Evolution in Bayesian Games II: Stability of Purified Equilibria[*]

William H. Sandholm
Department of Economics
University of Wisconsin
1180 Observatory Drive
Madison, WI 53706
whs@ssc.wisc.edu
www.ssc.wisc.edu/~whs

tel.: 608-263-3858
fax: 608-263-3876

October 16, 2006

[*] I thank Jeff Ely and Larry Samuelson, as well as an anonymous referee and Associate Editor, for helpful comments. Financial support from the National Science Foundation (SES-0092145) is gratefully acknowledged.

Abstract

We study the evolutionary stability of purified equilibria of two-player normal form games, providing simple sufficient conditions for stability and for instability under the Bayesian best response dynamic.

JEL classification numbers: C72, C73

Running title: Stability of Purified Equilibrium

1. Introduction

In a mixed equilibrium of a normal form game, each player is indifferent between his equilibrium strategy and all other strategies with the same support. This raises the question of why we should expect players to randomize in precisely the fashion that their equilibrium strategies dictate.

To address this issue, Harsanyi [12] shows that every equilibrium of almost every normal form game can be viewed as a strict equilibrium of a Bayesian game created by slightly perturbing the payoffs of the normal form game. In fact, Harsanyi proves that these purified equilibria exist regardless of the distribution of payoff noises so long as these noises become small.

In this paper, we establish general sufficient conditions for the stability and instability of purified equilibria of two-player normal form games under evolutionary dynamics. To do so, we introduce a population interpretation of Harsanyi's perturbed game. Instead of viewing this game as being played by a small group of players, each of whom has many possible type realizations, we replace each player with a population of small agents, and assume that in each of these populations, the entire type distribution is realized at once.

In studying stability, we adopt the view of Binmore and Samuelson [2] and Ellison and Fudenberg [5] that the relevant payoff perturbations in real world environments need not be especially small. Therefore, none of our conditions for stability or instability requires payoff perturbations to vanish in size.

To pose our questions, we must fix our notions of stability and instability by specifying an evolutionary dynamic for Bayesian population games. For this, we rely on the Bayesian best response dynamic of Ely and Sandholm [6], under which the behavior of the subpopulation of agents of type θ adjusts in the direction type θ's current best response. Ely and Sandholm [6] show that the rest points of the Bayesian best response dynamic are precisely the Bayesian equilibria of the underlying game, and they prove that stability analysis for their infinite dimensional dynamic reduces to stability analysis of a finite dimensional dynamic that describes the evolution of aggregate behavior. The work of Ely and Sandholm [6] serves two roles in the present study: it provides the dynamic used to define evolutionary stability, and it supplies the techniques that allow the analysis of stability to be performed using finite dimensional methods.

To begin our analysis, we observe that the relevant finite dimensional dynamic is equivalent to the perturbed best response dynamic, a dynamic introduced in the study of the learning process known as stochastic fictitious play (Fudenberg and Kreps [8]). A variety of methods of analyzing the perturbed best response dynamic now exist, based alternatively on linearization, Lyapunov's direct method, and the theory of monotone dynamical systems—see Hopkins [20], Hofbauer [14], Hofbauer and Sandholm [17, 18], and Hofbauer and Hopkins [15]. By developing these tools, we are able to establish simple sufficient conditions for stability and instability of purified equilibria under the Bayesian best response dynamic, and to provide intuition about the features of games and type distributions that drive stability and instability. Two implications of our results are especially worthy of note: our sufficient conditions imply that every Nash equilibrium of every normal form game can be purified in a stable fashion, but that there are normal form games whose Nash equilibria cannot be purified in an unstable fashion.

Dynamic stability of purified equilibria was first studied by Ellison and Fudenberg [5], who consider the stability of purified equilibria under population fictitious play.[1] Like Harsanyi, Ellison and Fudenberg [5] parameterize the distribution of types by a scalar disturbance level.[2] They study local stability of purified equilibria as this level becomes small, providing necessary and sufficient conditions for stability of sequences of purified equilibria of 2 x 2 and 3 x 3 games. Our analysis extends Ellison and Fudenberg's on many fronts: we offer sufficient conditions for both stability and instability of purified equilibrium; our conditions apply to games with arbitrary numbers of strategies; and these conditions hold force whether or not payoff perturbations are small.

Ellison and Fudenberg [5] provide examples of games whose mixed equilibria only admit sequences of unstable purifications. While at first glance these examples might appear to contradict our general result on the existence of stable purifications, this apparent discrepancy is easily resolved. By using a parameterization to introduce the small noise limit, Ellison and Fudenberg [5] implicitly impose restrictions on payoff perturbations other than the restriction that these perturbations become small: in particular, their parameterization forces the densities of the smallest perturbations to grow without bound. We show that it is this restriction that drives their instability results; without it, stable purification is always possible.

In fact, our analysis affirms and even strengthens Ellison and Fudenberg's [5] main conclusions. These authors describe their results as "fairly supportive of the idea that

[1] In population fictitious play, agents in a large population always play best responses to their current beliefs, which are defined to equal the time average of the population's past behavior. Population fictitious play is equivalent after a reparameterization of time to our aggregate best response dynamic, and hence to the perturbed best response dynamic. For a detailed explanation of this point, see Ely and Sandholm [6].

[2] In particular, they describe some fixed distribution of payoff perturbations using a random vector $\tilde{\theta}$; to reduce the noise level, they suppose that the actual distribution of perturbations is given by the rescaled random vector $\varepsilon\tilde{\theta}$.

populations may learn to play certain mixed strategy equilibria" (p. 86). In this paper, we provide general, intuitive conditions for stability of purifications; we impose no restrictions on the number of available strategies and only limited restrictions on the nature of the distributions of types. Therefore, our analysis substantially augments the set of environments in which stable purification is known to occur.[3]

2. Bayesian Population Games and Purification of Equilibrium

We analyze the stability of purified equilibria in two different contexts: pairwise random matching of a single population to play a symmetric two player normal form game, and pairwise random matching of two populations to play a general two player normal form game. We begin by introducing the definitions we require.

2.1 Normal Form Games

In a *symmetric two player normal form game,* the two players have the same strategy set $S^1 = \{1, \dots, n^1\}$ and the same payoff matrix $A \in \mathbf{R}^{n \times n}$; A_{ij} is the payoff a player obtains if he plays i and his opponent plays j. Mixed strategies for each player are elements of the simplex $X^1 = \{x^1 \in \mathbf{R}^{n^1}_+ : \sum_{i \in S} x^1_i = 1\}$. For reasons that will become clear in the next subsection, we sometimes omit the superscripts from our notation when working in this symmetric setting.

In a (*standard*) *two player normal form game,* players 1 and 2 have strategy sets $S^1 = \{1, \dots, n^1\}$ and $S^2 = \{1, \dots, n^2\}$ and payoff matrices $A \in \mathbf{R}^{n^1 \times n^2}$ and $B \in \mathbf{R}^{n^1 \times n^2}$. When the players select strategies $i \in S^1$ and $j \in S^2$, they obtain payoffs of A_{ij} and B_{ij}, respectively. Player p's mixed strategies are elements of $X^p = \{x^p \in \mathbf{R}^{n^p}_+ : \sum_{i \in S^p} x^p_i = 1\}$, while mixed strategy profiles $x = (x^1, x^2)$ are elements of $X^1 \times X^2$.

2.2 Population Games Defined By Random Matching

We consider two models of random matching in large populations corresponding to the two classes of normal form games described above. Under *single population* (or *symmetric*) *matching,* pairs of agents are chosen at random from a unit mass population to play the symmetric normal form game $A \in \mathbf{R}^{n^1 \times n^1}$. Under *two population* (or *standard*) *matching,* one agent is chosen at random from each of two unit mass populations, and

[3] We know of three other papers that consider purification in evolutionary contexts. Binmore and Samuelson [3] use static evolutionary stability concepts to analyze the tension between the instability of mixed equilibria when players may condition behavior on roles and the stabilizing effects of payoff perturbations. Sandholm [25] studies the robustness of purified equilibria to the evolution of preferences. Finally, Echenique and Edlin [4] address the instability of purified mixed equilibria in supermodular games.

the pair of agents then face one another in the normal form game $(A, B) \in \mathbf{R}^{n^1 \times n^2} \times \mathbf{R}^{n^1 \times n^2}$.

To define a common notation for the one and two population cases, we let p (equal to 1 or 2) denote the number of populations in the situation at hand and let $\mathcal{P}$ (equal to $\{1\}$ or $\{1, 2\}$) denote the set of populations. In both cases, we let $X \subset \mathbf{R}^n$ denote the set of *social states;* elements of X specify the distribution of strategies in each population $p \in \mathcal{P}$. In the single population case, $n = n^1$, and $X = X^1$ is just the simplex; in the two population case, $n = n^1 + n^2$, and $X = X^1 \times X^2$ is the product of two simplices.

Agents evaluate strategies by computing their expected payoffs at the current social state. In the single population case, the (expected) payoff to choosing strategy $i \in S$ at social state x is

$$F_i(x) = \sum_{j \in S} A_{ij} x_j \,.$$

Therefore, the payoffs to all strategies are described by the vector field F: $X \to \mathbf{R}^n$, defined by

$$F(x) = Ax.$$

In the two population case, the payoffs to strategies $i \in S^1$ and $j \in S^2$ are

$$F_i^1(x^2) = \sum_{j \in S} A_{ij} x_j^2 \quad \text{and} \quad F_j^2(x^1) = \sum_{i \in S} x_i^1 B_{ij} \,.$$

Thus, the payoffs to all strategies in both populations are described by the vector field F: $X \to \mathbf{R}^n$, where

$$F(x) = \begin{pmatrix} F^1(x^2) \\ F^2(x^1) \end{pmatrix} = \begin{pmatrix} 0 & A \\ B' & 0 \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \end{pmatrix}.$$

We sometimes write $F^p(x)$ for $F^p(x^{-p})$ when it is convenient to do so.

We conclude by reviewing the relevant definition of equilibrium for each setting. State $x^* \in X^1$ is a *symmetric Nash equilibrium* of the symmetric normal form game A if

$$x_i^* > 0 \text{ implies that } F_i(x^*) \geq F_{i'}(x^*) \text{ for all } i' \in S^1.$$

Similarly, state $x^* = ((x^*)^1, (x^*)^2)$ is a *Nash equilibrium* of the normal form game (A, B) if

$$(x^*)_i^1 > 0 \text{ implies that } F_i^1(x^*) \geq F_{i'}^1(x^*) \text{ for all } i' \in S^1, \text{ and}$$
$$(x^*)_j^2 > 0 \text{ implies that } F_j^2(x^*) \geq F_{j'}^2(x^*) \text{ for all } j' \in S^2.$$

2.3 Bayesian Population Games

In a *Bayesian population game,* different agents within each population $p \in \mathcal{P}$ have different payoff functions. In the present model, an agent in population p who chooses strategy i receives a payoff of $F_i^p(x) + \theta_i^p$. The first term in this sum, $F_i^p(x)$, is the *common payoff* to strategy i; it is determined by the underlying normal form game and is the same for all agents in population p. The second term, θ_i^p, is the agent's *idiosyncratic payoff* to strategy i. It is a component of the agents' *type,* a vector that we denote by $\theta^p \in \Theta^p = \mathbf{R}^{n^p}$.[4]

The measure μ^p on Θ^p represents the (realized) distribution of types in population $p \in \mathcal{P}$. We assume that μ^p satisfies the smoothness assumptions utilized in Ely and Sandholm [6] and Hofbauer and Sandholm [17], so that the results from those papers can be applied in our analysis.[5] We let $\mu = \{\mu^p\}_{p\in\mathcal{P}}$ denote the profile of type distributions.[6] The *Bayesian population game* generated by the common payoff function F and the profile of type distributions μ is denoted $BG(F, \mu)$.

A complete description of population p's behavior is given by a *Bayesian strategy* $\sigma^p \in \Sigma^p = \{\sigma^p\colon \Theta^p \to X^p\}$. The strategy distribution $\sigma^p(\theta^p) \in X^p$ describes the behavior of the subpopulation of players from population p who are of type θ^p. Bayesian strategies that specify the same strategy distribution for almost all subpopulations are considered equivalent. If we let $\Sigma = \prod_{p\in\mathcal{P}} \Sigma^p$, then each $\sigma = \{\sigma^p\}_{p\in\mathcal{P}} \in \Sigma$ is called a *Bayesian strategy profile.*

Aggregate behavior in population p is described by $E^{\mu^p}\sigma^p = \int_{\Theta^p} \sigma^p(\theta^p)\,d\mu^p(\theta^p) \in X^p$, while $E^{\mu}\sigma = \{E^{\mu^p}\sigma^p\}_{p\in\mathcal{P}} \in X$ describes aggregate behavior in the whole society. We omit superscripts from the expectation operators when no confusion will arise.

The *Bayesian best response function* $B^p\colon X \to \Sigma^p$ is a map from social states to optimal Bayesian strategies for population p. These optimal Bayesian strategies involve best responses by almost all types $\theta^p \in \Theta^p$:

$$B^p(x)(\theta^p) = \underset{y^p\in X^p}{\operatorname{argmax}}\; y^p \cdot (F^p(x) + \theta^p).$$

We let $B(x) = \prod_{p\in\mathcal{P}} B^p(x)$ denote the profile of best response functions.

