ISSN 1471-0498

DEPARTMENT OF ECONOMICS

DISCUSSION PAPER SERIES

DYNAMICALLY STABLE PREFERENCES

Thomas W.L. Norman

Number 207

October 2004

Manor Road Building, Oxford OX1 3UQ

Dynamically Stable Preferences*

Thomas W. L. Norman

All Souls College, Oxford

October 5, 2004

Abstract

This paper models the indirect evolution of the preferences of a population of fully rational agents repeatedly matched to play a symmetric 2×2 game in biological fitnesses. Each agent is biased in favor of one of the strategies, and receives a noisy signal of his and his opponent's bias. With sufficiently accurate signals, the resulting global game selects a unique outcome, allowing preference biases to be shaped by the replicator dynamics. Stability analysis in this setting requires the extension of recent techniques for evolution on infinite strategy spaces, introducing new setwise stability concepts. In coordination games, the interval of preference biases supporting the Pareto-dominant equilibrium is Lyapunov stable and weakly attracting, by virtue of constituting a "strongly uninvadable set." In Prisoners' Dilemmas that satisfy Kandori and Rob's (*Games and Economic Behavior* **22**, 1998, 30–60) "marginal bandwagon property," meanwhile, an interval of biases supporting efficient cooperation is a "neutrally uninvadable set," and thus Lyapunov stable. *Journal of Economic Literature* Classification Number: C70, C72.

Key Words: evolution; preferences; global games; replicator dynamics; continuous strategy spaces; evolutionary stability.

> "Man with all his noble qualities. . . still bears in his bodily frame the indelible stamp of his lowly origin."
>
> Charles Darwin, *The Descent of Man* (1871)

*I gratefully acknowledge discussions with David Myatt and Chris Wallace, as well as with seminar participants at Oxford University. The usual disclaimer applies. Email thomas.norman@all-souls.oxford.ac.uk

1 Introduction

Ideas and techniques from evolutionary biology have been fruitfully applied to the modelling of adaptive economic behavior in evolutionary game theory. Such models have provided powerful support for fully rational game-theoretic equilibria as the eventual outcomes reached by groups of boundedly rational individuals. However, the fit between biological models and traditional microeconomic assumptions remains an uneasy one, with evolutionary agents seemingly distant ancestors of *homo economicus*. This paper applies evolutionary methods at a more biological level, modelling the evolution of the *preferences* of fully rational agents playing symmetric 2×2 games in "biological fitnesses."

The model adopts the "indirect evolutionary approach" originating in the work of Güth (1995) and Güth and Yaari (1992), but with antecedents in Frank (1987) and Schelling (1960) for example. This approach assumes that individuals play rationally for given preferences, but that these preferences need not coincide with game fundamentals (fitnesses) and are subject to evolutionary selection according to their success. Divergence from fitness-maximizing preferences is then possible because of the resulting effect on opponents' play, as has been demonstrated very generally by Heifetz, Shannon, and Spiegel (2003).[1]

As Samuelson (2001b) notes, the indirect evolutionary approach raises two key questions. First, what preferences would emerge evolutionary victors if the whole range of possible preferences were allowed to compete, rather than some subset chosen for the example at hand? Second, can non-fitness-maximizing preferences emerge in the absence of preference observability? Dekel, Ely, and

[1]See also, for example, Benos (1998), Bergman and Bergman (2000), Bolle (2000), Bester and Güth (1998), Eshel, Samuelson and Shaked (1988), Fershtman and Weiss (1997, 1998), Heifetz and Segev (2003), Guttman (2000), Huck, Kirchsteiger, and Oechssler (1999), Huck and Oechssler (1999), Koçkesen, Ok and Sethi (2000a, 2000b), Kyle and Wang (1997), Possajennikov (2000), Rotemberg (1994), and Sethi and Somanathan (2001).

Yilankaya (2004) provide an answer to these questions with their conclusion that an outcome is stable only if it is efficient, given observability of preferences. This nontrivial role for non-fitness-maximizing preferences breaks down, however, if preferences are unobservable, in which case the stability concept employed coincides with the notion of neutrally stable strategies. Ok and Vega-Redondo (2001), meanwhile, also find that complete information is crucial to departures from "individualistic" preferences, whilst Ely and Yilankaya (2001) and Güth and Peleg (2001) reach similar conclusions.

