it is more appropriate to view the average outcome within the round as random, located at one of the robust long-run distributions receiving positive weight in μ^* — rather than μ^* itself. Accordingly we must treat Π_T as a random variable, equal to the average payoff in *some* robust distribution. The only restriction thus imposed here is that no weight is placed on a non-robust distribution, i.e., which receives zero weight in μ^*.

Formally, we posit that the distribution over the states of players in a given round-T will be randomly selected from the finite set $\mathcal{D}(H_T)$. Using $\rho(\mu, H_T)$ to denote the probability that the long-run distribution for round T will be $\mu \in \mathcal{D}(H_T)$,

$$\Pi_T = \int h d\mu \quad \text{with probability} \quad \rho(\mu, H_T), \tag{9}$$

where h, it will be recalled, represents the vector of payoff functions. No restriction need be imposed on the family of probability distributions $\rho(.,.)$, except that its support is $\mathcal{D}(H_T)$, i.e., every robust long-run distribution is selected with positive probability.

One might ask: why the restriction to robust distributions? It is possible, of course, that the behavior states of the process within any round may spend time in the neighborhood of a nonrobust long run distribution. However, with sufficiently small trembles the *proportion* of such time will be arbitrarily small relative to the proportion spent at or near robust long run distributions. This is the basis of the restriction imposed here.

[19] We discuss this issue in further detail in Section 10 below.

Together with (8), equation (9) defines a Markov process over aspirations across successive rounds (with state space $\mathbb{R}^2$). In fact, given (8), the state space can be restricted to a compact subset of $\mathbb{R}^2$, formed by the convex hull of the set of pure strategy payoffs, augmented by the initial aspirations. The Markov process for aspirations is well-defined if $\mathcal{P}_{H_T}$ is weakly ergodic for all T with probability one. This will indeed be the case if *either* the reinforcement rules satisfy (SDD), or if initial aspirations are intermediate or low. For future reference, we note that in the case of intermediate aspirations, the state space can be further restricted:

PROPOSITION **5** *Provided initial aspirations are intermediate, (8) and (9) define a Markov process over the set of intermediate aspirations.*

The proof of this result is simple, and follows from Proposition 2.[20]

4 Steady States and Stable Distributions

We now study the steady states of the aspiration-updating process. An analysis of convergence to these steady states is postponed to the next section. For reasons that will soon become apparent, we are particularly interested in *deterministic* steady states of the aspiration dynamic.

Say that H is a *steady state aspiration* if $\int h d\mu = H$ for every $\mu \in \mathcal{D}(H)$. It is a *pure steady state aspiration* if in addition there is a robust distribution $\mu \in \mathcal{D}(H)$ which is concentrated on a pure action pair.

A steady state aspiration H^* corresponds exactly to a deterministic steady state for the process defined by (8) and (9). Irrespective of which distribution in $\mathcal{D}(H^*)$ actually results in a given round, players will achieve an average payoff exactly equal to their aspirations, and so will carry the same aspirations into the next round. Conversely, given (9) it is clear that *every* distribution in $\mathcal{D}(H^*)$ must generate average payoffs exactly equal to aspirations H^* in order for the latter to remain steady with probability one.

While the notion of a steady state aspiration is conceptually clear, it is nevertheless hard to verify this property directly for any given aspiration vector H in a given game, owing to the difficulty in obtaining a general characterization of the entire set $\mathcal{D}(H)$ of

[20] In any round where aspirations are intermediate, the average payoff corresponds to some pure strategy state which Pareto-dominates their aspirations. The aspirations of both players in the next round will then partially move up towards the achieved payoff of the previous round, so they continue to be intermediate. Notice that the same assertion cannot be made of low aspirations, or even the union of intermediate and low aspiration pairs. It is possible that a low starting aspirations pair could lead to an aspirations update that is neither low nor intermediate (using the precise sense in which these terms have been defined).

robust distributions for an arbitrary aspiration pair H. Moreover, one is particularly interested in predicting the behavior of players rather than just the payoffs they achieve. For both these reasons, we now develop an analogous notion of stability of distributions over the state space (of behavior, rather than aspirations), which is easier to verify in the context of any given game.

Specifically, the aim of this Section is to develop an analogous steady state (or stability) notion in the space of distributions over Γ, the set of choice probability vectors. Special attention will thereafter be devoted to a particular class of such steady states, which are concentrated on the play of a pure strategy pair, which we shall call *pure stable outcomes (pso)*. The main result of this section (Proposition 6 below) will be to relate steady states in the aspiration space with those in the behavior space. In particular it will be shown that pso payoffs will correspond exactly to pure steady state aspirations. The following Section will then be devoted to results concerning convergence of behavior to pso's (and analogous convergence of aspirations to pure steady state aspirations), while subsequent Sections will be devoted to characterizing pso's based *only* on knowledge of the payoff functions of the game.

The steady state notion over behavior exploits the characterization of the set of robust distributions provided in part (ii) of Proposition 4. Say that a long-run distribution μ' of the untrembled process can be *reached following a single perturbation* from distribution μ if starting from some state γ in the support of μ, a single perturbation of the state of some player will cause the empirical distribution under $\mathcal{P}(H)$ to converge weakly to the distribution μ' (with positive probability). Next define a set $\mathcal{S}$ of long run distributions to be *SP-closed* (that is, closed under a single perturbation) if for every $\mu \in \mathcal{S}$, the set of long run distributions that can be reached from μ is contained in $\mathcal{S}$, and if every $\mu' \in \mathcal{S}$ can be reached from some $\mu \in \mathcal{S}$.

Proposition 4(ii) implies that the set $\mathcal{D}(H)$ is SP-closed for any pair of aspirations H. This is because it consists of all the long-run distributions that receive positive weight in a measure invariant with respect to the process $\mathcal{Q}.\mathcal{R}$, i.e., where the state of one randomly chosen player is trembled just once, followed by the untrembled process thereafter. Hence if we start from any robust distribution, a single tremble will cause the process either to return to the same distribution, or transit to some other robust distribution. Conversely, any robust distribution can be reached following a single perturbation of some other robust distribution.

To be sure, SP-closure is not "minimal" in the sense that there may be strict subsets of SP-closed sets which are SP-closed. If a set does satisfy this minimality requirement, call it *SP-ergodic*. By weak ergodicity, there are can only be a finite number of long run distributions. Combined with Proposition 4(ii), this implies that $\mathcal{D}(H)$ can be partitioned into disjoint SP-ergodic subsets $\mathcal{S}_1(H), \mathcal{S}_2(H), \ldots, \mathcal{S}_K(H)$. Of course this is provided

that $\mathcal{R}(H)$ is well-defined.[21] If in addition $\mathcal{Q}.\mathcal{R}$ has a unique invariant distribution, then $\mathcal{D}(H)$ is itself SP-ergodic. This motivates the following definition of stability of a distribution over behavior states.

Say that a measure μ over Γ is *stable* if

(i) μ is a long-run distribution of $\mathcal{P}$ for aspirations $H = \int h d\mu$, and

(ii) μ belongs to a set $\mathcal{S}$ of long-run distributions of $\mathcal{P}$ which is SP-ergodic, for which every $\mu' \in \mathcal{S}$ satisfies $\int h d\mu' = H$.

Notice that this definition makes no reference at all to set $\mathcal{D}(H)$ of robust long run distributions. Part (i) says that μ is a long-run distribution of $\mathcal{P}$ which is *consistent* with the aspirations H, i.e., generates an average payoff equal to H. To be sure, this consistency property is required by the condition that H is a steady state aspiration. The steady state property additionally demands that *every* long-run distribution in $\mathcal{D}(H)$ is consistent. Instead, (ii) imposes the milder condition that every distribution in the same SP-ergodic subset as μ is consistent. On the other hand, (ii) is stronger than what the steady state property for aspirations by itself requires, by insisting that μ belong to an SP-ergodic set of long-run distributions of $\mathcal{P}$. This condition is always met when $\mathcal{R}$ is well-defined (in that case this property is true for every element of $\mathcal{D}(H)$ by virtue of Proposition 4(ii), so ceases to have any bite).

The justification for this definition of stability of a distribution over behavior states, then, is not that it produces an exact correspondence with the notion of a steady state aspiration. Rather, conditions (i) and (ii) are easier to check for any candidate distribution in any given game. The main convenience of this definition is that it avoids any reference to robust distributions, i.e., the set $\mathcal{D}(H)$ of distributions 'selected' by the process of vanishing trembles. Specifically, checking for stability of a distribution μ over Γ requires that we go through the following steps:

(1) First calculate the average payoff H for each player under μ, and then check the consistency property: is μ a long-run distribution of the untrembled process in any round where players have aspirations H?

(2) Next find the set of all long-run distributions μ' that can be reached from μ (with aspirations fixed at H) following a single random perturbation.

(3) Then check that every such μ' generates a payoff vector of H.

[21] In general, $\mathcal{D}(H)$ can be partitioned into a collection of a collection of nonempty SP-ergodic sets, and a 'transient' set containing distributions which cannot be reached from any distribution in a SP-ergodic set.

(4) Finally, ensure that starting from any such μ', it is possible to return to μ following a sequence of single random perturbations (i.e., there is a sequence of long-run distributions $\mu^1, \mu^2, \ldots, \mu^N$ with $\mu^1 = \mu'$ and $\mu^N = \mu$, such that μ^k can be reached from μ^{k-1} following a single random perturbation).

This procedure avoids the need to find the entire set of selected distributions $\mathcal{D}(H)$, which is typically difficult.