[4] This specification of types follows Ellison and Fudenberg [5]. In contrast, Harsanyi allows an agent's type to specify an idiosyncratic payoff that depends not only on his own choice of strategy, but also on his opponent's choice of strategy.

[5] In particular, we assume that μ^p admits a bounded density, that the function B^p (defined below) is Lipschitz continuous, and that the function C^p (also defined below) is continuously differentiable at all equilibrium payoff vectors π^p.

[6] Of course, "profiles" in the single population case actually consist of a single element.

The Bayesian strategy profile $\sigma \in \Sigma$ is a *Bayesian equilibrium* of $BG(F, \mu)$ if $\sigma = B(E\sigma)$. Put differently, a Bayesian strategy profile is a Bayesian equilibrium if almost all subpopulations play best responses to the current population state.

2.4 Purification

In order to define purified equilibria of normal form games, we must say what it means for a Bayesian game to be close to a normal form game. In the symmetric case, we say that $BG(F, \mu)$ is an *ε-approximation* of the normal form game A if

(*i*) $F(x) = Ax$, and

(*ii*) $\mu\{\theta\colon |\theta| > \varepsilon\} < \varepsilon$.

The second requirement asks that most of the mass in the type distribution μ be placed near the origin. Similarly, in the standard case we call $BG(F, \mu)$ an ε-approximation of the normal form game (A, B) if

(*i*) $F(x) = \begin{pmatrix} 0 & A \\ B' & 0 \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \end{pmatrix}$, and

(*ii*) $\mu^p\{\theta^p : |\theta^p| > \varepsilon\} < \varepsilon$ for $p \in \{1, 2\}$.

In either setting, we say that the Bayesian strategy σ is an *ε-purification* of the Nash equilibrium x^* if

(*i*) $BG(F, \mu)$ is an ε-approximation of F,

(*ii*) σ is a Bayesian equilibrium of $BG(F, \mu)$, and

(*iii*) $|x - x^*| < \varepsilon$, where $x = E^\mu \sigma$.

Our three conditions for σ to be a purification of x^* require (*i*) that the game with perturbed payoffs be close to those of the original game, (*ii*) that σ be a Bayesian equilibrium of the perturbed game, and (*iii*) that the strategy distribution $x = E^\mu \sigma$ induced by σ be close to x^*.

In his analysis of purification, Harsanyi fixes an arbitrary profile of smooth type distributions μ^p, and then uses these "fixed" distributions to define sequences of type distributions μ^p_ε parameterized by ε.[7] He then proves that for all sufficiently small

[7] In particular, μ^p_ε is defined by $\mu^p_\varepsilon(\varepsilon T^p) = \mu^p(T^p)$ for all measurable $T^p \subseteq \Theta^p$. Ellison and Fudenberg [5] also employ this parameterization.

values of ε and for almost every normal form game, all Nash equilibria of the normal form game have purifications in the approximating Bayesian game $BG(F, \mu_\varepsilon)$. In fact, if one allows oneself to tailor the type distributions to the game and equilibrium in question, then it is easy to show that every Nash equilibrium x^* of every normal form game can be *exactly* purified. For example, if we assume that the mass of agents biased in favor of each pure strategy $i \in S^p$ is equal to the mass placed on this pure strategy in x^*, then the Bayesian strategy in which each agent plays his favored strategy is a purification of x^*.[8]

The main reason for restricting attention to small perturbations is to prove that in almost all cases, purified versions of x^* exist independently of the choice of the "fixed" type distribution μ^p. In this paper, our goal is to determine the extent to which heterogeneity in preferences existing in real life settings can generate stable polymorphic equilibria. For this reason, none of the results to follow depend on the noise level ε being small. In fact, as we discuss below, our stability results are most robust when the noise level ε is not especially small.[9]

3. The Bayesian Best Response Dynamic and Stable Purification

In order to study the evolutionary stability of purified equilibria, we must introduce an evolutionary dynamic on Σ, the space of Bayesian strategy profiles. This role is filled by the Bayesian best response dynamic of Ely and Sandholm [6], an extension of the best response dynamic of Gilboa and Matsui [11] to Bayesian games with a continuum of types.

3.1 The Bayesian Best Response Dynamic

The *Bayesian best response dynamic* is defined on Σ by

$$\text{(B)} \qquad \dot{\sigma}^p = B^p(E\sigma) - \sigma^p .$$

Solutions to this dynamic are defined in terms of the L^1 norm on Σ^p:

$$\|\sigma^p\| \equiv \sum_{i=1}^{n^p} E|\sigma_i^p| .$$

[8] More precisely: For each population $p \in \mathcal{P}$ and each strategy $i \in S^p$, define the set of types $\Theta_i^p \subset \Theta^p$ by $\Theta_i^p = \{\theta^p\colon\ \theta_i^p > 0 \text{ and } \theta_j^p \leq 0 \text{ for } j \neq i\}$, and suppose that the measure μ^p satisfies $\mu^p(\Theta_i^p) = (x^*)_i^p$. Then it is easy to see that the Bayesian strategy σ in which agents of types in Θ_i^p play strategy $i \in S^p$ is a Bayesian equilibrium, and that this equilibrium exactly purifies x^*, in the sense that $E^\mu\sigma = x^*$.

[9] In not focusing exclusively on the small noise limit, we take the view of Ellison and Fudenberg [5] that "a substantial degree of heterogeneity is an important feature of real world learning" (p. 112).

Ely and Sandholm [6] show that solutions to (B) from every initial condition in Σ exist and are unique, and that the rest points of this dynamic are precisely the Bayesian equilibria of $BG(F, \mu)$.[10]

Since Bayesian strategy profiles are infinite dimensional objects, dynamics on the space Σ are difficult to analyze directly. To contend with this problem, Ely and Sandholm [6] establish close connections between the Bayesian best response dynamic (B) on the function space Σ and the *aggregate best response dynamic*

$$\text{(AB)} \qquad \dot{x}^p = E(B^p(x)) - x^p$$

on $X \subset \mathbf{R}^n$. In particular, they prove (*i*) that aggregate behavior x^* is a rest point of (AB) if and only if the Bayesian strategy $\sigma^* = B(x^*)$ is a rest point of (B), and hence a Bayesian equilibrium; (*ii*) that the dynamic (AB) describes the evolution of aggregate behavior $E\sigma \in X$ under the dynamic (B); and (*iii*) that x^* is stable under (AB) if and only if $B(x^*)$ is stable under (B), where "stable" can refer to Lyapunov, asymptotic, or global asymptotic stability.

3.2 The Perturbed Best Response Dynamic

Since the composition $E \circ B^p \colon X \to X^p$ used to define the dynamic (AB) involves an intermediate step through the function space Σ, computations involving $E \circ B$ can be cumbersome. Much of the analysis to follow relies on the fact that this composition can be expressed in a simpler way. Following Hofbauer and Sandholm [16], we define the *choice probability function* $C^p \colon \mathbf{R}^{n^p} \to X^p$ for the distribution μ^p by

$$C_i^p(\pi^p) = \mu^p\left(\theta^p : \, i \in \operatorname*{argmax}_{j \in S^p} \pi_j^p + \theta_j^p\right).$$

In words, $C_i^p(\pi^p)$ is the proportion of types in population p who find strategy i optimal when the common payoff vector is π^p. We then define the *perturbed best response function* $\tilde{B}^p \colon X \to X^p$ for the pair (F, μ) by the composition $\tilde{B}^p = C^p \circ F^p$. It follows from these definitions that

$$\begin{aligned} E(B_i^p(x)) &= \int_{\Theta^p} B_i^p(x)(\theta^p)\, d\mu^p(\theta^p) \\ &= \mu^p(\theta^p : B_i^p(x)(\theta^p) = 1) \end{aligned}$$

[10] Ely and Sandholm's [6] results are stated for a single population model, but they extend immediately to multipopulation settings.

$$= \mu^p(\theta^p : \arg\max_{k\in S^p} F_k^p(x) + \theta_k^p = i)$$
$$= \tilde{B}_i^p(x).$$

That is, the composition $E \circ B^p$ is identical to $\tilde{B}^p$. This observation immediately implies

Proposition 3.1: *The aggregate best response dynamic* (AB) *is equivalent to the* perturbed best response dynamic

(P) $$\dot{x}^p = \tilde{B}^p(x) - x^p.$$

The perturbed best response dynamic first appears in Fudenberg and Kreps's [8] model of stochastic fictitious play, where it arises as a description of the expected motion of the time average of play. Hopkins [20], Hofbauer [14], Hofbauer and Sandholm [17, 18] and Hofbauer and Hopkins [15] use techniques ranging from linearization and Lyapunov's direct method to the theory of monotone dynamical systems to study the stability of rest points of (P). In the remainder of this paper, we build on these analyses in order to examine the stability of purified equilibria under the Bayesian best response dynamic (B).

3.3 Stable Purification via Linearization

Suppose we are given a normal form game, a Nash equilibrium x^*, the type distributions μ^p, and a purified equilibrium σ. How can we determine whether σ is stable under the Bayesian best response dynamic (B)?

As noted above, stability of the Bayesian equilibrium σ under the dynamic (B) is equivalent to stability of the distribution $x = E^\mu\sigma$ under the perturbed best response dynamic (P). A standard method of determining local stability of rest points of ordinary differential equations like (P) is linearization. In the symmetric case, the dynamic (P) is defined on X by

(P1) $$\dot{x} = \tilde{B}(x) - x,$$

so the linearization of the dynamic can be expressed in terms of the derivative matrix $D\tilde{B}(x)$. In the standard case, the dynamic (P) is defined on $X = X^1 \times X^2$ by

(P2) $$\dot{x} = \begin{pmatrix} \tilde{B}^1(x^2) \\ \tilde{B}^2(x^1) \end{pmatrix} - \begin{pmatrix} x^1 \\ x^2 \end{pmatrix},$$

so the linearization involves the two derivative matrices $D\tilde{B}^1(x^2)$ and $D\tilde{B}^2(x^1)$.

These derivative matrices provide some information that is superfluous to our stability analyses. Focusing for the moment on the symmetric case, the derivative matrix $D\tilde{B}(x)$ tells us how the perturbed best response $\tilde{B}(x)$ changes as we move away from $x \in X$ in any direction $v \in \mathbf{R}^n$: by Taylor's formula, $\tilde{B}(x+v) = \tilde{B}(x) + D\tilde{B}(x)v + o(|v|)$. But since the population's mass is fixed at one, the only relevant displacement directions are those that leave the population's mass constant. These directions are contained in the set $\mathbf{R}^n_0 = \{z \in \mathbf{R}^n\colon \sum_{i \in S} z_i = 0\}$, the *tangent space* for the single population state space X. We sometimes write TX instead of $\mathbf{R}^n_0$ for emphasis. In the two population case, both populations' masses are fixed at one, so the tangent space for X is $TX = TX^1 \times TX^2 = \mathbf{R}^{n^1}_0 \times \mathbf{R}^{n^2}_0$.

In light of these observations, we define

$$\begin{aligned} \bar{\lambda} &= \max\{Re(\lambda)\colon \lambda \text{ is an eigenvalue of } D\tilde{B}(x) \text{ restricted to } \mathbf{R}^n_0\}; \text{ and} \\ \bar{\bar{\lambda}} &= \max\{Re(\sqrt{\lambda})\colon \lambda \text{ is an eigenvalue of } D\tilde{B}^1(x^2)\, D\tilde{B}^2(x^1) \text{ restricted to } \mathbf{R}^{n^1}_0\} \\ &= \max\{Re(\sqrt{\lambda})\colon \lambda \text{ is an eigenvalue of } D\tilde{B}^2(x^1)\, D\tilde{B}^1(x^2) \text{ restricted to } \mathbf{R}^{n^2}_0\}. \end{aligned}$$

Lemma 3.2 shows that except in knife edge cases, the values of these quantities determine the stability of purified equilibria.

Lemma 3.2: *Let σ be a purification of x^* with distribution $x = E\sigma$.*

(i) $(p = 1)$ If $\bar{\lambda} < 1$, then the purification σ is asymptotically stable; if $\bar{\lambda} > 1$ it is unstable.

(ii) $(p = 2)$ If $\bar{\bar{\lambda}} < 1$, then the purification σ is asymptotically stable; if $\bar{\bar{\lambda}} > 1$ it is unstable.

Proof: Theorems 5.5 and 6.4 of Ely and Sandholm [6] and Proposition 3.1 above tell us that to determine the stability of σ under the Bayesian dynamic (B), it is enough to determine the stability of x under the perturbed best response dynamics (P1) and (P2). We consider the two cases individually.

(*i*) By standard results, the rest point x is stable under the dynamic (P1) if the eigenvalue of its linearization with largest real part has a real part less than zero, and it is unstable if this real part is greater than zero. The linearization of (P1) is $D\tilde{B}(x) - I$; since this dynamic is defined on X, we are only concerned with the properties of this linearization on the tangent space $TX = \mathbf{R}^n_0$. Since $(\lambda - 1, z)$ is an eigenvalue/eigenvector pair for $D\tilde{B}(x) - I$ if and only if (λ, z) is an eigenvalue/eigenvector pair for $D\tilde{B}(x)$, the result follows immediately from the definition of $\bar{\lambda}$.