Dekel, Ely, and Yilankaya (2004) consider an intermediate case between complete and incomplete information where players know the population distribution of preferences and observe their opponents' preferences with a probability $p \in (0, 1)$. They demonstrate local robustness of their findings for the polar cases to small changes in p (except for sufficiency of neutral stability in the unobservable preferences case). Heifetz, Shannon, and Spiegel (2003) also consider a number of intermediate cases between complete and incomplete information in the context of an example with a unique Bayesian equilibrium: first, that preferences are observable in some fraction of interactions but unobserved in others; second, that preferences are observed with noise; and third, that players can make a costly signal of preferences. They find that non-fitness-maximizing preferences may remain evolutionarily viable in these imperfect observability scenarios.

Whilst a good deal is known then about the stability properties of preferences under indirect evolution, there remains a key obstacle to a general *dynamic* analysis: the possibility of multiple equilibria precludes the existence of a well-defined function from preferences to fitnesses. Hence, existing studies have made use of static stability concepts, partial dynamic analysis or cases with a unique equilibrium. Sandholm (2001) overcomes this problem by nesting

a model of (rapid) behavioral evolution within a model of (sluggish) preference evolution, thus providing a full dynamic analysis of "two-speed" indirect evolution. In particular, Sandholm's players have idiosyncratic preference *biases* in a two-strategy game, and can only learn the aggregate distribution of behavior—they have no knowledge of opponents' preferences. In this setting, aggregate behavior adjusts smoothly to the population analog of mixed-strategy equilibrium in equilibration games such as Hawk-Dove, whilst discrete jumps are possible in coordination games, with limit behavior depending on initial conditions.

This paper considers the intermediate case between complete and incomplete information where, having publicly observed the population distribution of preferences for a given underlying fitness game, players privately observe a noisy signal of match-specific preferences. Common knowledge of the fundamentals fails in this setting, giving a global game (Carlsson and van Damme 1993, Morris and Shin 2002) in payoffs. The unique equilibrium selection resulting from such games allows expected fitness to be expressed as a (well-defined) function of preferences. This, in turn, means that the replicator dynamics (Taylor and Jonker 1978) are well defined on the space of preferences, allowing an investigation of their *dynamic* indirect evolution in a fully rational setting.

Specifically, it is assumed that the underlying fitness game is transformed by a preference bias in favor of one strategy or the other to give the (global) game in payoffs. The replicator dynamics then shape preference biases in the familiar manner, increasing the frequency of those that deliver high expected fitness relative to the population average. In order to analyze *continuous* biases—as is required by the global games framework—use is made of the replicator dynamics for infinite strategy spaces (Bomze 1990, 1991, Oechssler

and Riedel 2001, 2002). In particular, the variational norm is adopted as the measure of "distance" between populations, which are described by probability measures on the space of preference biases. Evolutionarily stable strategies are no longer asymptotically stable with infinite strategy spaces, but the stronger requirement of a homogeneous, "uninvadable" population does suffice (Oechssler and Riedel 2001). In the present model though, no populations satisfy this requirement—except possibly in Hawk-Dove fitness games, which have the character of an "arms race" in biases. Instead, a setwise generalization of uninvadability—a "strongly uninvadable set"—is defined and shown to give the infinite-dimensional analog of asymptotic stability for a set of probability measures on a closed interval of the real line.

This result may be of general interest, but for the present purposes it implies that the set of bias distributions giving the Pareto-dominant equilibrium of coordination fitness games—a strongly uninvadable set—satisfies the infinite-dimensional analog of asymptotic stability under the replicator dynamics. Moreover, the set of populations supporting the Pareto-dominated equilibrium are shown *not* to satisfy Lyapunov stability. In Prisoners' Dilemma fitness games, meanwhile, if Kandori and Rob's (1998) "marginal bandwagon property" (MBP) fails, then the set of bias distributions giving the usual inefficient outcome is strongly uninvadable, and hence satisfies the infinite-dimensional analog of asymptotic stability. However, if the MBP is satisfied for Prisoners' Dilemma fitness games, then there are no strongly uninvadable sets of biases, and a weaker setwise stability concept is needed. A "neutrally uninvadable set" is defined, and shown to imply Lyapunov stability under the replicator dynamics for a set of probability measures on a closed real interval. This is again a general result, but in the model at hand it implies that a set of bias distributions giving efficient cooperation is neutrally uninvadable, and hence