A particular case of a stable distribution is one which is entirely concentrated on some pure strategy state, which corresponds to the notion of a pure steady state aspiration. Thus say that a pure action pair $c \in \mathcal{A} \times \mathcal{B}$ is a *pure stable outcome* (pso) if the degenerate measure $\mu = \delta_c$ concentrated on the pure strategy state c is stable.

A pso combined with a pure steady state aspiration is nonrandom in all relevant senses: behavior and payoffs for both players are deterministic. Our results in the subsequent section will justify our interest in such outcomes. But before we proceed further, it is useful to clarify the exact correspondence between our notion of steady state aspiration (which pertains to steady state *payoffs*) and that of stable distributions over the state space (which pertains to the steady state *behavior*). These results follow up on and extend the informal discussion above.

PROPOSITION **6** *(a) If $c \in \mathcal{A} \times \mathcal{B}$ is a pso then $h(c)$ is a pure steady state aspiration. Conversely, if $h(c)$ is a pure steady state aspiration, then some $c' \in \mathcal{A} \times \mathcal{B}$ with $h(c') = h(c)$ is a pso.*

(b) More generally, if H is a steady state aspiration, then there exists $\mu \in \mathcal{D}(H)$ which is stable.

(c) If μ is a stable distribution with aspirations $H = \int h d\mu$, if $\mathcal{R}$ is well-defined, and $\mathcal{Q}.\mathcal{R}$ has a unique invariant measure, then H is a steady state aspiration.

Part (a) of the proposition states that pure stable outcomes correspond to pure steady state aspirations. Hence in order to study the pure steady states of the aspiration dynamic it suffices to examine the set of pure stable outcomes of the game. The next section will present some convergence results justifying the interest in such pure stable outcomes as representing the long run limit of the process of adaptation of aspirations and behavior.

The remaining parts of Proposition 6 consider the relationship between steady state aspirations and (possibly mixed) stable distributions, and show that the correspondence between the two notions does not extend generally. Parts (b) and (c) assert that there always exists a (pure or mixed) stable distribution corresponding to a steady state aspiration, but the reverse can be assured to be true only under additional conditions. These

are useful insofar as a complete characterization of all (pure or mixed) stable distributions enable us to identify *all* the steady state aspirations (rather than just the pure steady states), as will be the case for certain games considered in Section 8.

5 Convergence to Pure Steady States

A major goal in this paper — especially in the light of part (a) of Proposition 6, — is to characterize those action pairs which are pure stable outcomes. We postpone this task for a while and first settle issues of *convergence* to a steady state. In discussing such issues, we will also come away with further justification for focusing on pure steady states — and therefore on pure stable outcomes.

We begin with a sufficient condition for convergence that bases itself on the location of initial aspirations. It turns out that this condition is also sufficient for deriving convergence to a *pure* steady state.

PROPOSITION **7** *Suppose that initial aspirations H_0 are intermediate. Then the subsequent sequence of aspiration pairs H_T converges almost surely to a pure steady state aspiration as $T \to \infty$. Moreover, for all sufficiently large T, every long-run distribution μ_T over the behavior states in round T is concentrated on some pso.*

Proposition 7 comes with an interesting corollary: *any* pso is almost surely the limit of the process for a suitable (nonnegligible) set of initial aspirations:

PROPOSITION **8** *Take any pso c^*. There exists an open ball $\mathcal{T}(c^*)$ in $\mathbb{R}^2$ such that whenever initial aspirations H_0 lie in $\mathcal{T}(c^*)$, H_T converges to $H^* \equiv h(c^*)$ almost surely, and the long-run distribution μ_T is concentrated on c^* (or some payoff-equivalent pure strategy state) for all large T.*

The detailed proofs are presented in the appendix. But it is useful here to sketch the main idea underlying Proposition 7. By Proposition 2, if H is an intermediate aspiration pair, then the long run distributions corresponding to H are all concentrated on pure action pairs that are MS relative to H. Hence limit average payoffs in the round — no matter how we select from the set of long run distributions — must be (almost surely) no less than H. Because aspirations are bounded above by the maximum of the initial aspirations and the highest feasible payoff in the game, the resulting sequence H_T is a submartingale, and thus converges almost surely. It follows that $H_{T+1} - H_T$ converges to 0 almost surely.

Now $H_{T+1} - H_T = (1 - \tau)[\Pi_T - H_T]$. So it follows that $\Pi_T - H_T$ also converges to 0 almost surely. That is, Π_T must converge as well. Since for any T, Π_T lies in a finite set (by virtue of Proposition 2 once again) it follows that for large T, $\Pi_T = h(c^*)$

for some pure action pair c^*. To complete the proof, it suffices to show that c^* is a pso; the argument for this draws on our characterization of pso's in the next Section and is presented in the appendix.

Proposition 8 is likewise proved in the appendix. For any pso c^* it is shown that for initial aspirations sufficiently close to (but below) $h(c^*)$, convergence to the pure steady state aspiration $h(c^*)$ will occur almost surely.

These propositions establish convergence to pure steady states from initial aspirations that are intermediate. Whether convergence occurs from non-intermediate aspirations remains an open question. We end this section with an observation for the general case.

Consider *symmetric games*, so that $\mathcal{A} = \mathcal{B}$ and $f(a,b) = g(b,a)$ for all (a,b). Make the following two assumptions. First, suppose that there is some symmetric pure action pair $c^* = (a^*, a^*)$ which is Pareto-efficient amongst all mixed strategy pairs. Second, assume that players begin with the same aspirations, and update these by convexifying past aspirations with the *average* payoff received in the current round over *both* players:

$$H_{T+1} = \tau H_T + (1-\tau)\frac{\Pi_{A,T} + \Pi_{B,T}}{2} \tag{10}$$

where $\Pi_{i,T}$ now denotes the average payoff of player $i = A, B$ in round T, and H_T (with some abuse of notation) is now a scalar which stands for the common aspiration of both players.[22] Then the following proposition is true.

PROPOSITION **9** *Consider a symmetric game satisfying the description given above. Then (irrespective of initial aspirations as long as they are the same across players) aspirations almost surely converge to a pure steady state aspiration, and any associated sequence of long-run distributions converges weakly to a pso.*

6 Characterization of Pure Stable Outcomes

The preceding results justify focusing on pure stable outcomes. However, the definition of stability is extremely abstract — referring not just to the consistency of a single long run distribution of the process, but to the *stability* of that distribution, which involves checking all *other* long run distributions that can be reached from it following a single perturbation. The purpose of this section is, therefore, to provide a simple yet near-complete characterization of pure stable outcomes in terms only of the payoff matrix of the game.

Roughly speaking, the characterization states the following:

[22] We continue to assume, of course, that the unperturbed process is always weakly ergodic. This can be directly verified using Proposition 2 if $H_T \leq \pi^*$, but a similar property for $H_T > \pi^*$ would require an assumption such as SDD for the reinforcement rules.

An action pair c is a pso if and only if it is *either* individually rational and Pareto-efficient, *or* a particular type of Nash equilibrium which we shall call "protected".

The "if-and-only-if" assertion is not entirely accurate, but the only reason for the inaccuracy has to do with weak versus strict inequalities, which matters little for generic games.

Now turn to a more formal account. Say that an action pair $c \equiv (a, b) \in \mathcal{A} \times \mathcal{B}$ is *protected* if for all $(a', b') \in \mathcal{A} \times \mathcal{B}$:

$$f(a, b') \geq f(a, b) \quad \text{and} \quad g(a', b) \geq g(a, b). \tag{11}$$

In words, c is protected if unilateral deviations by any player do not hurt the other player.

An action pair $c = (a, b)$ is a *protected Nash equilibrium* if it is protected, it is a Nash equilibrium, and no unilateral deviation by either player can generate a (weak) Pareto improvement. More formally: for all $(a', b') \in \mathcal{A} \times \mathcal{B}$:

$$f(a', b) \leq f(a, b) \leq f(a, b') \tag{12}$$

$$g(a, b') \leq g(a, b) \leq g(a', b). \tag{13}$$

and

$$f(a', b) = f(a, b) \Longrightarrow g(a', b) = g(a, b) \quad \text{and} \quad g(a, b') = g(a, b) \Longrightarrow f(a, b') = f(a, b) \tag{14}$$

A protected Nash equilibrium, then, is a pure strategy Nash equilibrium with the "saddle point" property that unilateral deviations do not hurt the other player (nor generate a Pareto improvement). The corresponding actions (resp. payoffs) are pure strategy maxmin actions (resp. payoffs) for either player. Examples include mutual defection in the Prisoners Dilemma and any pure strategy equilibrium of a zero-sum game.

Next, say that an action pair $c = (a, b)$ is *individually rational* (IR) if $(f(c), g(c)) \geq (\underline{F}, \underline{G})$, where it may be recalled that $\underline{F}$ and $\underline{G}$ denote the (pure strategy) maxmin payoffs for players A and B respectively. It is *strictly* IR if the above inequality holds strictly in both components. Finally, an action pair $c = (a, b)$ is *efficient* if there is no other pure action pair which (weakly) Pareto dominates it.

We now present our characterization results.

PROPOSITION **10** *If $c \in \mathcal{A} \times \mathcal{B}$ is a pso, it must be IR, and is either efficient or a protected Nash equilibrium.*

The converse requires a mild strengthening of the IR and Nash properties.