(*ii*) In this case, the relevant comparison is between 1 and the largest real part of an eigenvalue of $D\tilde{B}(x)$ restricted to $TX = \mathbf{R}^{n^1}_0 \times \mathbf{R}^{n^2}_0$. But Lemma A.1 in the Appendix implies that this largest real part is given by $\bar{\bar{\lambda}}$ (and also shows that the two definitions of $\bar{\bar{\lambda}}$ offered above are equivalent). This completes the proof of Lemma 3.2. ■

In the next three sections, we use Lemma 3.2 to ascertain the properties of games and type distributions that determine the local stability of purified equilibria.

4. Stability and Instability of Purified Equilibria in 2 x 2 Games

We first consider 2 x 2 games. Given the 2 x 2 payoff matrices A and B, define

$$\begin{aligned}
\delta_A(y) &= (A_{11} - A_{21})\, y_1 + (A_{12} - A_{22})\, y_2, \\
\delta'_A &= (A_{11} + A_{22}) - (A_{12} + A_{21}), \\
\delta_B(y) &= (B_{11} - B_{12})\, y_1 + (B_{21} - B_{22})\, y_2, \text{ and} \\
\delta'_B &= (B_{11} + B_{22}) - (B_{12} + B_{21}).
\end{aligned}$$

The function δ_A specifies the payoff of strategy 1 relative to that of strategy 2 under payoff matrix A given opponents' behavior $y = (y_1, y_2)$. The constant δ'_A describes the rate of change of this relative payoff as we simultaneously increase y_1 and decrease y_2. The function $\delta_B(y)$ and the constant δ'_B serve analogous roles for the payoff matrix B.

In the symmetric setting, δ'_A is positive when A is a coordination game, where increasing the use of strategy 1 makes this strategy relatively more attractive; δ'_A is negative when A is a Hawk-Dove game, where the reverse is true. In the standard setting, δ'_A and δ'_B have the same sign when (A, B) is a coordination game, with the sign determining whether coordination is on diagonal or off-diagonal strategies; δ'_A and δ'_B have different signs in games like Matching Pennies whose unique Nash equilibrium is in mixed strategies, with the signs determining the direction in which best responses cycle.

To state our stability result for 2 x 2 games, we require one additional definition: let $g^p : \mathbf{R} \to \mathbf{R}$ denote the density of the difference $d^p = \theta_2^p - \theta_1^p$ under the type distribution μ^p. As usual, we write g for g^1 when discussing the symmetric case.

Theorem 4.1: *Let σ be a purification of x^* with distribution $x = E\sigma$, and suppose that $n^1 = n^2 = 2$.*

(i) ($p = 1$) If $g(\delta_A(x))\,\delta'_A < 1$, then σ is stable; if $g(\delta_A(x))\,\delta'_A > 1$, σ is unstable.

(ii) ($p = 2$) If $g^1(\delta_A(x^2))\, g^2(\delta_B(x^1))\,\delta'_A\, \delta'_B < 1$, then σ is stable; if $g^1(\delta_A(x^2))\, g^2(\delta_B(x^1))\,\delta'_A\, \delta'_B > 1$, σ is unstable.

Proof: (*i*) By definition,

$$C_1(\pi) = \mu(\theta\colon \ \pi_1 + \theta_1 \geq \pi_2 + \theta_2) = \mu(\theta\colon \ \theta_2 - \theta_1 \leq \pi_1 + \pi_2) = \int_{-\infty}^{\pi_1 - \pi_2} g(\tau)\,d\tau, \text{ and}$$
$$C_1(\pi) + C_2(\pi) = 1.$$

Therefore, since $\tilde{B}(x) = (C \circ F)(x) = C(Ax)$, we can compute that

$$D\tilde{B}(x) = DC(Ax)\,A = g(\delta_A(x))\ \Xi\, A, \text{ where } \Xi = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

One can easily verify that the vector $(1,-1)'$, which spans $\mathbf{R}^2_0$, is an eigenvector of $D\tilde{B}(x)$ with eigenvalue $g(\delta_A(x))\,\delta'_A$. The conclusion therefore follows from Lemma 3.2(*i*).

(*ii*) Reasoning analogous to that used above shows that

$$D\tilde{B}^1(x^2)\, D\tilde{B}^2(x^1) = DC^1(Ax^2)\,A\,DC^2(B'x^1)\,B' = g^1(\delta_A(x^2))\,g^2(\delta_B(x^1))\Xi A \Xi B'.$$

The vector $(1,-1)'$ is an eigenvector of this matrix with eigenvalue λ = $g^1(\delta_A(x^2))\,g^2(\delta_B(x^1))\,\delta'_A\,\delta'_B$. Since λ is real, $\lambda - 1$ has the same sign as $Re(\sqrt{\lambda}) - 1$, so the conclusion follows from Lemma 3.2(*ii*). ■

In the symmetric setting, stability of purified equilibrium depends on the product $g(\delta_A(x))\,\delta'_A$, which is the lone relevant eigenvalue of the derivative matrix $D\tilde{B}(x)$. Here, $g(\delta_A(x))$ represents the density of the set of agents who are indifferent at the equilibrium, while δ'_A summarizes information about the game's incentive structure. In the standard setting, stability is determined by the value of $\lambda(x) = g^1(\delta_A(x^2))\,g^2(\delta_B(x^1))\,\delta'_A\,\delta'_B$.[11]

Let us focus for now on purified mixed equilibria of coordination games. In the symmetric setting, A is a coordination game when δ'_A is positive. Thus, Theorem 4.1(*i*) tells us that a purified equilibrium of A is stable if the density $g(\delta_A(x))$ is less than $(\delta'_A)^{-1}$, and that it is unstable if $g(\delta_A(x))$ is greater than $(\delta'_A)^{-1}$. In the standard setting, (A, B) is a coordination game when $\delta'_A\delta'_B$ is positive. Since $\lambda(x)$ contains the product of the two preference densities, stability in this case only requires the density of indifferent players in one of the two populations to be small.

We illustrate the stability criterion for the symmetric case with an example.[12] Suppose that a population of agents is randomly paired to play the coordination game

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

[11] This expression is also derived in the proof of Proposition 2 of Ellison and Fudenberg [5].

[12] The intuition behind the stability criterion for the standard case is somewhat more complicated. It is described in our working paper, Sandholm [26].

This game has two pure equilibria and the mixed equilibrium $x^* = (\frac{1}{2}, \frac{1}{2})$. Without diversity in preferences, x^* is unstable under any reasonable evolutionary dynamic.

To introduce diversity in preferences, let us first suppose that g, the density of the bias in favor of strategy 2, is uniform on $[-\frac{1}{2}, \frac{1}{2}]$:

$$g(d) = \begin{cases} 1 & \text{if } d \in [-\frac{1}{2}, \frac{1}{2}], \\ 0 & \text{otherwise.} \end{cases}$$

The top part of Figure 1 presents the perturbed best response function $\tilde{B}$ induced by g. (In particular, it graphs $\tilde{B}_1$ as a function of x_1, under the assumption that $x_2 \equiv 1 - x_1$.) Since all agents prefer strategy 2 whenever $x_1 < \frac{1}{4}$, $\tilde{B}_1(x_1) = 0$ at such states; similarly, $\tilde{B}_1(x_1) = 1$ whenever $x_1 > \frac{3}{4}$. The uniformity of g implies that as x_1 varies from $\frac{1}{4}$ to $\frac{3}{4}$, the proportion of agents who prefer strategy 1 increases linearly from 0 to 1. The slope of $\tilde{B}_1$ in this middle region, $\tilde{B}_1'(x_1) = 2 = g(\delta_A(x))\,\delta_A'$, is the relevant eigenvalue of the derivative matrix $DB(x^*)$ from the proof of Theorem 4.1(i).

The bottom part of Figure 1 presents the perturbed best response dynamic (P), here described by $\dot{x}_1 = \tilde{B}_1(x_1) - x_1$. At states to the left of x_1^*, $\dot{x}_1$ is negative, so motion is leftward toward $x_1 = 0$, while at states to the right of x_1^*, motion is rightward toward $x_1 = 1$; the rest point x_1^* is therefore unstable. We can also reach this conclusion via the eigenvalue analysis from Theorem 4.1(i): since $\frac{d}{dx_1}(\tilde{B}_1(x_1) - x_1)\big|_{x_1 = x_1^*} = \tilde{B}_1'(x_1^*) - 1 = 1 > 0$, x_1^* is unstable. We will generalize this approach in Theorem 5.1, where we offer sufficient conditions for instability of purified equilibria for games with many strategies.

Next, let us perform the same exercise for a type distribution with density

$$g(d) = \begin{cases} \frac{7}{4} & \text{if } d \in [-\frac{1}{2}, -\frac{1}{4}] \cup [\frac{1}{4}, \frac{1}{2}], \\ \frac{1}{4} & \text{if } d \in (-\frac{1}{4}, \frac{1}{4}), \\ 0 & \text{otherwise.} \end{cases}$$

In this case, as x_1 varies from 0 to 1, the slope of $\tilde{B}_1$ follows the sequence $\{0, \frac{7}{2}, \frac{1}{2}, \frac{7}{2}, 0\}$. As Figure 2 shows, the perturbed best response dynamic has five rest points; this time, the rest point $x_1^* = \frac{1}{2}$ is stable. Once again, stability can also be determined by an analysis of eigenvalues: the slope of $\dot{x}_1$ at x_1^* is given by $\frac{d}{dx_1}(\tilde{B}_1(x_1) - x_1)\big|_{x_1 = x_1^*} = \tilde{B}_1'(x_1^*) - 1 = \frac{1}{2} - 1 = -\frac{1}{2} < 0$, implying stability. In Theorem 6.1, we extend this stability analysis via linearization to games with arbitrary numbers of strategies.

While linearization can tell us whether an equilibrium is locally stable, it says nothing about the size of a stable equilibrium's basin of attraction. For instance, it

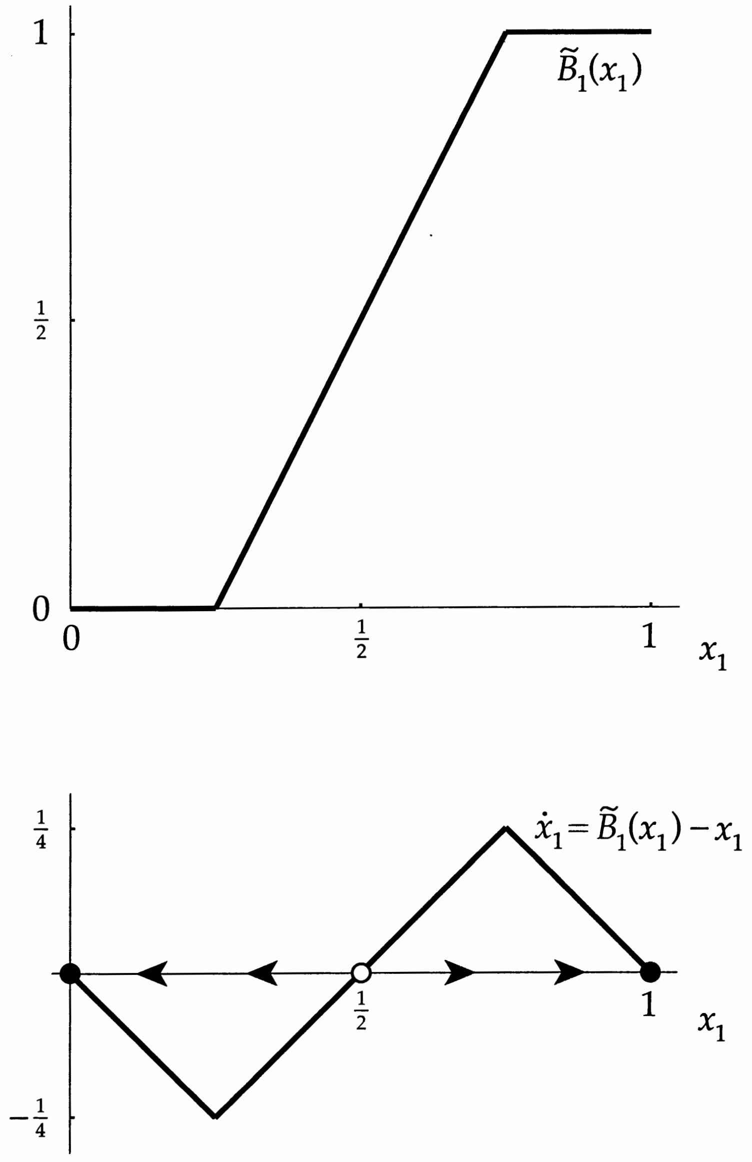

Figure 1

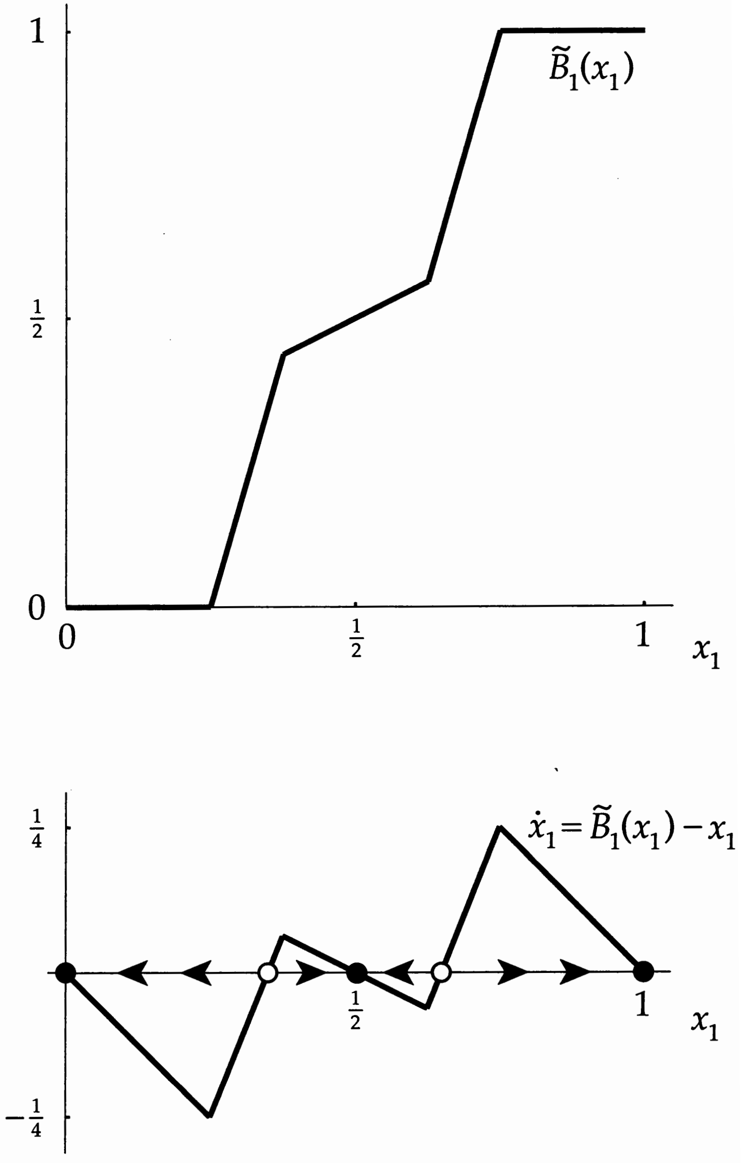

Figure 2

follows easily from the foregoing analysis (especially Figure 2) that if g is symmetric, satisfies $g(0) < (\delta'_A)^{-1} = \frac{1}{2}$, and has support $[-\varepsilon, \varepsilon]$, then the rest point $x_1^* = \frac{1}{2}$ is stable, with a basin of attraction contained in the interval $[x_1^* - \frac{\varepsilon}{2}, x_1^* + \frac{\varepsilon}{2}]$. Therefore, smaller perturbations lead to smaller basins of attraction.

In our view, the relevant scale for payoff perturbations depends on the application in question. Whether the perturbations are large enough relative to the likely behavior disturbances for a purified equilibrium to be considered stable is thus contingent upon the context at hand.

These analyses of coordination games might create the impression that introducing the right payoff perturbations can enable one to reverse the stability of any mixed equilibrium. In fact, while we can always create stable equilibria from unstable ones, we cannot always do the opposite. Theorem 4.1(*i*) shows that under symmetric matching, purified equilibria of Hawk-Dove games are always stable: since δ'_A is negative in these games, the eigenvalue $g(\delta_A(x))\delta'_A$ is negative as well, implying local stability. Similarly, Theorem 4.1(*ii*) implies that under standard matching, purified equilibria are stable in all games satisfying $\delta'_A\, \delta'_B < 0$: that is, whenever best responses cycle. Theorem 6.3 offers a substantial generalization of these results: it exhibits classes of games with arbitrary numbers of strategies in which purified equilibria are always *globally* stable, regardless of the distributions of types.

5. Sufficient Conditions for Instability of Purified Equilibria

In the stability analyses of 2 x 2 games, the crucial pieces of data from the preference distributions are the densities of indifferent agents, as it is the indifferent agents who are the first to react to small disturbances in behavior. When each individual has only two strategies, the relevant density is simply the value of the function $g^p: \mathbf{R} \to \mathbf{R}_+$ evaluated at the equilibrium payoff. But when players can choose among more than two strategies, the set of types who are currently indifferent between a given pair of strategies forms a multidimensional subspace of $\mathbf{R}^n$, with the relevant subspace depending on which pair of strategies is considered. To extend our results to games with many strategies, we must find more general ways of describing the density of indifferent agents.

As a step in this direction, recall from Section 3.2 the definition of the choice probability function $C^p: \mathbf{R}^{n^p} \to X^p$:

$$C_i^p(\pi^p) = \mu^p\left(\theta^p : \; i \in \operatorname*{argmax}_{j \in S^p} \pi_j^p + \theta_j^p\right).$$

$C^p_i(\pi^p)$ is the mass of agents who find strategy $i \in S^p$ optimal when the common payoff vector is π^p. Consequently, the density of indifferent agents is captured by the sensitivity of $C^p(\pi^p)$ to slight changes in π^p. This sensitivity is captured in turn by the derivative matrix $DC^p(\pi^p)$: $\mathbf{R}^{n^p} \to TX^p = \mathbf{R}^{n^p}_0$. For any small perturbation $v^p \in \mathbf{R}^{n^p}$, the vector $DC^p(\pi^p)\, v^p \in \mathbf{R}^{n^p}_0$ describes the changes in the probabilities with which the strategies in S^p are chosen when payoffs shift from π^p to $\pi^p + v^p$: by Taylor's formula, $C^p(\pi^p + v^p) = C^p(\pi^p) + DC^p(\pi^p)v^p + o(|v^p|)$.

We obtain scalar measures of the sensitivity of choice probabilities to changes in payoffs by considering the eigenvalues of $DC^p(\pi^p)$.[13] Define

$$\bar{\lambda}_{DC^p(\pi^p)} = \max\{\lambda\colon\ \lambda \text{ is an eigenvalue of } DC^p(\pi^p)\} \text{ and}$$

$$\underline{\lambda}_{DC^p(\pi^p)} = \min\{\lambda\colon\ \lambda \text{ is an eigenvalue of } DC^p(\pi^p) \text{ restricted to } \mathbf{R}^{n^p}_0\}.$$

It is known that the matrix $DC^p(\pi^p)$ is symmetric and positive semidefinite, and hence that its eigenvalues are real and weakly positive. It follows that the maximal eigenvalue of $DC^p(\pi^p)$ provides a measure of the maximal sensitivity of choice probabilities to changes in payoffs. To understand the measure of minimal sensitivity, bear in mind that increasing each strategy's payoff by the same amount does not alter choice probabilities. This observation is expressed in terms of the derivative matrix as $DC^p(\pi^p)\mathbf{1} = \mathbf{0}$, where $\mathbf{1} = (1,\ldots,1)'$ and $\mathbf{0} = (0,\ldots,0)'$. To obtain our measure of minimal sensitivity to *consequential* changes in payoffs, we restrict attention to payoff changes in the set $\mathbf{R}^{n^p}_0$, which contains all vectors orthogonal to the vector $\mathbf{1}$. If μ^p has full support on $\mathbf{R}^{n^p}$, one can show that $DC^p(\pi^p)$ is actually positive *definite* with respect to $\mathbf{R}^{n^p}_0$, which implies that $\underline{\lambda}_{DC^p(\pi^p)}$ is strictly greater than zero.

The statement of our general instability result requires a few additional definitions. If $M \in \mathbf{R}^{n^p \times n^p}$ is a square matrix, we let $S(M) = \frac{1}{2}(M + M')$ denote its symmetric part. If the matrix $N \in \mathbf{R}^{n^p \times n^p}$ is symmetric and maps $\mathbf{R}^{n^p}_0$ into itself, we define

$$\underline{\lambda}_N = \min\{\lambda\colon\ \lambda \text{ is an eigenvalue of } N \text{ restricted to } \mathbf{R}^{n^p}_0\}.$$

Finally, we let $\Phi = I - \frac{1}{n^p}\mathbf{1}\mathbf{1}' \in \mathbf{R}^{n^p \times n^p}$ denote the orthogonal projection of $\mathbf{R}^{n^p}$ onto $\mathbf{R}^{n^p}_0$.

Theorem 5.1: *Let σ be a purification of $x^* \in int(X)$ with distribution $x = E\sigma$.*

(*i*) (p = 1) *Suppose that A is positive definite on $\mathbf{R}^n_0$. If $\underline{\lambda}_{DC(Ax)}\, \underline{\lambda}_{S(\Phi A)} > 1$, then the purification σ is a repeller: all nearby trajectories move away from σ. Suppose in addition that*

[13] For analyses of the properties of the derivative matrix $DC^p(\pi^p)$, see Anderson, de Palma, and Thisse [1] or Hofbauer and Sandholm [16].

A is symmetric; then if $\bar{\lambda}_{DC(Ax)}\,\underline{\lambda}_{\Phi A \Phi} > 1$*, the purification* σ *is unstable.*

(*ii*) ($p = 2$) *Suppose that* (*A*, *B*) *is a potential game* ($B = A$) *and that* x^* *is isolated. If either* $\underline{\lambda}_{DC^1(Ax^2)}\,\bar{\lambda}_{DC^2(A'x^1)}\,\underline{\lambda}_{\Phi A \Phi A' \Phi} > 1$ *or* $\bar{\lambda}_{DC^1(Ax^2)}\,\underline{\lambda}_{DC^2(A'x^1)}\,\underline{\lambda}_{\Phi A' \Phi A \Phi} > 1$*, then the purification* σ *is unstable.*

The proof of this result is provided in the Appendix. As an aside, we note that our analysis relies on a simple transformation that allows calculations involving linear and bilinear operators on $\mathbf{R}^n_0$ to be performed using operators on the more convenient space $\mathbf{R}^{n-1}$ (Lemmas A.2 and A.3). This transformation may prove useful in analyses of other evolutionary dynamics.

In the symmetric setting ($p = 1$), if agents are randomly matched to play a game with a symmetric positive definite payoff matrix, the resulting population game is a potential game with a strictly convex potential function. Any interior Nash equilibrium is a local minimizer of this function; since all reasonable evolutionary processes increase potential, interior Nash equilibria are dynamically unstable under a wide range of evolutionary dynamics.[14] The second statement in Theorem 5.1(*i*) shows that as long as a single eigenvalue of $DC(Ax)$ is sufficiently large—in other words, as long as choice probabilities are sensitive to one direction of payoff change—the purification of any interior Nash equilibrium is unstable as well.

If A is positive definite but not symmetric, interior Nash equilibria are unstable under a more restrictive class of evolutionary dynamics.[15] Under this weaker restriction on the payoff matrix, part (*i*) of the theorem shows that the purified equilibrium σ is unstable if *all* relevant eigenvalues of $DC(Ax)$ are sufficiently large. However, this stronger condition on $DC(Ax)$ implies that σ is a source: all solutions to (B) starting near σ leave the vicinity of σ.

In the standard setting ($p = 2$), random matching in a normal form potential game (A, A) also generates a potential game. In this case, isolated interior equilibria are saddle points of the population game's potential function, and so are unstable under a wide range of evolutionary dynamics.[16] In this case, as long as there is *one* population whose choice probabilities are sensitive to *some* change in payoffs, the purified equilibrium is unstable.

[14] The potential function f for a potential game F satisfies $\nabla f \equiv F$. Since in the present case $F(x) = Ax$ for a symmetric matrix A, $f(x) = \frac{1}{2} x \cdot Ax$; since A is positive definite, f is strictly convex. For derivations of the remaining claims stated above, see Sandholm [24].

[15] See Hopkins [20].

[16] In this case, the potential function for F is $f(x) = x^1 \cdot Ax^2$. The saddle point property is proved in the Appendix (Proposition A.7). Theorem 5.1(*ii*) extends immediately to any normal form potential game as defined by Monderer and Shapley [23].

6. Sufficient Conditions for Stability of Purified Equilibria

To state our first sufficient conditions for stability, we recall a definition from matrix analysis. The scalar $\lambda \in \mathbf{R}_+$ is a *singular value* of $M \in \mathbf{R}^{k \times l}$ if λ is a nonnegative square root of an eigenvalue of $MM' \in \mathbf{R}^{k \times k}$ (if $k \leq l$) or of $M'M \in \mathbf{R}^{l \times l}$ (if $l < k$). Let $\bar{s}_M$ denote the largest singular value of M. Our analysis relies on the following three properties of singular values:[17]

(S1) If $M \in \mathbf{R}^{k \times l}$ and $N \in \mathbf{R}^{l \times m}$, then $\bar{s}_{MN} \leq \bar{s}_M \, \bar{s}_N$.

(S2) If $M \in \mathbf{R}^{m \times m}$, and if $\tilde{\lambda}_M \in \mathbf{C}$ is the eigenvalue of M with largest modulus, then $\left|\tilde{\lambda}_M\right| \leq \bar{s}_M$.

(S3) If $M \in \mathbf{R}^{m \times m}$ is symmetric, then the singular values of M are the absolute values of the eigenvalues of M.

Theorem 6.1 provides sufficient conditions for stability of purified equilibria.