PROPOSITION **11** *Suppose that one of the following holds: (i) c is a protected Nash equilibrium which is also strict Nash, or (ii) c is efficient and strictly IR. Then c is a pso.*

The argument underlying part (ii) of Proposition 11 is easy to explain in the context of a game with generic payoffs. If c is efficient and strictly IR, then aspirations $h(c)$ are intermediate, and c is the only action pair which is mutually satisfactory relative to these aspirations. Proposition 2 assures us that there is a unique long-run distribution of the untrembled process with aspirations $h(c)$, hence is the only element of $\mathcal{D}(h(c))$. Then c satisfies the requirements of a pso.

Somewhat more interesting is the case of a candidate pso that is *not* efficient. Can such a pso exist? Our answer is in the affirmative, provided the candidate in question possesses the protected Nash property. Intuitively, a protected Nash equilibrium is stable with respect to single random perturbations: since it is protected, a perturbation of one player will not induce the other player to change her state at all. And given that it is a Nash equilibrium, the original deviation cannot benefit the deviator. So if the state of one player changes at all owing to a perturbation, it must involve that player shifting weight to a payoff equivalent action, leaving payoffs unaltered. If the equilibrium is strict Nash, the deviator must return to the original action, ensuring that it is not possible to transit to any other long run distribution following a single tremble. This explains why a strict protected Nash equilibrium constitutes a pso.

On the other hand, inefficient actions that lack the protected Nash property cannot survive as pso's, which is the content of Proposition 10. They *may* serve as attractors (and as pure long run distributions) for certain initial aspirations. But there *must* be other positive-probability attractors: e.g., any Pareto-dominating action pair. Moreover, it is possible to transit to the latter from a non-protected-Nash outcome following a single perturbation. Since the two long run distributions are Pareto-ordered, an inefficient action pair cannot be a pso if it is not protected Nash.

Notice that our characterization is not complete: there is some gap between the necessary and sufficient conditions for a pso. Nevertheless the preceding results cannot be strengthened further. Consider the following examples.

	C	D
C	(2,2)	(1,1)
D	(1,1)	(1,1)

In this example the action pair (D, D) is a protected Nash equilibrium, but it is not a pso (owing to the fact that it is not a strict Nash equilibrium). The reason is that $\delta_{C,D}$ can be reached from $\delta_{D,D}$ following a single perturbation, and $\delta_{C,C}$ can be reached from $\delta_{C,D}$ following a single perturbation. Hence $\delta_{C,C}$ must be included in any stochastically closed subset of $\mathcal{D}(1,1)$ to which $\delta_{D,D}$ belongs. Since the two distributions do not generate the same mean payoffs, (D, D) cannot be a pso.

The next example shows an efficient IR action pair (C, C) which is not a pso. The reason is that following a single perturbation of the row player's state to one which puts positive probability weight on D, he could thereafter converge to the pure action pair D, whereupon the column player must obtain a payoff of 0 rather than 1.

	C	D
C	(0,1)	(0,0)
D	(0,0)	(0,0)

These examples show that Proposition 11 cannot be strengthened. On the other hand, Proposition 10 cannot be strengthened either: the fact of being a pso does not allow us to deduce any strictness properties (such as strict IR or strict Nash).

7 PSO Existence and other Stable Distributions

The very last example of the preceding section actually displays a game in which no pso exists. From any of the pure strategy states where at least one player selects D, it is possible to reach the (C, C) pure strategy state following a sequence of single perturbations. Moreover, we have already seen that it is possible to "escape" (C, C) by means of a unilateral perturbation. This shows that no pso can exist.

Nevertheless, pso's can be shown to exist in generic games (where any distinct action pair generates distinct payoffs for both players):

PROPOSITION **12** *A pso exists in any generic game.*

We end this section with some remarks on stable distributions in general. Observe first that all *degenerate* stable distributions must be pso's.

PROPOSITION **13** *Every distribution which is stable and degenerate (either with respect to behavior states or payoffs) must be a pso.*

The reasoning underlying this is simple. Consider first the possibility that a degenerate distribution places all its weight on some (non-pure-strategy) state in which payoffs randomly vary. Then there must exist some player and a pair of resulting outcomes which yield payoffs that are above and below his aspirations. Given inertia, the former outcome must cause a revision in the state of this player, contradicting the assumption that the distribution is concentrated on a single state. Hence if there is more than action pair that can result from the distribution, they must all generate exactly the same payoff

for each player. Consistency requires this constant payoff equals each player's aspiration. This implies that any action played will be positively reinforced; with positive probability the player will subsequently converge to the corresponding pure strategy, which is an "absorbing" event, contradicting the hypothesis that we started with a long-run distribution (all states in the suppprt of which communicate).

Finally, we describe some properties of non-degenerate stable distributions.

PROPOSITION **14** *Let μ be a stable distribution (with aspirations H). Then:*

(i) μ is individually rational: $H \geq (\underline{F}, \underline{G})$.

(ii) If the payoff to at least one player under μ is stochastic, there is no action pair c^* such that $h(c^*) \geq H$.

Recall from part (b) of Proposition 6 that for every steady state aspiration there exists a corresponding stable distribution. Proposition 14 thus helps restrict the set of steady state aspirations. Spevcifically, part (ii) states that stable distributions cannot generate random payoffs if the corresponding aspirations are Pareto dominated by some pure action pair. And (i) shows that pure strategy maxmin payoffs provide a lower bound. As we shall see in the next section, the combination of these propositions permit sharp predictions for a wide range of games.

8 Applications

8.1 Common Interest, Including Games of Pure Coordination

In a game of *common interest*, there is a pure action pair c^* which strictly Pareto dominates all others. Games of pure coordination constitute a special case: $f(a,b) = g(a,b) = 0$ whenever $a \neq b$, positive whenever $a = b$, and all symmetric action pairs are strictly Pareto-ordered.

Proposition 14 implies that nondegenerate stable distributions with stochastic payoffs to either player cannot exist, as they would be Pareto-dominated by c^*. Hence all stable distributions must be pso's.

Since c^* is efficient and strictly IR, Proposition 11 implies that c^* is a pso. Proposition 10 implies that the only other candidates for a pso must be protected Nash equilibria. The following game is an example of a game of common interest with an inefficient pso (comprised of (M, M)), besides the efficient pso (T, L).

	L	M	D
T	(3,3)	(0,2)	(0,0)
M	(2,0)	(2,2)	(2,0)
B	(0,0)	(0,2)	(0,0)

In the special case of coordination games, however, there cannot be any Nash equilibrium which is protected, as unilateral deviations cause lack of coordination which hurts both players. *In pure coordination games, therefore, there is a unique stable distribution, concentrated entirely on the efficient outcome* c^*.

8.2 The Prisoners' Dilemma and Collective Action Games

Proposition 10 implies that the Prisoners' Dilemma has exactly two pso's: one involving mutual cooperation (since this is efficient and strictly IR), and the other involving mutual defection (since this is a protected strict Nash equilibrium).

More generally, consider the following class of collective action problems: each player selects an effort level from the set $\mathcal{A} = \mathcal{B} = \{e_1, e_2, \ldots, e_n\}$, with $e_i > e_{i-1}$ for all i. These efforts determine the level of collective output or success of the pair. Collective output is increasing in the effort of each player. The payoff of each player equals a share of the collective output, minus an effort cost which is increasing in the personal level of effort.

In such collective action games, an increase in the effort of any player increases the payoff of the other player. The maxmin payoff for each player thus corresponds to the maximum of the player's payoff with respect to his own effort level, assuming the other player is selecting minimal effort e_1. Let e_j denote the best response to e_1.

Now suppose that there exists a symmetric effort pair (e_m, e_m) with $m > 1$ which is efficient and Pareto dominates all other symmetric effort pairs. Then $f(e_m, e_m) > f(e_j, e_j) \geq f(e_j, e_1)$ if $j \neq m$, while $f(e_m, e_m) = f(e_j, e_j) > f(e_j, e_1)$ if $j = m$. So (e_m, e_m) is strictly IR, and thus constitutes a pso.

If there is an inefficient pso, it must be a protected Nash equilibrium. Any Nash equilibrium in which some player is choosing higher effort than e_1 is not protected. Hence the only candidate for an inefficient pso is the pair (e_1, e_1). If this is a strict Nash equilibrium then it is a pso. If it is not a Nash equilibrium then all pso's are efficient. In general, however, intermediate levels of effort are ruled out. Hence *a pso is either efficient, or involves minimal effort* e_1 *by both players.*

8.3 Oligopoly

Consider two firms involved in quantity or price competition, each with a finite number of alternative price or quantity actions to select from. Each firm is free to "exit" (e.g., by choosing a sufficiently high price or zero quantity) and earn zero profit. Suppose that the demand functions satisfy the relatively weak conditions required to ensure that in any pure strategy Nash equilibrium where a firm earns positive profit, there exists a deviation (e.g., involving larger quantity or lower price) for the other firm which drives the first firm into a loss. This implies that each firm's maxmin profit is zero. Hence any

collusive (i.e., efficient from the point of view of the two firms) action pair generating positive profit for both firms is a pso.

If there is any other pso, it must be a zero profit protected Nash equilibrium (e.g. a competitive Bertrand equilibrium in a price-setting game without product differentiation). If such a zero profit equilibrium does not exist (e.g., when there is quantity competition or product differentiation), *all pso's must be collusive.*

8.4 Downsian Electoral Competition

Suppose there are two parties contesting an election, each selecting a policy platform from a policy space $\mathcal{P} \equiv \{p_1, \ldots, p_n\}$, a set of points on the real line. There are a large number of voters, each with single-peaked preferences over the policy space and a unique ideal point. Let f_i denote the fraction of the population with ideal point p_i, and let p_m denote the median voter's ideal point. In the event that both parties select the same position they split the vote equally. The objective of each party is monotone increasing (and continuous) in vote share. Every policy pair is efficient, and the median voter's ideal policy p_m is a maxmin action for both parties. Hence maxmin payoffs correspond to a 50-50 vote split. This game has a unique pso involving the Downsian outcome where both parties select p_m, since this is the only pure action pair which is IR. Indeed, this is the unique stable distribution of the game, as it cannot have a nondegenerate stable distribution owing to Proposition 14.