Theorem 6.1: *Let σ be a purification of x^* with distribution $x = E\sigma$.*
(i) ($p = 1$) If $\bar{\lambda}_{DC(Ax)} \, \bar{s}_{\Phi A \Phi} < 1$, then the purification σ is stable.
(ii) ($p = 2$) If $\bar{\lambda}_{DC^1(Ax^2)} \, \bar{\lambda}_{DC^2(B'x^1)} \, \bar{s}_{\Phi A \Phi} \, \bar{s}_{\Phi B \Phi} < 1$, then the purification σ is stable.

Since $\bar{s}_{\Phi A \Phi} \leq \bar{s}_{\Phi} \, \bar{s}_A \, \bar{s}_{\Phi} = \bar{s}_A$ and $\bar{s}_{\Phi B \Phi} \leq \bar{s}_{\Phi} \, \bar{s}_B \, \bar{s}_{\Phi} = \bar{s}_B$ by properties (S1) and (S3), the expressions $\bar{s}_{\Phi A \Phi}$ and $\bar{s}_{\Phi B \Phi}$ in the statement of the theorem can be replaced with $\bar{s}_A$ and $\bar{s}_B$. Of course, doing so makes our sufficient conditions for stability more demanding.

The proof of Theorem 6.1 can be found in the Appendix. It follows immediately from this theorem that stable purifications always exist.

Corollary 6.2: *For each $\varepsilon > 0$, every Nash equilibrium of every two player normal form game admits a stable ε-purification. In general, the size of the basin of attraction of the stable purified equilibrium depends on the value of ε.*

Proof: Consider an exact purification of the form described in Section 2.4, in which measure $(x^*)_i^p$ agents are biased toward strategy i, and impose the additional restriction that none of these agents' biases towards i are less than $\frac{\varepsilon}{2}$. Then when evaluated at the equilibrium payoffs, DC^1 and DC^2 both equal the zero matrix. The result therefore

[17] Properties (S1) and (S2) follow from Theorems 3.3.4 and 3.3.2 of Horn and Johnson [22], while property (S3) follows easily from the definition of a singular value. Singular values are best known for their role in the *singular value decomposition*: every real matrix M can be written as $V\Sigma W'$, where V and W are orthogonal, the off-diagonal elements of Σ are zero, and the diagonal elements of Σ are the singular values of A. For further discussion, see Horn and Johnson [22, Chapter 3].

follows from Theorem 6.1. ■

The sufficient conditions for stable purification in Theorem 6.1 are stated in terms of the maximal sensitivities of choice probabilities to changes in payoffs and the maximal singular values of the payoff matrices. Like Theorem 4.1 for 2 x 2 games, Theorem 6.1 guarantees stability whenever the maximal sensitivity of choice probabilities in the lone population (when p = 1) or in one of the two populations (when p = 2) is sufficiently small. Since it is possible to adjust the distributions of types to make these sensitivities as small as desired, we conclude in Corollary 6.2 that any Nash equilibrium of any normal form game can be purified in a stable fashion. Note that in contrast with our instability result (Theorem 5.1), these stability results do not require the normal form game in question to be drawn from a particular class of "well behaved" games.

Unlike Theorem 4.1 for 2 x 2 games, Theorem 6.1 does not enable us to find classes of games whose purified equilibria are stable regardless of the distribution of types. In the 2 x 2 case, we were able to reach such conclusions because the statistics $\delta'_A \in \mathbf{R}$ and $\delta'_B \in \mathbf{R}$ capture both the magnitude and the "direction" of the effects of changes in opponents' behavior on own payoffs. For certain specifications of these "directions", our sufficient conditions for stability were satisfied by default. In the present case, the singular values $\bar{s}_A \in \mathbf{R}_+$ and $\bar{s}_B \in \mathbf{R}_+$ only describe the magnitudes of payoff effects, so distribution-free stability results cannot be obtained in a similar fashion. However, by relying on Lyapunov functions rather than linearization as the basis for our analysis, we can obtain the following stability result. Its proof can be found in the Appendix.

Theorem 6.3: *Let σ be a purification of x^* with distribution $x = E\sigma$, where each type distribution μ^p is smooth and has full support on $\mathbf{R}^{n^p}$.*

(*i*) ($p = 1$) *If A is negative semidefinite with respect to $\mathbf{R}^n_0$, then σ is globally asymptotically stable. In particular, these statements are true if A admits an interior ESS or an interior NSS, or if A is symmetric zero sum ($A = -A'$).*

(*ii*) ($p = 2$) *If (A, B) is zero sum ($B = -A$), then σ is globally asymptotically stable.*

7. An Example

As an illustration of the use of our techniques, we consider the stability of purified equilibria in Rock-Paper-Scissors games. Consider a symmetric two player normal form game with payoff matrix

$$A = \begin{pmatrix} 0 & -l & w \\ w & 0 & -l \\ -l & w & 0 \end{pmatrix}.$$

Here, w (the benefit from winning a match) and l (the cost of losing) are both positive. We distinguish three cases: $w > l$, which we call *good RPS;* $w = l$, which we call *standard RPS;* and $w < l$, which we call *bad RPS.*

In all three cases, A has a unique symmetric Nash equilibrium, $x^* = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, whose basic stability properties are well known.[18] State x^* is an ESS in good RPS, an NSS in standard RPS, and neither of these in bad RPS. In the first two cases, x^* is globally asymptotically stable under the best response dynamic of Gilboa and Matsui [11]; in the last, the best response dynamic approaches a limit cycle from almost all initial conditions.

What are the stability properties of purified versions of x^*? If the type distribution μ is smooth and has full support, then Theorem 6.3 tells us that in both good RPS and standard RPS, there is a unique purified equilibrium, an equilibrium that is globally asymptotically stable under the Bayesian best response dynamic. In bad RPS, purified equilibria can be either stable or unstable depending on the choice of type distribution. To employ our results from Sections 5 and 6, we compute

$$\Phi A = \frac{1}{3}\begin{pmatrix} l-w & -2l-w & l+2w \\ l+2w & l-w & -2l-w \\ -2l-w & l+2w & l-w \end{pmatrix} = \Phi A \Phi \text{ and } S(\Phi A) = \frac{l-w}{6}\begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}.$$

The latter matrix has an eigenvalue of 0 for direction $\mathbf{1}$ and two eigenvalues of $\frac{l-w}{2}$ for directions in $TX = \mathbf{R}^3_0$. Thus, $\underline{\lambda}_{S(\Phi A)} = \frac{l-w}{2}$, and so Theorem 5.1 tells us that a purified equilibrium σ with $E\sigma = x$ is unstable (in fact, a source) under the Bayesian best response dynamic whenever $\underline{\lambda}_{DC(Ax)} > \frac{2}{l-w}$. On the other hand, the matrix $\Phi A \Phi$ has two singular values of $\sqrt{l^2 + lw + w^2}$ and one of 0, so Theorem 6.1 implies that σ is stable whenever $\bar{\lambda}_{DC(Ax)} < 1/\sqrt{l^2 + lw + w^2}$.

To take the analysis one step further, we focus on a convenient family of type distributions μ. Suppose that under μ, the type components θ_i are i.i.d. with the extreme value distribution $P(\theta_i \leq b) = \exp(-\exp(-\eta^{-1} b - \gamma))$, where $\eta > 0$ is a parameter called the *noise level,* and $\gamma \approx .5772$ is Euler's constant. Then choice probabilities are given by the *logit choice rule,*

[18] See Gaunersdorfer and Hofbauer [10] and Weibull [27].

$$C_i(\pi) = \frac{\exp(\eta^{-1}\pi_i)}{\sum_k \exp(\eta^{-1}\pi_k)},$$

and the perturbed best response dynamic (P) becomes the well known *logit dynamic,* introduced by Fudenberg and Levine [9]:

$$\dot{x}_i = \frac{\exp(\eta^{-1}(Ax)_i)}{\sum_k \exp(\eta^{-1}(Ax)_k)} - x_i.$$

To ease the interpretation of the results to come, we note that if θ_i follows the extreme value distribution defined above, then $E(\theta_i) = 0$ and $SD(\theta_i) = \pi\eta / \sqrt{6} \approx 1.2826\ \eta$.[19]

By symmetry, these i.i.d. type distributions generate exact purification: for each $\eta > 0$, there is a purified equilibrium σ with $E\sigma = x^* = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Since x^* is a rest point of (P), we know that $C(Ax^*) = x^*$. Using this fact, we can compute $DC(Ax^*)$:

$$\begin{aligned} DC(Ax^*) &= \eta^{-1}\left(\text{diag}(C(Ax^*)) - C(Ax^*)C(Ax^*)'\right) \\ &= \frac{\eta^{-1}}{9}\begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}. \end{aligned}$$

This matrix has an eigenvalue of 0 for direction **1** and two eigenvalues of $\frac{\eta^{-1}}{3}$ for directions in $\mathbf{R}^3_0$. Therefore, Theorem 5.1 ensures that the purification σ is unstable if the noise level η is less than $\frac{l-w}{6}$, while Theorem 6.1 ensures that it is stable if η exceeds $\frac{1}{3}\sqrt{l^2 + lw + w^2}$.

Since we have specified the type distribution explicitly, we can check the stability of the purification directly using Lemma 3.2. Applying the analysis of the matrix $DC(Ax^*)$ above, we find that

$$DC(Ax^*)A = DC(Ax^*)\Phi A = \tfrac{\eta^{-1}}{3}\Phi A.$$

This matrix has an eigenvalue of 0 for direction **1** and eigenvalues of $\frac{\eta^{-1}}{6}\left((l-w) \pm \mathrm{i}\sqrt{3}(l+w)\right)$ for directions in $\mathbf{R}^3_0$. Therefore, Lemma 3.2 tells us that the purification σ is stable if $\eta > \frac{l-w}{6}$ and is unstable if $\eta < \frac{l-w}{6}$. We thus conclude that while the stability bound we obtained using Theorem 6.1 is loose, the instability bound obtained from Theorem 5.1 is tight.

We illustrate these results in Figures 3 and 4, which present phase diagrams of the

[19] See Anderson, de Palma, and Thisse [1] and Hofbauer and Sandholm [16].

logit dynamic in bad RPS with $w = 1$ and $l = 2$. In Figure 3, we set the noise level to $\eta = .1$; evidently, the purified equilibrium is unstable, with all solution trajectories other than the one at x^* appearing to converge to a limit cycle. In Figure 4, we raise the noise level to $\eta = .2$; in this case, the purified equilibrium becomes a global attractor. It is noteworthy that in this example, perturbations with a standard deviation as low as $\pi / 6\sqrt{6} \approx .2138$ are enough to create a stable purified equilibrium.

Appendix

We begin with a preliminary result used to prove Lemma 3.2(*ii*) (with $Q = D\tilde{B}^1(x^2)$ and $R = D\tilde{B}^2(x^1)$). Similar results can be found, for example, in Hofbauer and Sigmund [19, Section 17.4].

Lemma A.1 *Consider the matrix*

$$P = \begin{pmatrix} 0 & Q \\ R & 0 \end{pmatrix} \in \mathbf{R}^{(n^1+n^2)\times(n^1+n^2)},$$

where Q has range $\mathbf{R}_0^{n^1}$ and R has range $\mathbf{R}_0^{n^2}$. If λ is an eigenvalue of P restricted to $\mathbf{R}_0^{n^1} \times \mathbf{R}_0^{n^2}$, then λ^2 is an eigenvalue of QR restricted to $\mathbf{R}_0^{n^1}$, and is also an eigenvalue of RQ restricted to $\mathbf{R}_0^{n^2}$. If $\lambda \neq 0$, each converse implication also holds. Hence,

$$\begin{aligned} &\max\{Re(\lambda)\text{: } \lambda \textit{ is an eigenvalue of } P \textit{ restricted to } \mathbf{R}_0^{n^1} \times \mathbf{R}_0^{n^2}\} \\ &\quad = \max\{Re(\sqrt{\lambda})\text{: } \lambda \textit{ is an eigenvalue of } QR \textit{ restricted to } \mathbf{R}_0^{n^1}\} \\ &\quad = \max\{Re(\sqrt{\lambda})\text{: } \lambda \textit{ is an eigenvalue of } RQ \textit{ restricted to } \mathbf{R}_0^{n^2}\}. \end{aligned}$$

Proof: To prove the first claim, suppose that $Px = \lambda x$ for some $x \in \mathbf{C}_0^{n^1} \times \mathbf{C}_0^{n^2}$ and some $\lambda \in \mathbf{C}$. Evaluating this equality yields $Qx^2 = \lambda x^1$ and $Rx^1 = \lambda x^2$. Hence, $QRx^1 = Q\lambda x^2 = \lambda^2 x^1$ and $RQx^2 = R\lambda x^1 = \lambda^2 x^2$. To prove the first converse implication, suppose that $QRx^1 = \lambda^2 x^1$ for some $x^1 \in \mathbf{C}_0^{n^1}$ and some $\lambda \in \mathbf{C} - \{0\}$. If we set $x^2 = \lambda^{-1} Rx^1 \in \mathbf{C}_0^{n^2}$, then $Rx^1 = \lambda x^2$ and $Qx^2 = \lambda^{-1} QRx^1 = \lambda^{-1}(\lambda^2 x^1) = \lambda x^1$, so $Px = \lambda x$. The proof for the claim concerning RQ is similar.

We now use the first set of claims to establish the two equalities. If all relevant eigenvalues of P all have zero real part, then all relevant eigenvalues of QR and RQ are real and nonpositive, so the equalities follow from the first set of claims.

On the other hand, suppose there exists a relevant eigenvalue of P with nonzero real part. Then there exists one, say λ^*, with strictly positive real part: if $\lambda \neq 0$ with $Re(\lambda) < 0$ is a relevant eigenvalue of P, then λ^2 is a relevant eigenvalue of QR, and hence $-\lambda$ is

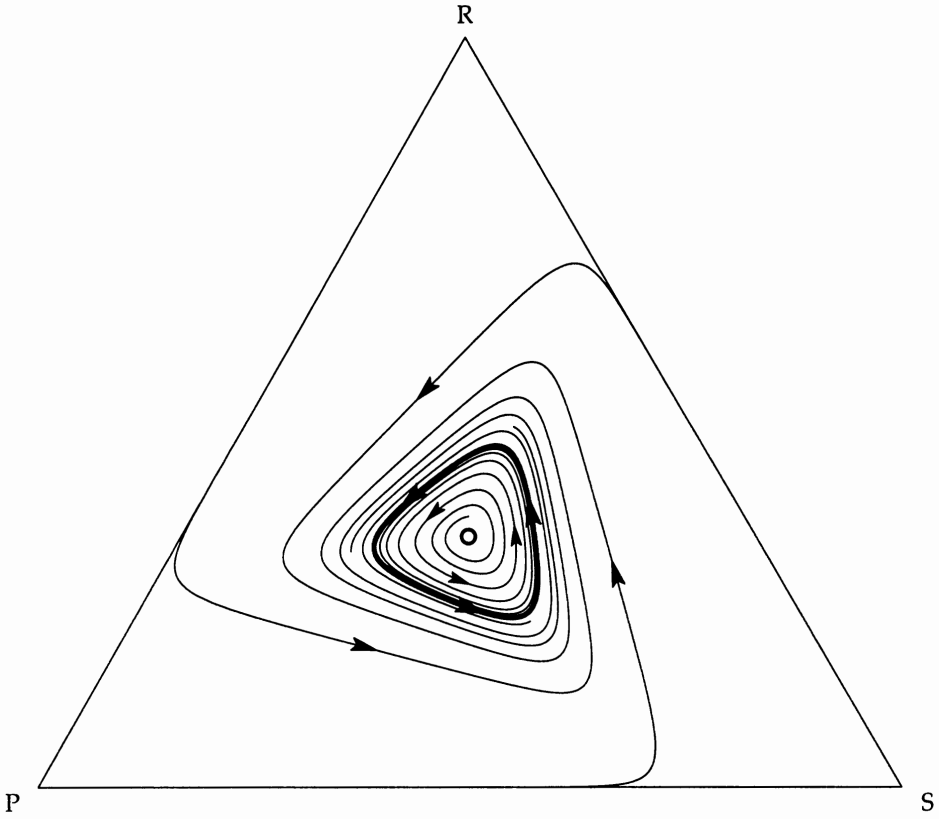

Figure 3: The logit(.1) dynamic in bad RPS

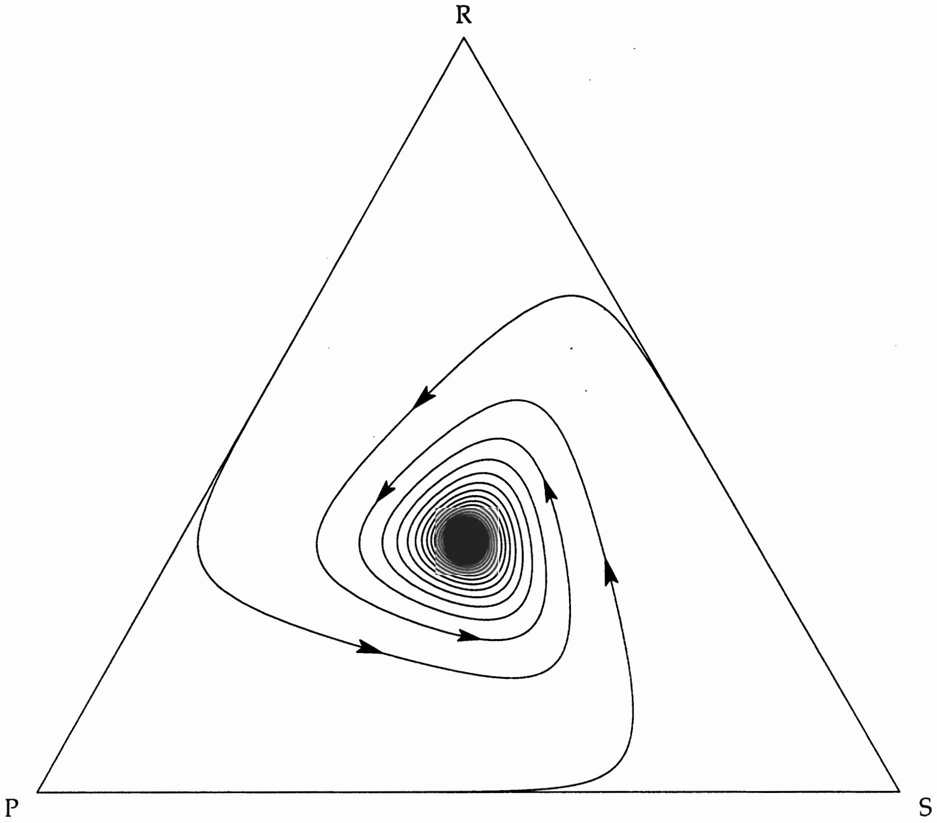

Figure 4: The logit(.2) dynamic in bad RPS

also a relevant eigenvalue of P (since $(-\lambda)^2 = \lambda^2$ is an eigenvalue of QR). Furthermore, every eigenvalue of QR and RQ is either the square of an eigenvalue of P or is zero. Since $Re(\lambda^*) > 0$, a zero eigenvalue cannot determine max $Re(\sqrt{\lambda})$ for QR or RQ, and so the equalities once again hold. ■

We now introduce a transformation that enables us to convert analyses of operators on $\mathbf{R}_0^n$ into analyses of operators on the more convenient space $\mathbf{R}^{n-1}$. Define the matrix $R \in \mathbf{R}^{n\times n}$ by

$$R = \begin{pmatrix} \frac{\sqrt{n}+n(n-2)}{n(n-1)} & \frac{\sqrt{n}-n}{n(n-1)} & \cdots & \frac{\sqrt{n}-n}{n(n-1)} & -\frac{1}{\sqrt{n}} \\ \frac{\sqrt{n}-n}{n(n-1)} & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \frac{\sqrt{n}-n}{n(n-1)} & \vdots \\ \frac{\sqrt{n}-n}{n(n-1)} & \cdots & \frac{\sqrt{n}-n}{n(n-1)} & \frac{\sqrt{n}+n(n-2)}{n(n-1)} & -\frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{n}} & \cdots & \cdots & \cdots & \frac{1}{\sqrt{n}} \end{pmatrix}.$$

R rotates span$\{\mathbf{1}, e_n\}$ about its orthogonal complement by an angle of $\cos^{-1}(\frac{1}{\sqrt{n}})$. (Here e_n is the nth standard basis vector in $\mathbf{R}^n$.) It is easy to verify that $R\mathbf{1} = \sqrt{n}\, e_n$ and that $R\,\mathbf{R}_0^n = \mathbf{R}_\dagger^n$, where $\mathbf{R}_\dagger^n = \{y \in \mathbf{R}^n\colon\ y_n = 0\}$. Since R is a rotation matrix, it is orthogonal, and hence $R' = R^{-1}$. (For more on rotation matrices, see Friedberg, Insel, and Spence [7, Section 6.10].)

Next, let $J \in \mathbf{R}^{(n-1)\times n}$ be defined by $J = (I\ \ \mathbf{0})$. Premultiplication of $y \in \mathbf{R}^n$ by J drops y's last component. It is easy to see that $J'J$ is the orthogonal projection of $\mathbf{R}^n$ onto $\mathbf{R}_\dagger^n$, and that $JJ' = I$ is the identity on $\mathbf{R}^{n-1}$.

Finally, define the matrix $Z \in \mathbf{R}^{(n-1)\times n}$ by $Z = JR$, so that Z is the matrix R with its final row removed. The following lemmas describe the properties of Z that make it useful in linearization analysis.

Lemma A.2: (*i*) *Z defines a bijective map from $\mathbf{R}_0^n$ to $\mathbf{R}^{n-1}$.*

(*ii*) *$Z'Z = R'J'JR = \Phi$, the orthogonal projection of $\mathbf{R}^n$ onto $\mathbf{R}_0^n$.*

(*iii*) *$ZZ' = JRR'J' = JJ' = I$, the identity on $\mathbf{R}^{n-1}$.*

Lemma A.3: *Suppose that $M \in \mathbf{R}^{n\times n}$ maps $\mathbf{R}_0^n$ into itself, and let $x \in \mathbf{R}_0^n$ (or $\mathbf{C}_0^n$). Then (λ, x) is an eigenvalue/eigenvector pair for M if and only if (λ, Zx) is an eigenvalue/eigenvector pair for ZMZ'.*

Proofs: The claims in Lemma A.2 are geometrically obvious and easy to verify by direct calculation; we use them to prove Lemma A.3. First, observe that if $Mx = \lambda x$ for some $x \in \mathbf{R}^n_0$, then $M\Phi x = \lambda\Phi x$, which implies that $ZM\Phi x = \lambda Z\Phi x$; then statements (*ii*) and (*iii*) above allow us to conclude that $ZMZ'Zx = \lambda ZZ'Zx = \lambda Zx$. Conversely, if $ZMZ'Zx = \lambda Zx$ for some $x \in \mathbf{R}^n_0$, then statement (*ii*) tells us that $ZMx = \lambda Zx$. Since Mx and x are both in $\mathbf{R}^n_0$, we conclude from statement (*i*) that $Mx = \lambda x$. ■

In addition to the definitions of maximal and minimal eigenvalues introduced in the text, the analyses to follow utilize the following definitions. When the matrices $N \in \mathbf{R}^{k\times k}$ (with $k = n$ or $n - 1$) and $Q \in \mathbf{R}^{(n-1)\times(n-1)}$ are symmetric, we let

$$\bar{\lambda}_N = \max\{\lambda\text{: } \lambda \text{ is an eigenvalue of } N\} \text{ and}$$
$$\underline{\lambda}_Q = \min\{\lambda\text{: } \lambda \text{ is an eigenvalue of } Q\}.$$

The Proof of Theorem 5.1(i)

To begin the proof of the first claim of Theorem 5.1(*i*), we establish that if $x = E\sigma$ is a source of (P) = (AB) (i.e., that if all solutions of (P) starting at some $x_0 \neq x$ with $|x - x_0| \leq \varepsilon$ permanently leave the ε-neighborhood of x), then σ is a source of (B). To see this, observe that if $\{\sigma_t\}$ is a solution to (B) with $\|\sigma_0 - \sigma\| \leq \varepsilon$, then it follows from Ely and Sandholm [6, Lemma 2.1] that

$$|E\sigma_0 - x| = |E\sigma_0 - E(B(x))| = |E\sigma_0 - E\sigma| \leq \|\sigma_0 - \sigma\| < \varepsilon.$$

Hence, the solution to (AB) starting from $x_0 = E\sigma_0$ permanently leaves the ε-neighborhood of x. By Ely and Sandholm [6, Theorem 5.2], this solution is actually the trajectory $\{E\sigma_t\}$, so using Ely and Sandholm [6, Lemma 2.1] once again, we find that

$$\|\sigma_t - B(x)\| \geq |E\sigma_t - E(B(x))| = |E\sigma_t - x|.$$

Thus, $\{\sigma_t\}$ eventually leaves the ε-neighborhood of the rest point $B(x)$ forever, establishing our claim.

To show that x is a source of (P) = (AB), it is enough by Lemma 3.2(*i*) to show that if $\underline{\lambda}_{DC(Ax)}\, \underline{\lambda}_{S(\Phi A)} > 1$, then all eigenvalues of the matrix $D\tilde{B}(x) = DC(Ax)A$ with respect to directions in $\mathbf{R}^n_0$ have real part greater than 1. If $DC(Ax)$ is not positive definite on $\mathbf{R}^n_0$ (i.e., if it is only positive semidefinite on $\mathbf{R}^n_0$), then Lemma A.3 and the Rayleigh-Ritz Theorem (Horn and Johnson [21, Thoerem 4.2.2] imply that $\underline{\lambda}_{DC(Ax)} = 0$, so the antecedent inequality does not hold. Our conclusion therefore follows from the following lemma, which builds on results of Hines [13], Hofbauer and Sigmund [19, Section 16.4], and Hopkins [20].