9 Related Literature

The paper most closely related to this one is our own earlier work on *consistent aspirations* (Bendor, Mookherjee and Ray (1992, 1995)) which did not offer a dynamic model of aspiration adjustment, replacing it instead with the requirement that aspirations be consistent with the long-run behavior they induce (for small trembles). Moreover, the learning rules studied in those papers were significantly narrower, while the results were more restricted.[23]

As already discussed, Erev and Roth (1998) specialize the Luce (1959) model to study aspiration-based learning, though the appropriate state space for their process is one of scores rather than choice probabilities.

Karandikar, Mookherjee, Ray and Vega-Redondo (1998) (henceforth KMRV) consider an explicit model of aspiration adjustment in which aspirations evolve simultaneously with the behavior states of players. This increases the dimensionality of the state space.

[23] No characterization of long run consistent equilibria was provided (except for specific 2 by 2 games of coordination and cooperation under strong restrictions on the learning rules); the only general result establised was that cooperative outcomes form equilibria with consistent aspirations.

The resulting complexity of the dynamic analysis necessitated restriction to a narrow class of 2 by 2 games of coordination and cooperation, and to a particular class of reinforcement learning rules. In particular, that paper assumed that players' states are represented by pure strategies, and that players switch strategies randomly only in the event of dissatisfaction (with a probability that depends on the degree of dissatisfaction). The current paper replaces the simultaneous evolution of aspirations and behavior states with a sequenced two-way dynamic; this permits the analysis to be tractable enough to apply to arbitrary finite games and a very large class of reinforcement learning rules. Nevertheless, our notion of stability in this paper owes considerably to Proposition 4(ii), which in turn is based on Theorem 2 in KMRV.

Börgers and Sarin (2000) also consider simultaneous adjustments of behavior and aspirations. However, they restrict attention to single-person decision problems under risk, rather than games.

Kim (1995a) and Pazgal (1997) apply the Gilboa-Schmeidler case-based theory to repeated interaction games of coordination and cooperation. Kim focuses on a class of 2 by 2 games, whereas Pazgal examines a larger class of games of mutual interest, in which there is an outcome which strictly Pareto dominates all others. Actions are scored on the basis of their cumulative past payoff relative to the current aspiration, and players select only those actions with the highest score. Aspirations evolve in the course of play: aspirations average *maximal* experienced payoffs in past plays (in contrast to the KMRV formulation of aspirations as a geometric average of past payoffs). Both Kim and Pazgal show that cooperation necessarily results in the long run if initial aspiration levels lie in prespecified ranges.

Dixon (2000) and Palomino and Vega-Redondo (1999) consider models where aspirations are formed not just on own payoff experience in the past, but also those of one's peers. Dixon considers a set of identical but geographically separated duopoly markets; in each market a given pair of firms repeatedly interact. The aspirations of any firm evolve in the course of the game, but are based on the profit experiences of firms across *all* the markets. Firms also randomly experiment with different actions. If the current action meets aspirations then experimentation tends to disappear over time; otherwise they are bounded away from zero. In this model, play converges to joint profit maximizing actions in all markets, regardless of initial conditions. Palomino and Vega-Redondo consider a non-repeated-interaction setting (akin to those studied in evolutionary game theory) where pairs are randomly selected from a large population in every period to play the Prisoners' Dilemma. The aspiration of each player is based on the payoff experiences of the entire population. They show that in the long run, a positive fraction of the population will cooperate.

10 Extensions

Our model obtains general insights into the nature of reinforcement learning, but a number of simplifying assumptions were invoked in the process. Dropping these assumptions would constitute useful extensions of the model. In this section, we speculate on the consequences of relaxing some of the most important assumptions.

Our analysis relies heavily on the sequenced dynamic between aspirations and behavior. While it may not be unreasonable to suppose that aspirations adjust more slowly than behavior, the sequenced nature of the process implies that the two differ by an order of magnitude. Simultaneous adaptation — perhaps at a higher rate for behavior — may be a more plausible alternative. This is exactly the approach pursued in KMRV (besides Börgers and Sarin (2000) in a single person environment). However, the simultaneous evolution of behavior and aspirations results in a single dynamic in a higher dimensional state space, which is extremely complicated. KMRV were consequently able to analyze only a particular class of 2×2 games. Moreover, they restricted attention to a very narrow class of reinforcement learning rules, where player's behavior states are represented by pure strategies.

Whether such analyses can be extended more generally remains to be seen. But we can guess at some possible differences by examining the relationship between the results of KMRV and this paper for the Prisoner's Dilemma. In KMRV, there is a unique long run outcome (as aspirations are updated arbitrarily slowly) where both players cooperate most of the time. That outcome is a pso in our model, but there is an *additional* pso concentrated on mutual defection. This is *not* a long run outcome in the KMRV framework. The reason is that starting at mutual defection (and aspirations consistent with this outcome), as one player deviates (accidentally) to cooperation, the other player temporarily experiences a higher payoff. In the KMRV setting, *this serves to temporarily raise the latter's aspiration.* Hence, when the deviating player returns to defection, the non-deviating player is no longer satisfied with the mutual defection payoff. This destabilizes the mutual defection outcome.[24]

This suggests that there *are* differences between models where aspirations adjust in a sequenced rather than simultaneous fashion. Nevertheless there is a close connection between stable outcomes of the two formulations in the Prisoner's Dilemma: the stable outcomes with simultaneously adjusting aspirations is a refinement of the set of stable outcomes with sequentially adjusting aspirations. Whether this relationship extends to more general games is an interesting though challenging question for future research.

At the same time, the sequenced model may well be a better approximation to the

[24]The mutual cooperation outcome (with corresponding aspirations) does not get destabilized in this fashion, and so *is* robust with respect to a single random perturbation, unlike the mutual defection outcome. In contrast, when aspirations are held fixed, neither mutual defection nor mutual cooperation get destabilized by a single random perturbation.

learning process. While the simultaneously evolving aspiration model may appear descriptively more plausible, the finer results of that model are nevertheless driven by the assumption that players respond differently to arbitrarily small disappointments compared with zero disappointment. This is exactly why the mutual defection outcome is destabilized in KMRV. It may be argued that actions plausibly experience negative reinforcement only if the extent of disappointment exceeds a small but minimal threshold. In that case the mutual defection outcome in the Prisoner's Dilemma would not be destabilized with simultaneously (but slowly) adjusting aspirations, and the stable outcomes generated by the two formulations would tend to be similar.[25] If this is true more generally, the sequenced dynamic formulation may be an acceptable "as-if" representation of the outcomes of the more complicated simultaneously-evolving-aspiration model.

The sequenced approach implies, in particular, that there are an infinite number of plays within any round, and then an infinity of rounds; the long run results pertain really to an "ultra" long run. (Note, however, that Propositions 7 and 8 establish convergence to a pso in a finite number of rounds, while behavior distributions within any round converge quickly.) This may limit the practical usefulness of the theory; for instance, with respect to the interpretation of experimental evidence. While we have considerable sympathy with this criticism, it should be pointed out that the time scale is no different from formulations standard in the literature, e.g., Kandori, Mailath and Rob (1993). At any given stage of the game, they assume that players are matched randomly with one another an infinite number of times, thus allowing the theoretical distribution of matches to represent exactly the empirical distribution of matches. This experience is used by players to update their states to the next stage of the game. Our use of an infinite number of plays within a given round serves exactly the same purpose: to represent the empirical average payoff by the average payoff of the corresponding theoretical long run distribution. If players actually play finitely many times within a given round, there will be random discrepancies between the empirical and theoretical average, resulting in additional randomness in the aspiration dynamic. Examining the long-run consequences of these discrepancies would be worthwhile in future research (just as Robson and Vega-Redondo (1996) showed how to extend the Kandori-Mailath-Rob theory analogously). Likewise, the consequences of allowing small aspiration trembles, or reversing the order of trembles and aspiration revisions, would need to be explored.

We simplified the analysis considerably by considering games with deterministic payoffs (though, to be sure, payoffs are allowed to be stochastic if mixed "strategies" are employed). But this is one case in which simplification brings a significant conceptual gain. In a deterministic one-player decision problem our learning process *does* yield long

[25] Of course, the modification of the negative reinforcement assumption would modify the analysis of the sequenced dynamic as well, but for finite generic games the introduction of a small minimal threshold of disappointment for actions to be negatively reinforced would not change the long-run outcomes.

run optimality. Yet in games non-Nash outcomes are possible: i.e., each player is unable to reach the best outcome in the single-person decision problem "projected" by the strategy of the other player. These two assertions mean that the source of non-Nash play is the interaction between the learning of different players. It is not because the learning rule is too primitive to solve deterministic single-person problems.

This neat division breaks down when the single-person problem is itself non-deterministic. As Arthur (1993) has observed, it is more difficult for reinforcement learners to learn to play their optimal actions in the long-run, even in a single person environment, an issue explored more thoroughly by Börgers, Morales and Sarin (1998). It is for this reason that the deterministic case is instructive. Nevertheless, an extension of the model to games with random payoffs would be desirable in future research.