Lemma A.4: *Suppose that $D \in \mathbf{R}^{n\times n}$ is symmetric, satisfies $D\mathbf{1} = \mathbf{0}$, is positive definite on $\mathbf{R}^n_0$, and maps $\mathbf{R}^n_0$ into itself. In addition, suppose that $A \in \mathbf{R}^{n\times n}$ is positive definite with respect to $\mathbf{R}^n_0$. Let $\underline{\lambda}_{DA}$ be the smallest real part of an eigenvalue of DA restricted to $\mathbf{R}^n_0$. Then $\underline{\lambda}_{DA} \geq \underline{\lambda}_{D}\, \underline{\lambda}_{S(\Phi A)}$.*

Proof: Suppose that

$$DA(v + \mathrm{i}w) = (\alpha + \mathrm{i}\beta)(v + \mathrm{i}w) \tag{1}$$

for some v, $w \in \mathbf{R}^n_0$ and α, $\beta \in \mathbf{R}$. Since D is positive definite on $\mathbf{R}^n_0$ and maps $\mathbf{R}^n_0$ to itself, it is one-to-one on $\mathbf{R}^n_0$, so there exist y, $z \in \mathbf{R}^n_0$ such that $Dy = v$ and $Dz = w$. Moreover, since $D\mathbf{1} = \mathbf{0}$ and since Φ projects $\mathbf{R}^n$ onto $\mathbf{R}^n_0$, we can rewrite equation (1) as

$$D\Phi A(v + \mathrm{i}w) = (\alpha + \mathrm{i}\beta)(v + \mathrm{i}w) = (\alpha + \mathrm{i}\beta)D(y + \mathrm{i}z).$$

Since D is one-to-one on $\mathbf{R}^n_0$, it follows that

$$\Phi A(v + \mathrm{i}w) = (\alpha + \mathrm{i}\beta)(y + \mathrm{i}z).$$

Since D is symmetric, premultiplying by $(v - \mathrm{i}w)'$ yields

$$(v - \mathrm{i}w)'\, \Phi A(v + \mathrm{i}w) = (\alpha + \mathrm{i}\beta)\,(v - \mathrm{i}w)'\,(y + \mathrm{i}z) = (\alpha + \mathrm{i}\beta)\,(y + \mathrm{i}z)'\,D(y - \mathrm{i}z).$$

If we equate the real parts of the previous expression and collect terms, we obtain

$$v'\,\Phi A v + w'\,\Phi A w = \alpha(y'\,Dy + z'\,Dz) = \alpha(v'\,D^{-1}\,v + w'\,D^{-1}\,w), \tag{2}$$

where D^{-1} inverts D on $\mathbf{R}^n_0$. Since $v \in \mathbf{R}^n_0$ and since

$$S(Z\Phi A\,Z') = \tfrac{1}{2}(Z\Phi AZ' + ZA'\Phi Z') = Z(\tfrac{1}{2}(\Phi A + A'\Phi))Z' = ZS(\Phi A)\,Z',$$

Lemma A.2, the Rayleigh-Ritz Theorem, and Lemma A.3 imply that

$$\begin{aligned} v'\,\Phi A v &= v'\,\Phi\Phi A\Phi v \\ &= v'\,Z'\,(ZA\Phi Z')Zv \\ &= v'\,Z'\,S(ZA\Phi Z')Zv \\ &= v'\,Z'\,(Z\,S(A\Phi)\,Z')Zv \\ &\geq \underline{\lambda}_{Z\,S(\Phi A)Z'}\; v'\,Z'\,Zv \\ &= \underline{\lambda}_{S(\Phi A)}\; v'\,\Phi v \\ &= \underline{\lambda}_{S(\Phi A)}\; v'\,v. \end{aligned} \tag{3}$$

At the same time, if we let $\tilde{\lambda}_{D^{-1}}$ denote the largest eigenvalue of D^{-1} restricted to $\mathbf{R}^n_0$, then $\tilde{\lambda}_{D^{-1}} = (\underline{\lambda}_D)^{-1}$, so applying the Rayleigh-Ritz Theorem again yields

$$(4) \qquad v'\,D^{-1}\,v \le \alpha\tilde{\lambda}_{D^{-1}}\;v'\,v = (\underline{\lambda}_D)^{-1}\;v'\,v.$$

Combining expressions (2), (3), and (4), we find that

$$\begin{aligned}\underline{\lambda}_{S(\Phi A)}\,(v'\,v + w'\,w) &\le v'\,\Phi A v + w'\,\Phi A w\\ &= \alpha(v'\,D^{-1}\,v + w'\,D^{-1}\,w)\\ &\le \alpha(\underline{\lambda}_D)^{-1}\;(v'\,v + w'\,w).\end{aligned}$$

We therefore conclude that $\alpha \ge \underline{\lambda}_D\;\underline{\lambda}_{S(\Phi A)}$, and hence that $\underline{\lambda}_{DA} \ge \underline{\lambda}_D\;\underline{\lambda}_{S(\Phi A)}$. ❑

To prove the second claim in Theorem 5.1(*i*), suppose that A is symmetric and is positive definite on $\mathbf{R}^n_0$. Then the claim immediately follows from Lemma 3.2(*i*) and Lemma A.5 below. The lemma builds on a well known fact from matrix analysis: if $D \in \mathbf{R}^{n\times n}$ is symmetric and $A \in \mathbf{R}^{n\times n}$ is symmetric positive definite, then DA and D have the same inertia—that is, the same numbers of positive, negative, and zero eigenvalues (see Horn and Johnson [21, Theorem 7.6.3]).

Lemma A.5: *Suppose that* $D \in \mathbf{R}^{n\times n}$ *is symmetric, satisfies* $D\mathbf{1} = \mathbf{0}$, *and maps* $\mathbf{R}^n_0$ *into itself, and suppose that* $A \in \mathbf{R}^{n\times n}$ *is symmetric and that it is positive definite on* $\mathbf{R}^n_0$. *Then the eigenvalues of DA with respect to* $\mathbf{R}^n_0$ *are real, and the largest one,* $\bar{\lambda}_{DA}$, *is at least* $\bar{\lambda}_D\;\underline{\lambda}_{\Phi A\Phi}$.

Proof: The eigenvalues of DA with respect to $\mathbf{R}^n_0$ are the same as the eigenvalues of $DA\Phi$ with respect to $\mathbf{R}^n_0$; by Lemma A.3, these in turn are equal to the eigenvalues of $ZDA\Phi Z'$ with respect to $\mathbf{R}^{n-1}$. Since $D\mathbf{1} = \mathbf{0}$, we can use Lemma A.2 to express this last product as

$$ZDA\Phi Z' = ZD\Phi\Phi A\Phi Z' = ZDZ'Z\Phi A\Phi Z' = \hat{D}\hat{A},$$

where $\hat{D} = ZDZ' \in \mathbf{R}^{(n-1)\times(n-1)}$ is symmetric and $\hat{A} = Z\Phi A\Phi Z' \in \mathbf{R}^{(n-1)\times(n-1)}$ is symmetric positive definite.

Since $\hat{A}$ has a symmetric positive definite square root $\hat{A}^{1/2}$, $\hat{D}\hat{A}$ is similar to

$$\hat{A}^{1/2}\hat{D}\hat{A}\hat{A}^{-1/2} = \hat{A}^{1/2}\hat{D}\hat{A}^{1/2}$$

Therefore, Ostrowski's Theorem (Horn and Johnson [21, Theorem 4.5.9]), a quantitative version of Sylvester's Law of Inertia, implies that the eigenvalues of $\hat{D}\hat{A}$ are real, and that $\bar{\lambda}_{\hat{D}\hat{A}} \geq \bar{\lambda}_{\hat{D}}\, \underline{\lambda}_{\hat{A}}$, where $\bar{\lambda}_{\hat{D}\hat{A}}$ is the largest eigenvalue of $\hat{D}\hat{A}$. Since we established earlier that the eigenvalues of DA restricted to $\mathbf{R}^n_0$ are the same as the eigenvalues of $\hat{D}\hat{A}$, and since $\bar{\lambda}_D = \bar{\lambda}_{\hat{D}}$ and $\underline{\lambda}_{\Phi A \Phi} = \underline{\lambda}_{\hat{A}}$ by Lemma A.3, we are able to conclude that $\bar{\lambda}_{DA} \geq \bar{\lambda}_D\, \underline{\lambda}_{\Phi A \Phi}$. ❑

This completes the proof of Theorem 5.1(*i*). ■

The Proof of Theorem 5.1(*ii*)

To establish the first sufficient condition stated in part (*ii*) of the theorem, it is enough by Lemma 3.2(*ii*) to show that if $\underline{\lambda}_{DC^1(Ax^2)}\, \bar{\lambda}_{DC^2(A'x^1)}\, \underline{\lambda}_{\Phi A \Phi A' \Phi}$ strictly exceeds 1, then $D\tilde{B}^1(x^2)\, D\tilde{B}^2(x^1) = DC^1(Ax^2)\, A\, DC^2(A'x^1)\, A'$ has a real eigenvalue that is strictly greater than one.

If $DC^1(Ax^2)$ is only positive semidefinite on $\mathbf{R}^n_0$, then Lemma A.3 and the Rayleigh-Ritz Theorem imply that $\underline{\lambda}_{DC^1(Ax^2)} = 0$, and so that $\underline{\lambda}_{DC^1(Ax^2)}\, \bar{\lambda}_{DC^2(A'x^1)}\, \underline{\lambda}_{\Phi A \Phi A' \Phi} = 0$; we may therefore assume that $DC^1(Ax^2)$ is positive definite on $\mathbf{R}^n_0$. In addition, since (A, A) has an isolated interior equilibrium, A must be square (i.e., n^1 and n^2 must be equal—see Hofbauer and Sigmund [19, Section 17.4]); it must also be true that when $z \in \mathbf{R}^n_0$ is nonzero, Az is not a multiple of $\mathbf{1}$, or equivalently, $\Phi A z \neq \mathbf{0}$ (otherwise, if x^* is an interior equilibrium, then so is $x^* + \varepsilon z$). Therefore, ΦA defines a bijective map from $\mathbf{R}^n_0$ to itself.

In light of these observations, the sufficiency of the first inequality in part (*ii*) of the theorem is a consequence of the following lemma, which builds on results of Hofbauer and Hopkins [15]. The other sufficient condition is obtained by applying this result to the product $D\tilde{B}^2(x^1)\, D\tilde{B}^1(x^2)$.

Lemma A.6: *Let $D^1 \in \mathbf{R}^{n \times n}$ and $D^2 \in \mathbf{R}^{n \times n}$ be symmetric, map $\mathbf{R}^n_0$ into itself, and map the vector $\mathbf{1}$ to the origin, and suppose that D^1 is positive definite on $\mathbf{R}^n_0$ and that D^2 is positive semidefinite on $\mathbf{R}^n_0$. In addition, suppose that $A \in \mathbf{R}^{n \times n}$ is such that ΦA is a bijective map from $\mathbf{R}^n_0$ to itself. Then the eigenvalues of $D^1 A D^2 A'$ with respect to $\mathbf{R}^n_0$ are real, and the largest one, $\bar{\lambda}_{D^1 A D^2 A'}$, is at least $\underline{\lambda}_{D^1}\, \bar{\lambda}_{D^2}\, \underline{\lambda}_{\Phi A \Phi A' \Phi}$.*

Proof: The eigenvalues of $D^1 A D^2 A'$ with respect to $\mathbf{R}^n_0$ are the same as the eigenvalues of $D^1 A D^2 A' \Phi$ with respect to $\mathbf{R}^n_0$; by Lemma A.3, these in turn are equal to the eigenvalues of $ZD^1AD^2A'\Phi Z'$ with respect to $\mathbf{R}^{n-1}$. Now since $D^1\mathbf{1} = \mathbf{0}$ and $D^2\mathbf{1} = \mathbf{0}$, and since the range of D^2 is contained in $\mathbf{R}^n_0$, we can write

$$\begin{aligned} ZD^1AD^2A'\Phi Z' &= ZD^1\Phi\Phi A\Phi D^2\Phi A'\Phi Z' \\ &= ZD^1Z'Z\Phi AZ'ZD^2Z'ZA'\Phi Z' \end{aligned}$$

$$= \hat{D}^1 \hat{A} \hat{D}^2 \hat{A}',$$

where $\hat{D}^1 = Z D^1 Z'$ is symmetric positive definite on $\mathbf{R}^{n-1}$, $\hat{D}^2 = Z D^2 Z'$ is symmetric positive semidefinite on $\mathbf{R}^{n-1}$, and $\hat{A} = Z\Phi AZ'$ has full rank on $\mathbf{R}^{n-1}$.