Finally, our model was restricted to the case of only two players, and extensions to the case of multiple players would be worthwhile. It is easily verified that the underlying model of behavioral dynamics *within* any given round (based on Propositions 1, 3 and 4) extend to a multiplayer environment. Hence the subsequent dynamic model of aspirations across rounds is also well-defined, based on (9). Proposition 2 however needs to be extended. With an arbitrary (finite) number of players, it can be verified that the following extension of Proposition 2 holds.

Let $\mathcal{N}$ denote the set of players $\{1, 2, \ldots, n\}$, and $\mathcal{J}$ a *coalition*, i.e., nonempty subset of $\mathcal{N}$. Let the action subvector a_J denote the vector of actions for members of $\mathcal{J}$, and a_{-J} the action vector for the complementary coalition $\mathcal{N} - \mathcal{J}$. Also let A_J denote the aspiration subvector, and $\pi_J(a_J, a_{-J})$ the payoff function for members of $\mathcal{J}$. Then say that a_J is *jointly satisfactory (JS)* for $\mathcal{J}$ given aspirations A_J if $\pi_J(a_J, a_{-J}) \geq A_J$ for all possible action vectors a_{-J} of the complementary coalition $\mathcal{N} - \mathcal{J}$.

In the two player case, if $\mathcal{J}$ is a singleton coalition this corresponds to the notion of a *uniformly satisfactory* action. If $\mathcal{J}$ is the grand coalition it corresponds to the notion of a *mutually satisfactory* action pair. This motivates the following extension of the definition of an intermediate aspiration: A, a vector of aspirations for $\mathcal{N}$, is an *intermediate* aspiration if (i) there exists an action tuple a which is JS for $\mathcal{N}$ relative to aspirations A; and (ii) there does not exist coalition $\mathcal{J}$, a proper subset of $\mathcal{N}$, which has an action subvector a'_J which is JS relative to aspirations A.

Then the following extension of Proposition 2 holds with arbitrarily many players: starting from an intermediate aspiration, the process within any round will almost surely converge to some action tuple which is JS for the grand coalition. In turn this implies that starting from an intermediate level, aspirations must converge to some efficient pso, so Proposition 7 will extend as well.

Moreover, it is easy to see that the sufficient conditions for an action pair to constitute a pso in the two player case, continue to be sufficient in the multiplayer case. Specifically: (i) take any Pareto efficient action tuple a with the property that aspirations $A = \pi(a)$ are intermediate (which generalizes the notion of strict IR). Then a is a pso. (ii) Any

protected strict Nash equilibrium (defined by the same property that unilateral deviations are strictly worse for the deviator, and do not hurt other players) is a pso.[26]

However, further work is needed to identify how the necessary conditions for a pso extend to multiplayer settings.[27] But the preceding discussion indicates that some of the key results of the two player analysis do extend straightforwardly to a multiplayer setting, such as the possibility that players learn to cooperate in the n-player Prisoners Dilemma, efficiently coordinate in n-person coordination games, and collude in oligopolistic settings.

11 Concluding Comments

Our analysis of the long-run implications of reinforcement learning in repeated interaction settings yields new insights. In particular, our findings are sharply distinguished not only from models of rational or best-response learning, but also from evolutionary models. These differences stem from differences both in the nature of the learning rules as well as the interaction patterns typically assumed in these models.

The first distinctive feature is that reinforcement learning models permit convergence to (stage game) non-Nash outcomes, in a manner that appears quite robust across different specifications of the game and the precise nature of reinforcement. This does not result from a specification of reinforcement learning that prevents players from converging to optimal choices in a single person deterministic environment. Rather, players do not converge to best responses owing to a game-theoretic feature, resulting from the interaction in the learning processes of different players. Experimentation with alternative actions may generate temporary payoff gains relative to cooperation, but this change will make the other player dissatisfied and so induces her to deviate in turn, which erodes the gains from the original deviation. A period of trial and error with different actions follows, until they find their way back to a cooperative action pair. Despite the myopic adjustment of behavior states, in effect they respond to others' deviations in a way that mimics a repeated game strategy.

It is also clear that repeated interaction between the *same* set of players is important to ensure such a result. If instead players were randomly selected from a large population to play the game once, and thereafter returned to the population to be matched with other players in later rounds — as in the standard evolutionary models — such interactive dynamic responses would typically not arise.

[26] Part (i) follows from applying the extension of Proposition 2 described above. Part (ii) follows from the fact that single random perturbations of the state of any player starting from a protected strict Nash equilibrium will lead back to the same degenerate distribution concentrated on that action tuple.

[27] Such an extension would also be important in checking whether the exact correspondence between pure steady state aspirations and pso's extend.

Second, convergence to Nash outcomes is possible, but only to a very special subset of Nash outcomes: those which have the protected or saddle-point property. Indeed, our main characterization of pso's established that any inefficient pso must *necessarily* take this form. The saddle-point property ensures that unilateral deviations do not allow subsequent periods of simultaneous trembles by both players, which is necessary for them to discover and gravitate towards more efficient outcomes. Conversely, any protected Nash equilibrium which is also strict is a pso. This saddle-point property is familiar from zero sum games, and explains why pure strategy Nash equilibria in such games are salient when players are reinforcement learners. The median voter outcome in electoral competition models is a particular example.

Third, the particular predictions for games of coordination and cooperation are distinctive. For instance, properties such as risk-dominance do not play a role in games of coordination. These games typically lack protected Nash equilibria, so only payoff dominance considerations matter, and only efficient outcomes can result. In games such as the Prisoners' Dilemma, protected Nash equilibria do exist. Then payoff dominance considerations do not drive the selection: there are multiple Pareto-ordered pso's, suggesting that history plays a significant role in selecting long-run outcomes. It confirms the common intuition that historically low aspirations may be self-reinforcing, and thus subject to hysteresis. When there is an inefficient protected Nash equilibrium, this is robust with respect to small doses of trembles by either player, and thus tends to survive in the long-run.

Acknowledgements

Mookherjee and Ray acknowledge financial support from the National Science Foundation Grant SBR-9709254, and Ray from the John Simon Guggenheim Foundation. This paper has benefitted from comments of Stephen Morris and two anonymous referees, as well as seminar participants at Johns Hopkins, Texas A&M, the 1998 Stony Brook Workshop on Learning and the 1999 Decentralization Conference at New York University.

Appendix: Proofs

LEMMA **1** *(i) Suppose that for some $a \in \mathcal{A}$, player A receives PR no matter what B does; that is, $f(a,b) \geq F$ for all $b \in \mathcal{B}$. Then if player A takes action a at some date, the probability that he will play a forever thereafter (and hence that his state will converge to the pure strategy concentrated on a) is positive and bounded away from zero, irrespective of A's initial state.*

(ii) Similarly, suppose that for some action pair c both players receive PR. Then if c is played at some date, the probability that c will be played forever thereafter (and hence that the state will converge to the pure strategy state concentrated on c) is positive and bounded away from zero, irrespective of the initial state.

Proof. We prove part (i), the proof of (ii) being similar.

Let γ be the going state at any date (say 0) and suppose that A selects a. For $t \geq 1$, let p_t denote the infimum probability that action a is selected at t conditional on being selected at all dates $0, \ldots, t-1$ (where the infimum is taken over all possible actions chosen by B at those dates). Then the probability of an infinite run on a starting at date 1 is bounded below by $\Pi_{t=1}^{\infty} p_t$, which is positive if and only if $\sum_{t=1}^{\infty}[1-p_t] < \infty$. Moreover, the former infinite product is bounded away from zero (with respect to A's state), if the latter infinite sum has an upper bound which is uniform with respect to A's state. Since a is positively reinforced at every date t, $(1-p_{t+1}) \leq (1-\epsilon)(1-p_t)$, so that $(1-p_t) \leq (1-\epsilon)^t$. Hence the required infinite sum is bounded above by $\frac{1}{\epsilon}$. ■

LEMMA **2** *Consider a deterministic decision environment for Player A, e.g., one in which A's payoff function, say f^*, does not depend on player B's action b. Suppose that his aspiration F does not exceed his maximal payoff $\max_a f^*(a)$. Then almost surely player A's choice will converge to an action (and hence his state to a corresponding pure strtategy) which generates a payoff of at least F.*

Proof. If $f^*(a) \geq F$, call *a satisfactory*. Otherwise it is *unsatisfactory*. Given Lemma 1, it suffices to show that at any date t and any initial state α, the probability that some satisfactory action a will be selected at either t or $t+1$, is bounded away from zero.

If a satisfactory action is selected at t there is nothing to prove. So suppose an unsatisfactory action a' is selected at t. Define

$$\theta(a,a') \equiv \min_{\alpha \in \Delta(\mathcal{A})} (1-\epsilon) R^A(\alpha, a', f^*(a'), F)[a]$$

Also let θ denote the minimum over all $\theta(a,a')$ where $f^*(a') < F$. Because $f^*(a') < F$, it follows from assumption NR that $R^A(\alpha, a', f^*(a'), F)[a] > 0$ for every α. By continuity, the minimum of this expression over all α must also be positive. It follows that $\theta(a,a') > 0$. Since the action sets are finite, it also follows that $\theta > 0$, and we are done. ■

LEMMA **3** *If any player experiences NR at any date, then the probability weight on every action for that player at the next two dates is bounded away from zero.*

Proof. Every other action will be selected at the next date with probability bounded away from zero, owing to NR (and the argument that $\theta > 0$). And the same action will also be selected at the next date with probability at least ϵ, owing to inertia. Hence the weight on every action at the next date is bounded away from zero. Applying inertia again, the same is true for the subsequent date as well. ■

LEMMA **4** *For given aspirations $H = (F, G)$, suppose there exists an action pair $c = (a, b)$ satisfying any one of the following properties:*
(i) *c is MS relative to H,*
(ii) *a is uniformly satisfactory (US) for A: $f(a, b') \geq F$ for all b',*
(iii) *b is uniformly satisfactory (US) for B: $g(a', b) \geq G$ for all a'.*

Then for any initial state, the process $\mathcal{P}$ will almost surely either converge to a MS pure strategy state (satisfying (i)), or the state of one player will converge to a pure strategy concentrated on a US action (satisfying either (ii) or (iii)).

Proof. Given Lemma 1 it suffices to show that starting from any initial state, the probability of playing either some MS action pair, or one player playing a US action, within the next two dates, is bounded away from zero.

If no MS or US actions are played at date 0, the action pair $c' = (a', b')$ chosen at 0 violates (i), (ii) and (iii). Hence there exists b'' such that $f(a', b'') < F$.

Suppose first that A does have a US action a. If A received NR at date 0 by playing some non-US action, then she will play the US action at the next date with probability bounded away from zero (special case of Lemma 3), and we are done. So suppose that A received PR at date 0. Then since the action pair c' chosen at 0 does not satisfy (i), player B must receive NR. Then with probability bounded away from zero, player B will play b'' at date 1 while A will repeat a', whence A will receive an NR, and thereafter play the US action a at date 2 with probability bounded away from zero.

A similar argument obviously applies if B has a US action. So to complete the proof, suppose neither player has a US action. Let c be an MS action pair. Because the action pair played at 0 is not MS, at least one player received NR. Without loss of generality, suppose this is A. By Lemma 3, A will then play every action at each of the subsequent two dates with probability bounded away from zero. We are then done if B selected her component of an MS action at 0, or received an NR at 0. Suppose neither: she played b' at 0, and received a PR. Since b' is not an US action, there exists a'' such that $g(a'', b') < G$. Then let A play a'' at 1 and a at 2, while B repeats b' at 1 and thus obtains an NR at 1. Then the MS action pair c will be selected at 2 with probability bounded away from zero, which completes the proof of the lemma. ■

Given aspirations (F, G), call an action a for player A a *USU* (uniformly satisfactory for A and unsatisfactory for B) action if $f(a, b') \geq F, g(a, b') < G$ for all $b \in \mathcal{B}$.

COROLLARY TO LEMMA **4**. Given some aspiration pair, suppose either that a MS action pair exists, or one player has a USU action. Then almost surely $\mathcal{P}$ converges to a pure strategy state that is MS, or one player converges to a USU pure strategy.

[The proof is straightforward given Lemma 4. If player A converges to a US action a which is not USU, there is b for B such that (a, b) is MS. In that case, they will converge to a MS action pair, as B will keep trying different actions whenever dissatisfied, so must eventually discover and stick to a satisfactory action, such as b.]

Proof of Proposition 2. If aspirations are intermediate, then MS action pairs exist, while neither player has a US action. The result follows from Lemma 4. If aspirations are low, then once again MS action pairs exist (e.g., the maxmin action pair), and there are no USU actions. The result then follows from the Corollary to Lemma 4. ■

Proof of Proposition 3. This follows upon applying Theorem 16.2.5 in Meyn and Tweedie (1993), and verifying that the perturbed Markov process is strong Feller and open-set irreducible, hence is a T-chain. ■

Proof of Proposition 4. (i) follows from the tightness of the space of probability measures on a compact metric space in the topology of weak convergence (see, e.g., Parthasarathy (1967)). (ii) follows upon applying the reasoning in proof of Theorem 2 of Karandikar *et al* (1998). Specifically, all but the very last step in that proof uses the properties that $\mathcal{R}$ is well-defined, and that the perturbed Markov process is strong Feller and open set irreducible, to infer that μ^* must be invariant for $\mathcal{Q}.\mathcal{R}$. ■

Proof of Proposition 5. See main text.

In what follows, we deviate from proving propositions in the order stated in the text. [No circularities will be introduced thereby!] We first complete the proofs of the "characterization propositions" 10 and 11.

Proof of Proposition 10. First we show that if c is a pso then it is IR. Otherwise suppose (without loss of generality) that $f(c) < \underline{F}$. Starting from δ_c, and aspirations equal to $h(c)$, a single perturbation of A's state can cause him to converge to his maxmin pure strategy action with positive probability, applying Lemma 1. Hence it is possible to reach a non-mean-payoff-equivalent distribution from δ_c following a single perturbation, contradicting the hypothesis that c is a pso.

To complete the proof, we show that if a pso c is inefficient, it must be a protected Nash equilibrium.

Suppose c is not protected. Then there is an action, say a' for A, such that $g(a', b) < g(c)$. Now suppose that starting from δ_c, A's state is perturbed to a totally mixed state, and a' is played. Then B receives NR, so will play all actions at the next date with positive probability. This is true of A as well, because he is at a totally mixed state. It follows that a Pareto-dominating action pair c^* will be played at the next date with positive probability. In other words, it is possible to reach a Pareto-superior distribution δ_{c^*} from δ_c following a single perturbation, which contradicts the stability of δ_c. Therefore c is protected.

Finally, we show that c is Nash. Otherwise (given that we already know c is protected) there is a unilateral deviation by one player which generates a Pareto improvement. Once again, it is possible to transit from δ_c to another Pareto-superior distribution (the non-deviator being protected will not change her state, and the deviator can converge to the Pareto-dominating pure strategy action) following a single perturbation. This contradicts the stability of c. ■

Proof of Proposition 11. If c is a protected strict Nash equilibrium, then the only distribution that can be reached from δ_c (given aspirations $h(c)$) — following a single perturbation — is δ_c itself. For if A's state is perturbed once, B will continue with the pure strategy b in perpetuity since she is protected in any unilateral deviation by A. And given that B sticks to b, and a is a strict best response for A to b, any other action a' generates a lower payoff to A than his aspiration $f(c)$. Applying Lemma 2, it follows that A must re-converge to the pure strategy action a. Hence only δ_c can be reached from itself following a single perturbation, and $\{\delta_c\}$ constitutes a SP-ergodic set of long-run distributions of $\mathcal{P}_{h(c)}$, the untrembled process with aspirations $h(c)$. So c is a pso.

If c is efficient and strictly IR, then Proposition 2 implies that the only long-run distributions of $\mathcal{P}_{h(c)}$ are of the form δ_{c^*} with $h(c^*) = h(c)$. All these distributions can be reached from one another following a single perturbation: this is obvious for any pair representing a deviation by a single player. In the case of a pair c^1, c^2 where $a^1 \neq a^2$ and $b^1 \neq b^2$, note that the strict IR property implies that neither player is protected against certain unilateral deviations by the other player. Hence a perturbation of the state of any player will with positive probability cause a transition from one long-run distribution to the other (e.g., one player deviates in a way to hurt the other player and then at the subsequent date the alternative pure action pair will be played with positive probability, following which Lemma 1 ensures that an infinite run on the new action pair is possible). So the entire set of long-run distributions δ_{c^*} of the form $h(c^*) = h(c)$ is SP-ergodic, and c is a pso. ■

Proof of Proposition 6: Consider first part (a). If c is a pso, note first that no USU actions exist relative to aspirations $h(c)$ — otherwise a single perturbation can result in the play of an USU action, and eventual convergence to a long-run distribution of $\mathcal{P}_{h(c)}$

which is not payoff-equivalent to δ_c. This would contradict the hypothesis that c is a pso.

Thus — since no USU actions exist and since c is MS relative to aspirations $h(c)$ — the corollary to Lemma 4 assures us that every long-run distribution of $\mathcal{P}_{h(c)}$ is of the form $\delta_{c'}$ where $h(c') \geq h(c)$. And Proposition 2 ensures that $\mathcal{R}_{h(c)}$ is well-defined. Proposition 4(ii) then implies that $\mathcal{D}(h(c))$ is the support of an invariant distribution of $\mathcal{Q}_{h(c)}.\mathcal{R}_{h(c)}$, and so can be partitioned into a collection of disjoint, nonempty SP-ergodic subsets $\mathcal{S}_1, \ldots, \mathcal{S}_K$.

We must show that every $\mu \in \mathcal{D}(h(c))$ has mean payoff $h(c)$.

If c is efficient then every long-run distribution of $\mathcal{P}_{h(c)}$ is of the form $\delta_{c'}$ where $h(c') = h(c)$, so the same is true for every distribution in $\mathcal{D}(h(c))$. So suppose that c is inefficient. Then it must be a protected Nash equilibrium, by Proposition 10. Suppose there exists a long-run distribution in $\mathcal{D}(h(c))$ of the form $\delta_{c'}$ with $h(c') \geq h(c)$ and $h(c') \neq h(c)$. We claim that δ_c can be reached from $\delta_{c'}$ following a sequence of unilateral perturbations. Starting from c' (with aspirations $h(c)$), perturb the state of A to ensure that a is played with positive probability. Then a is a US action relative to aspirations $h(c)$ since c is protected. So A will (with positive probability) converge to the pure strategy corresponding to action a. And given that A selects a with probability one, Lemma 2 ensures that B must converge to an action b'' which is payoff equivalent to c, i.e., $h(a, b'') = h(c)$. If $b'' = b$ we are done. If not then note that δ_c can be reached from $\delta_{(a,b")}$ following a single (further) perturbation of B's state.