Since $\hat{A}$ has full rank on $\mathbf{R}^{n-1}$, $\hat{A} \hat{D}^2 \hat{A}'$ is congruent to $\hat{D}^2$. Hence, Ostrowski's Theorem implies that the eigenvalues of $\hat{A} \hat{D}^2 \hat{A}'$ are real and that

$$\overline{\lambda}_{\hat{A}\hat{D}^2\hat{A}'} \geq \overline{\lambda}_{\hat{D}^2} \underline{\lambda}_{\hat{A}\hat{A}'}.$$

Moreover, $\hat{D}^1 \hat{A} \hat{D}^2 \hat{A}'$ is similar to

$$(\hat{D}^1)^{-1/2} \hat{D}^1 \hat{A} \hat{D}^2 \hat{A}' (\hat{D}^1)^{1/2} = (\hat{D}^1)^{1/2} \hat{A} \hat{D}^2 \hat{A}' (\hat{D}^1)^{1/2}.$$

Therefore, employing Ostrowski's Theorem once again, we find that the eigenvalues of $\hat{D}^1 \hat{A} \hat{D}^2 \hat{A}'$ are real and that the maximal eigenvalue $\overline{\lambda}_{\hat{D}^1\hat{A}\hat{D}^2\hat{A}'}$ satisfies

$$\overline{\lambda}_{\hat{D}^1\hat{A}\hat{D}^2\hat{A}'} \geq \underline{\lambda}_{\hat{D}^1} \overline{\lambda}_{\hat{A}\hat{D}^2\hat{A}'} \geq \underline{\lambda}_{\hat{D}^1} \overline{\lambda}_{\hat{D}^2} \underline{\lambda}_{\hat{A}\hat{A}'}.$$

We established above that the eigenvalues of $D^1 A D^2 A'$ with respect to $\mathbf{R}^n_0$ are the same as the eigenvalues of $\hat{D}^1 \hat{A} \hat{D}^2 \hat{A}'$. In addition, Lemma A.3 implies that $\underline{\lambda}_{\hat{D}^1} = \underline{\lambda}_{D^1}$, that $\overline{\lambda}_{\hat{D}^2} = \overline{\lambda}_{D^2}$, and, since

$$\hat{A} \hat{A}' = Z\Phi AZ'ZA'\Phi Z' = Z\Phi A\Phi A'\Phi Z',$$

that $\underline{\lambda}_{\hat{A}\hat{A}'} = \underline{\lambda}_{\Phi A\Phi A'\Phi}$. We therefore conclude that the eigenvalues of $D^1 A D^2 A'$ are real and that $\overline{\lambda}_{D^1 AD^2 A'} \geq \underline{\lambda}_{D^1} \overline{\lambda}_{D^2} \underline{\lambda}_{\Phi A\Phi A'\Phi}$. ❑

This completes the proof of Theorem 5.1(*ii*). ■

We note as an aside that the last terms of the two sufficient conditions from Theorem 5.1(*i*), $\underline{\lambda}_{\Phi A\Phi A'\Phi}$ and $\underline{\lambda}_{\Phi A'\Phi A\Phi}$, are equal. To see this, observe that the eigenvalues of $\Phi A\Phi A'\Phi$ with respect to $\mathbf{R}^n_0$ equal the eigenvalues of $Z\Phi A\Phi A'\Phi Z' = \hat{A} \hat{A}'$; $\hat{A} \hat{A}'$ is similar to $\hat{A}' \hat{A}$ (Horn and Johnson [21, Theorem 1.3.20]), which has the same eigenvalues as $\Phi A'\Phi A\Phi$ with respect to $\mathbf{R}^n_0$. Since $\hat{A} \hat{A}'$ is symmetric positive semidefinite (Horn and Johnson [21, Observation 7.1.6]), this argument also shows that $\underline{\lambda}_{\Phi A\Phi A'\Phi}$ and $\underline{\lambda}_{\Phi A'\Phi A\Phi}$ are nonnegative.

The following result on the instability of interior equilibria of potential games was noted in connection with Theorem 5.1(*ii*).

Proposition A.7: *Let (A, A) be a normal form potential game, let F be the population game that obtains when two unit mass populations are randomly matched to play (A, A), and let f(x) =* $x^1 \cdot A x^2$ *be the potential function for F. If x* is an isolated interior Nash equilibrium of F (and hence of (A, A)), then x* is a saddle point of f with respect to directions in TX.*

Proof: Proposition 3.1 of Sandholm [24] tells us that any interior Nash equilibrium x^* of F is a critical point of f, in the sense that $\nabla f(x^*)$ is orthogonal to $TX = \mathbf{R}_0^{n^1} \times \mathbf{R}_0^{n^2}$. To show that x^* is a saddle point of f, it is enough to show that the eigenvalues of $\nabla^2 f(x)$ with respect to TX come in pairs of the form $\pm\sqrt{\lambda_i}$ with $\lambda_i > 0$.

As we noted in the proof of Theorem 5.1(*ii*), that (A, A) admits an isolated interior equilibrium implies that A is square and has full rank on $\mathbf{R}_0^n$ ($\Phi A z \neq 0$ for all $z \in \mathbf{R}_0^n$). Now observe that

$$\nabla^2 f(x) = DF(x) = \begin{pmatrix} 0 & A \\ A' & 0 \end{pmatrix}.$$

It is easy to show that the eigenvalues of this matrix with respect to $\mathbf{R}_0^n \times \mathbf{R}_0^n$ are the same as the eigenvalues of

$$\begin{pmatrix} 0 & ZAZ' \\ ZA'Z' & 0 \end{pmatrix}$$

with respect to $\mathbf{R}^{n-1} \times \mathbf{R}^{n-1}$ (cf. Lemma A.3), and that the eigenvalues of the latter matrix are the positive and negative square roots of the eigenvalues of the symmetric matrix $ZAZ'ZA'Z' = ZA\Phi\Phi'A'Z'$ (cf. Lemma A.1). The full rank condition on A implies that the rank of $ZA\Phi$ is $n - 1$, and hence that the rank of $ZA\Phi\Phi'A'Z'$ is $n - 1$ (Horn and Johnson [21, Observation 7.1.6]). Therefore, all $n - 1$ eigenvalues of $ZA\Phi\Phi'A'Z'$ are strictly positive. Since the relevant eigenvalues of $\nabla^2 f(x)$ are the positive and negative square roots of these numbers, the proof is complete. ■

We conclude by proving our global stability results.

The Proof of Theorem 6.1

(*i*) By Lemma 3.2(*i*), it is enough to show that the eigenvalue of $D\tilde{B}(x) = DC(Ax)A$ with respect to $\mathbf{R}_0^n$ of largest modulus, or, equivalently, the eigenvalue of $D\tilde{B}(x)\Phi$ with largest modulus, has modulus less than one, since in this case all eigenvalues of $D\tilde{B}(x)\Phi$ have real parts with absolute value less than one. But since the matrix $DC(Ax)$ maps the vector $\mathbf{1}$ to the origin (so that $DC(Ax) = DC(Ax)\Phi$) and is symmetric positive semidefinite, properties (S1)-(S3) imply that

$$\left|\tilde{\lambda}_{D\tilde{B}(x)\Phi}\right| \le \bar{s}_{D\tilde{B}(x)\Phi} = \bar{s}_{DC(Ax)A\Phi} = \bar{s}_{DC(Ax)\Phi A\Phi} \le \bar{s}_{DC(Ax)}\,\bar{s}_{\Phi A\Phi} = \bar{\lambda}_{DC(Ax)}\,\bar{s}_{\Phi A\Phi} < 1.$$

(*ii*) By Lemma 3.2(*ii*), it is enough to show that the eigenvalue of $D\tilde{B}^1(x^2)\,D\tilde{B}^2(x^1) = DC^1(Ax^2)\,A\,DC^2(B'x^1)\,B'$ with respect to $\mathbf{R}_0^{n^1}$ of largest modulus has modulus less than one: in this case, the square root of this eigenvalue and of all other eigenvalues have modulus less than one, and so the real parts of these square roots have absolute value less than one. Since we are only concerned with eigenvalues with respect to $\mathbf{R}_0^{n^1}$, and since the range of $D\tilde{B}^2(x^1)$ is $\mathbf{R}_0^{n^2}$, it is sufficient to show that the eigenvalue of $D\tilde{B}^1(x^2)\,\Phi D\tilde{B}^2(x^1)\,\Phi$ with largest modulus has modulus less than one. Reasoning as above, we find that

$$\begin{aligned}
\left|\tilde{\lambda}_{D\tilde{B}^1(x^2)\Phi D\tilde{B}^2(x^1)\Phi}\right| &\le \bar{s}_{D\tilde{B}^1(x^2)\Phi D\tilde{B}^2(x^1)\Phi} \\
&= \bar{s}_{DC^1(Ax^2)A\Phi DC^2(B'x^1)B'\Phi} \\
&= \bar{s}_{DC^1(Ax^2)\Phi A\Phi DC^2(B'x^1)\Phi B'\Phi} \\
&\le \bar{s}_{DC^1(Ax^2)}\,\bar{s}_{\Phi A\Phi}\,\bar{s}_{DC^2(B'x^1)}\,\bar{s}_{\Phi B'\Phi} \\
&= \bar{\lambda}_{DC^1(Ax^2)}\,\bar{\lambda}_{DC^2(B'x^1)}\,\bar{s}_{\Phi A\Phi}\,\bar{s}_{\Phi B\Phi} \\
&< 1. \blacksquare
\end{aligned}$$

The Proof of Theorem 6.3

Let A or (A, B) be a normal form game from any of the classes mentioned in the statement of the theorem. Then random matching in this game generates a population game F whose derivative matrices $DF(x)$ are negative semidefinite with respect to TX. In other words, F is a *stable game* (Hofbauer and Sandholm [18]).

Using a Lyapunov function argument, Hofbauer and Sandholm [17, Theorem 3.1] show that the dynamic (P) admits a globally asymptotically stable rest point whenever the distributions μ^p are smooth and have full support and the game F is stable. Theorem 6.6 of Ely and Sandholm [6] shows that aggregate behavior x is globally asymptotically stable under (P) = (AB) if and only if the Bayesian strategy $B(x)$ is globally asymptotically stable under (B). Together, these facts establish our result. ■

References

[1] S. P. Anderson, A. de Palma, J.-F. Thisse, Discrete Choice Theory of Product Differentiation, MIT Press, Cambridge, 1992.

[2] K. Binmore, L. Samuelson, Muddling through: noisy equilibrium selection, J. Econ. Theory 74 (1997), 235-265.

[3] K. Binmore, L. Samuelson, Evolution and mixed strategies, Games Econ. Behav. 34 (2001), 200-226.

[4] A. Edlin, F. Echenique, Mixed equilibria are unstable in games of strategic complements, J. Econ. Theory 118 (2004), 61-79.

[5] G. Ellison, D. Fudenberg, Learning purified mixed equilibria, J. Econ. Theory 90 (2000), 84-115.

[6] J. C. Ely, W. H. Sandholm, Evolution in Bayesian games I: Theory, Games Econ. Behav. 53 (2005), 83-109.

[7] S. H. Friedberg, A. J. Insel, L. E. Spence, Linear Algebra, 2nd ed., Prentice Hall, Englewood Cliffs, NJ, 1989.

[8] D. Fudenberg, D. M. Kreps, Learning mixed equilibria, Games Econ. Behav. 5 (1993), 320-367.

[9] D. Fudenberg, D. K. Levine, Theory of Learning in Games, MIT Press, Cambridge, 1998.

[10] A. Gaunersdorfer, J. Hofbauer, Fictitious play, Shapley polygons, and the replicator equation, Games Econ. Behav. 11 (1995), 279-303.

[11] I. Gilboa, A. Matsui. Social stability and equilibrium, Econometrica 59 (1991), 859-867.

[12] J. C. Harsanyi, Games with randomly disturbed payoffs: a new rationale for mixed-strategy equilibrium points, Int. J. Game Theory 2 (1973), 1-23.

[13] W. G. S. Hines, Three characterizations of population strategy stability, J. Appl. Prob. 17 (1980), 333-340. Correction, R. Cressman, W. G. S. Hines, J. Appl. Prob. 21 (1984), 213-214.

[14] J. Hofbauer, From Nash and Brown to Maynard Smith: Equilibria, dynamics, and ESS, Selection 1 (2000), 81-88.

[15] J. Hofbauer, E. Hopkins, Learning in perturbed asymmetric games, Games Econ. Behav. 52 (2005), 133-152.

[16] J. Hofbauer, W. H. Sandholm, On the global convergence of stochastic fictitious play, Econometrica 70 (2002), 2265-2294.

[17] J. Hofbauer, W. H. Sandholm, Evolution in games with randomly disturbed

payoffs, forthcoming, J. Econ. Theory.

[18] J. Hofbauer, W. H. Sandholm, Stable games, unpublished manuscript, University of Vienna and University of Wisconsin, 2006.

[19] J. Hofbauer, K. Sigmund, Theory of Evolution and Dynamical Systems, Cambridge University Press, Cambridge, 1988.

[20] E. Hopkins, A note on best response dynamics, Games Econ. Behav. 29 (1999), 138-150.

[21] R. A. Horn, C. R. Johnson, Matrix Analysis, Cambridge University Press, Cambridge, 1985.

[22] R. A. Horn, C. R. Johnson, Topics in Matrix Analysis, Cambridge University Press, Cambridge, 1991.

[23] D. Monderer, L. S. Shapley, Potential games, Games Econ Behav. 14 (1996), 124-143.

[24] W. H. Sandholm, Potential games with continuous player sets, J. Econ. Theory 97 (2001), 81-108.

[25] W. H. Sandholm, Preference evolution, two-speed dynamics, and rapid social change, Rev. Econ. Dynamics 4 (2001), 637-679.

[26] W. H. Sandholm, Evolution in Bayesian games II: Stability of purified equilibria, SSRI Working Paper #2003-21R, University of Wisconsin, 2005.

[27] J. W. Weibull, Evolutionary game theory, MIT Press, Cambridge, 1995.