Hence δ_c must belong to the same SP-ergodic subset $\mathcal{S}_k$ as $\delta_{c'}$. This implies that $\delta_{c'}$ can also be reached from δ_c following a sequence of single perturbations, which contradicts the hypothesis that δ_c is a stable distribution. So every distribution in $\mathcal{D}(H)$ must generate payoff $h(c)$, implying that $h(c)$ is a steady state aspiration. It is also a pure steady state aspiration because every element of $\mathcal{D}(h(c))$ is concentrated on a pure strategy state. We have therefore established that if c is a pso then $h(c)$ is a pure steady state aspiration.

The second part of (a) follows directly from the definition of a pure steady state aspiration.

Turn now to part (b). H is a steady state aspiration implies that every $\mu' \in \mathcal{D}(H)$ satisfies $H = \int h d\mu'$. Note also that weak ergodicity of $\mathcal{P}$ implies that there exists a nonempty subset of $\mathcal{D}(H)$ which is SP-ergodic. (This is because the finite set of long-run distributions of $\mathcal{P}$ can always be partitioned into a collection of nonempty SP-ergodic sets, and a 'transient' subset). Then take any μ in a SP-ergodic subset of $\mathcal{D}(H)$: this is stable.

To prove (c), note that if μ is stable, it belongs to an SP-ergodic set $\mathcal{S}$ of long-run distributions of $\mathcal{P}$, where $H = \int h d\mu'$ for every μ' in $\mathcal{S}$. If $\mathcal{R}$ is well-defined and $\mathcal{Q}.\mathcal{R}$ has a unique invariant distribution, there is a unique SP-ergodic set of long-run distributions of $\mathcal{P}$. By Proposition 4(ii), $\mathcal{D}(H)$ must coincide with $\mathcal{S}$. Hence every μ' in $\mathcal{D}(H)$ generates

mean payoff H, so H is a steady state aspiration. ■

Proof of Proposition 7. Following the argument in the text, with probability one there is an integer T^* and pure action pair c^* such that $\Pi_T = h(c^*)$ for all $T > T^*$, and H_T converges to $h(c^*)$ from below. Since H_T is intermediate for all T, Proposition 2 implies that every long run distribution of $\mathcal{P}_{H_T}$ is concentrated on a pure strategy state that is MS relative to H_T (in particular, $\mathcal{R}_{H_T}$ is well-defined). Hence for $T > T^*$, $\mu_T = \delta_c$ for some c satisfying $h(c) = h(c^*)$. Since the action sets are finite, we can find T^{**} large enough so that for any $T > T^{**}$, there is an action pair c payoff-equivalent to c^* such that $\mu_T = \delta_c$, which has the property that $\delta_c \in \mathcal{D}(H_{T_n}), n = 1, 2, \ldots$ along some subsequence $T_n \to \infty$.

Take any such δ_c: it suffices to show it is a pso. To this end, we use the characterization in Proposition 11. First, note that c is strongly IR by construction (because $h(c) = h(c^*) \geq H_T >> (\underline{F}, \underline{G})$).

Next we show that c is efficient. Suppose not; then c is Pareto-dominated by some other action pair c'. We claim that in this case, c must be protected.

Otherwise, there is a player, say B, who is worse off if the other player (A) deviates to some action $\tilde{a}$, i.e., $g(\tilde{a}, b) < g(a, b)$. Since $H_T \to h(c^*) = h(c)$, it follows that $g(\tilde{a}, b) < G_T$ for T sufficiently large. For any such T, a single perturbation of δ_c (e.g., if A's state is perturbed to a totally mixed state putting positive probability weights on $\tilde{a}$ and a') will cause $\mathcal{P}_{H_T}$ to transit to $\delta_{c'}$ with positive probability. For instance, at the first date A could play $\tilde{a}$ while B plays b and receives NR, and at the next date c' is played, which is MS relative to aspirations $H_T \leq h(c)$. By Lemma 1 there will be an infinite run on c' thereafter with positive probability.

So $\delta_{c'}$ can be reached from δ_c following a single perturbation, when players' aspirations are H_T, for T sufficiently large. Since $\mathcal{R}_{H_T}$ is well-defined for all T, Proposition 4(ii) implies that $\delta_{c'}$ must be in $\mathcal{D}(H_T)$, and hence $\Pi(H_T) = h(c')$ with positive probability, for all large T. This contradicts the requirement that $\Pi(H_T) = h(c) \neq h(c')$ for all $T > T^*$ with probability one.

Therefore c is protected. But this conclusion contradicts the fact that c is strongly IR (protection implies that maxmin payoffs are at least as high as $h(c)$). It follows that our original supposition is erroneous, and that c is indeed efficient.

Now use Proposition 11 to infer that c is a pso. ■

Proof of Proposition 8: Let c^* be a pso. Define $\mathcal{I}(c^*, \epsilon) \equiv [f(c^*) - \epsilon, f(c^*)] \times [g(c^*) - \epsilon, g(c^*)]$. We first show that for $\epsilon > 0$ sufficiently small, USU actions cannot exist (relative to any aspiration $H \in \mathcal{I}(c^*, \epsilon)$).

Otherwise there is a sequence $\epsilon_n \to 0+$ and aspirations $H_n \in \mathcal{I}(c^*, \epsilon_n)$ such that one player (A, say) has a USU action a_n relative to H_n for each n. Hence $f(a_n, b) \geq f(c^*) - \epsilon_n$ and $g(a_n, b) < g(c^*)$ for all b and for all n. Since $\mathcal{A}$ is finite it follows that there exists

$a' \in \mathcal{A}$ such that $a_n = a'$ for infinitely many n. Hence $f(a', b) \geq f(c^*)$ and $g(a', b) < g(c^*)$ for all b. Then with aspirations fixed at $h(c^*)$, a single perturbation of δ_{c^*} would cause $\mathcal{P}$ to transit to a long-run distribution where A selects a' with probability one, and B is perpetually dissatisfied. Since this distribution must generate B a lower average payoff than δ_{c^*}, we contradict the hypothesis that c^* is a pso.

Hence for small enough ϵ, the Corollary to Lemma 4 ensures that $\mathcal{P}$ must converge to some MS action pair, for any aspirations $P \in \mathcal{I}(c^*, \epsilon)$. The finiteness of the action sets implies that for small enough ϵ, an action pair that is MS relative to some $P \in \mathcal{I}(c^*, \epsilon)$ is MS relative to $h(c^*)$. So for small ϵ, the only long-run distributions of $\mathcal{P}$ for any $P \in \mathcal{I}(c^*, \epsilon)$ are of the form δ_c where $h(c) \geq h(c^*)$.

If c^* is efficient then every MS action pair relative to aspirations $h(c^*)$ will be payoff-equivalent to c^*. Hence for small ϵ, every distribution in $\mathcal{D}(H)$ for every $P \in \mathcal{I}(c^*, \epsilon)$ is of the form δ_c where $h(c) = h(c^*)$. Then $H_0 \in \mathcal{I}(c^*, \epsilon)$ implies that with probability one, for any T: $\mu_T = \delta_c$ for some c payoff equivalent to c^*, and H_T converges to $h(c^*)$.

If c^* is inefficient, Proposition 10 implies it must be a protected Nash equilibrium. Hence given any $\epsilon > 0$ and any aspirations $H \in \mathcal{I}(c^*, \epsilon)$, it is not possible to reach a Pareto-superior $\delta_{c'}$ from δ_{c^*} following a single perturbation (if one player deviates unilaterally, the other player continues to be satisfied relative to aspirations set equal to $h(c^*)$ and hence relative to aspirations $H \leq h(c^*)$ as well; hence the latter player will not deviate from her component of c^*).

On the other hand for ϵ sufficiently small (eg., smaller than the smallest difference between two distinct payoffs for any player), and for any $P \in \mathcal{I}(c^*, \epsilon)$, it is possible to reach δ_{c^*} from a Pareto-superior $\delta_{c'}$ following a sequence of single perturbations. For instance, if A's state is perturbed and he plays a^*, he is guaranteed a payoff of at least $f(c^*) \geq F$, no matter what B does. Hence Lemma 1 implies that with positive probability he will converge to the pure strategy a^*. In this event, B's state must converge to a pure strategy b such that $g(a^*, b) = g(c^*)$ if ϵ is sufficiently small, by virtue of Lemma 2. Hence it is possible to reach $\delta_{(a^*,b)}$ from $\delta_{c'}$ following a single perturbation for any sufficiently small ϵ. Note finally that δ_{c^*} can be reached from $\delta_{(a^*,b)}$ following a single perturbation of B's state which makes her play b^* with positive probability.

It follows that if c^* is inefficient, then for sufficiently small ϵ and for any aspirations $H \in \mathcal{I}(c^*, \epsilon)$, there cannot be a distribution in $\mathcal{D}(H)$ which Pareto-dominates δ_c. Hence $\mathcal{D}(H)$ for any such H is of the form δ_c such that $h(c) = h(c^*)$, and now we can apply the same reasoning as in the case where c^* is efficient. ■

Proof of Proposition 9. If in round T the aspiration H_T lies in the interval $(\underline{F}, \pi^*]$, the result follows from an application of Proposition 7. If $H_T \leq \underline{F}$ then there exists at least one MS action pair (e.g., c^*) relative to H_T, and each player's maxmin action is uniformly satisfactory (i.e., generates payoff at least as large as the aspiration, no matter what the other player does). Then USU actions cannot exist and the corollary to Lemma

4 implies that the unperturbed process will almost surely converge to some pure strategy state which is MS relative to H_T. Hence in round T players must attain an average payoff between H_T and π^*, implying that $H_{T+1} \in (H_T, \pi^*)$. An argument analogous to that used in proving Proposition 7 then establishes the result.

It remains to consider the case where $H_T > \pi^*$ for all T. Since the average payoff in any round cannot exceed π^*, it follows that $H_{T+1} \leq \tau H_T + (1-\tau)\pi^*$, i.e., $H_{T+1} - \pi^* \leq \tau(H_T - \pi^*) \leq \tau^2(H_{T-1} - \pi^*) \leq \ldots \leq \tau^{T+1}[H_0 - \pi^*]$ for all T. So H_T must then converge to π^*, which in turn requires that the probability of c^* under μ_T converges to 1. Finally, note that π^* is a pure steady state aspiration, and c^* is a pso, since with aspirations at $\pi^* = h(c^*)$ every long-run distribution of the unperturbed process must generate a payoff of $h(c^*)$, by Proposition 2. So δ_{c^*} is the unique long-run distribution of $\mathcal{P}_{h(c^*)}$, hence the only member of $\mathcal{D}(h(c^*))$. ■

Proof of Proposition 12. Let $(\hat{F}, \hat{G})$ denote the payoffs resulting when both players seelct their maxmin actions. By definition of maxmin payoffs, $(\hat{F}, \hat{G}) \geq (\underline{F}, \underline{G})$. Since the game is generic, either $(\hat{F}, \hat{G}) >> (\underline{F}, \underline{G})$, or $(\hat{F}, \hat{G}) = (\underline{F}, \underline{G})$. In the former case, there exists an efficient, action pair that is strictly IR (take any efficient action pair that Pareto-dominates or is payoff-equivalent to the maxmin action pair). When $(\hat{F}, \hat{G}) = (\underline{F}, \underline{G})$, and are equal to players' aspirations H, USU actions cannot exist. Then the Corollary to Lemma 4 implies that every long-run distribution of $\mathcal{P}$ is of the form δ_c where $h(c) \geq H$.

If the maxmin action pair is efficient, every long-run distribution of $\mathcal{P}$ generates payoff H, and one of them must be stable (since the set of these distributions must contain an SP-ergodic set). If it is inefficient, the genericity of the game implies that there exists an efficient action pair, which is strictly IR, and so must be a pso. ■

Proof of Proposition 13. See main text.

Proof of Proposition 14. (i) follows from an argument analogous to the first step in the proof of Proposition 10. If (ii) is false there is a pure action pair c which Pareto dominates (F, G) and one player (say B) has a stochastic payoff in μ. Then there must be a state γ in the support of μ in which some action pair $c' = (a', b')$ is played with positive probability, where B is dissatisfied: $g(c') < G$. Starting with γ, perturb A's state to a totally mixed state, which assigns positive weight to a and a'. Suppose c' is played initially, causing B's choice of b' to receive NR. Then at the subsequent date c will be played with positive probability; by Lemma 1 there will be an infinite run on c with positive probability. Hence δ_c could be reached from μ following a single perturbation. Since they are not mean-payoff-equivalent, we contradict the hypothesis that μ is stable. ■

References

B. Arthur, "On Designing Economic Agents that Behave Like Human Agents," *Journal of Evolutionary Economics* **3**, 1993, 1–22.

J. Bendor, D. Mookherjee and D. Ray, "Aspirations, Adaptive Learning and Cooperation in Repeated Games", Discussion Paper, Planning Unit, Indian Statistical Institute, New Delhi, 1992.

———————, "Aspirations, Adaptive Learning and Cooperation in Repeated Games", Discussion Paper No. 9442, Center for Economic Research, Tilburg University, May 1994. Revised, mimeo, Department of Economics, Boston University, 1995.

———————, "Aspiration-Based Reinforcement Learning in Repeated Games: An Overview," mimeo, Department of Economics, Boston University, 2000.

K. Binmore and L. Samuelson, "Muddling Through: Noisy Equilibrium Selection," *Journal of Economic Theory*, 74, 1997, 235-265.

T. Börgers and R. Sarin, "Learning through Reinforcement and Replicator Dynamics," *Journal of Economic Theory* **77**, 1997, 1-14.

———————- , "Naive Reinforcement Learning with Endogenous Aspirations," *International Economic Review* **41**, 2000, 921–950.

T. Börgers, A. Morales and R. Sarin, "Simple Behavior Rules which Lead to Expected Payoff Maximizing Choices," mimeo, University College, London, 1998.

R. Bush and F. Mosteller, *Stochastic Models of Learning*, New York: John Wiley and Sons, 1955.

R. Bush, F. Mosteller and G. Thompson, "A Formal Structure For Multiple Choice Situations," in *Decision Processes*, edited by R.M. Thrall, C.H. Coombs and R.L. Davis, 1954, New York: Wiley.

C. Camerer and T. Ho, "Experience-weighted Attraction Learning in Normal Form Games," *Econometrica*, 67(4), 1999, 827-874.

J. Cross, "A Stochastic Learning Model of Economic Behavior," *Quarterly Journal of Economics*, 87 (1973), 239-266.

R. Cyert and J. March, *A Behavioral Theory of the Firm.* Englewood-Cliffs, NJ: Prentice-Hall, 1963.

H.D. Dixon, "Keeping up with the Joneses: Competition and the Evolution of Collusion", *Journal of Economic Behavior and Organization*, 2000, forthcoming.

I. Erev and A. Roth, "On the Need for Low Rationality, Cognitive Game Theory: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria," working paper, August 1995, Department of Economics, University of Pittsburgh.

————————-, "Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique Mixed Strategy Equilibria," *American Economic Review*, 88, 1998, 848-881.

W. Estes, "Individual Behavior in Uncertain Situations: An Interpretation in Terms of Statistical Association Theory," in *Decision Processes*, edited by R.M. Thrall, C.H. Coombs and R.L. Davis, 1954, New York: Wiley.

I. Gilboa and D. Schmeidler, "Case-based Decision Theory," *Quarterly Journal of Economics*, 110, 1995, 605-640.

————————, "Case-based Optimization," *Games and Economic Behavior*, 15 (1996), 1-26.

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

Y. Kim, "Satisficing, Cooperation and Coordination," mimeo, Department of Economics, Queen Mary and Westfield College, University of London, 1995a.

———-, "A Satisficing Model of Learning in Extensive Form Games," mimeo, Department of Economics, Yonsei University, Seoul, 1995b.

B. Lipman (1991), "How to Decide How to Decide How to ...: Modeling Limited Rationality," *Econometrica* **59**, 1105–1125.

R. Duncan Luce (1959), *Individual Choice Behavior*, John Wiley.

S.P. Meyn and R.L. Tweedie (1993) *Markov Chains and Stochastic Stability*, London, New York: Springer-Verlag.

D. Mookherjee and B. Sopher, "Learning Behavior in an Experimental Matching Pennies Game," *Games and Economic Behavior* **7**, 1994, 62–91.

————————, "Learning and Decision Costs in Experimental Constant Sum Games," *Games and Economic Behavior* **19**, 1997, 97–132.

K. Narendra and P. Mars, "The Use of Learning Algorithms in Telephone Traffic Routing: A Methodology," *Automatica*, 19(5), 1983, 495-502.

K. Narendra and M. Thathachar, *Learning Automata: An Introduction*, Englewood Cliffs: Prentice Hall, 1989.

R. Nelson and Winter, S. *An Evolutionary Theory of Economic Change.* Cambridge, Massachussetts: Harvard University Press, 1982.

M. F. Norman, *Markov Processes and Learning Models*, New York and London: Academic Press, 1972.

F. Palomino and F. Vega-Redondo, "Convergence of Aspirations and (Partial) Cooperation in the Prisoner's Dilemma", *International Journal of Game Theory*, 28(4), 1999, 465-488.

G. Papavassilopoulos, "Learning Algorithms for Repeated Bimatrix Nash Games with Incomplete Information," *Journal of Optimization Theory and Applications*, 62(3), 1989, 467-488.

K. R. Parthasarathy (1967), *Probability Measures on Metric Spaces*, New York: Academic Press.

A. Pazgal, "Satisficing Leads to Cooperation in Mutual Interest Games", *International Journal of Game Theory*, 26, 1997, 439-453.

A. Robson and F. Vega-Redondo, "Efficient Equilibrium Selection in Evolutionary Games with Random Matching," *Journal of Economic Theory*, 70(1), 1996, 65-92.

A. Roth and I. Erev, Learning in extensive-form games: experimental data and simple dynamic models in the intermediate term," *Games Econ. Behav.* **8**, 1995, 164–212.

R. Selten, "The Chain Store Pradaox," *Theory and Decision*, 9, 1978, 127-159.

————-, "Evolution, Learning and Economic Behavior," *Games and Economic Behavior* **3** (1991), 3–24.

R. Selten and R. Stoecker, End behavior in sequences of finite Prisoners' Dilemma supergames," *J. Econ. Behav. Organ.***7** (1986), 47–70.

H. Simon, "A Behavioral Model of Rational Choice," *Quarterly Journal of Economics*, 69 (1955), 99-118.

———- , *Models of Man*, 1957, New York.

———- , "Theories of Decision Making in Economics and Behavioral Science," *American Economic Review*, 49(1), 1959, 253-283.

P. Suppes and R. Atkinson, "Markov Learning Models for Multiperson Interactions," Stanford: Stanford University Press, 1960.

P. Young, "The Evolution of Conventions," *Econometrica*, 61(1), 1993, 57-84.