Lyapunov stable. The usual inefficient outcome, on the other hand, cannot be supported by such stable populations. Intuitively, Prisoners' Dilemmas that are not "too bad" (in the sense of the MBP) can be transformed into coordination games where efficient cooperation is the risk-dominant strategy (and is thus played with probability 1).

2 The Model

The Fitness Game A population of individuals $i = 1, \ldots, N$ are repeatedly matched to play a symmetric 2×2 game in (real-valued) *fitnesses*:

	1	2
1	f_{11} f_{11}	f_{21} f_{12}
2	f_{12} f_{21}	f_{22} f_{22}

which can be normalized to give

	1	2
1	f_1 f_1	0 0
2	0 0	f_2 f_2

where $f_1 = f_{11} - f_{21}$ and $f_2 = f_{22} - f_{12}$. This fitness game will fall within one of the usual symmetric 2×2 game classes: coordination ($f_1, f_2 > 0$); Hawk-Dove ($f_1, f_2 < 0$); or dominant-strategy ($f_1 f_2 < 0$).

Preference Biases To arrive at the payoffs that determine individual play, however, any given underlying fitness game is transformed by idiosyncratic *preference biases* for one strategy or the other, $x_i \in [a, b] \subset \mathbb{R}$, where $f_1, -f_2, 0 \in (a, b)$. This gives the following (asymmetric) game in *payoffs*:

	1	2
1	$f_{11} + x_C$ $f_{11} + x_R$	f_{21} $f_{12} + x_R$
2	$f_{12} + x_C$ f_{21}	f_{22} f_{22}

where R denotes the Row player and C the column player.[2] It is this *payoff game* that determines individual play, and thus outcomes; the players are quite insouciant of underlying fitnesses.

Preference Evolution The individual preference biases are determined by Taylor and Jonker's (1978) replicator dynamics, with the attendant complications arising from the infinite strategy space $S = [a, b]$ (Bomze 1990, 1991, Oechssler and Riedel 2001, 2002). The aggregate play of the population is described by a probability measure P on the measure space $(S, \mathscr{B})$, where $\mathscr{B}$ denotes the Borel σ-algebra on S. Δ denotes the simplex of all populations on S. The function $f : S \times S \to \mathbb{R}$ is defined to be a bounded, Borel-measurable function $f(x_i, x_j)$ mapping preference biases into (expected) fitnesses. The

[2]Note that, since x_i can be positive or negative, either strategy may be preferred by any given player.

average fitness of population P against population Q is then

$$E(P,Q) = \int_S \int_S f(x,y)Q(dy)P(dx). \tag{1}$$

Following Oechssler and Riedel (2001), the analysis takes place on the Banach space $\mathscr{M}^e(S,\mathscr{B})$ of signed measures endowed with the supremum or variational norm.[3]

Definition 1 *The* variational norm *on* $\mathscr{M}^e(S,\mathscr{B})$ *is given by*

$$\|\mu\| = \sup_f \left| \int f \, d\mu \right|,$$

where the sup *is taken over all measurable functions* $f : S \to \mathbb{R}$ *bounded by* 1, $\sup_{x\in S} |f(x)| \leq 1$.

For the present purposes, the variational norm's specialization to probability measures,[4]

$$\|P - Q\| = 2 \sup_{A\in\mathscr{B}} |P(A) - Q(A)|,$$

is particularly useful.

The replicator dynamics then, as usual, increase the frequency of strategies that are successful relative to the prevailing average fitness:

$$Q'(t)(A) = \int_A \sigma(x, Q(t))Q(t)(dx), \qquad Q(0) = P, \tag{2}$$

[3] ν is a *signed measure* on $(S,\mathscr{B})$ if there are two finite measures μ^1 and μ^2 such that for all sets $A \in \mathscr{B}$, $\nu(A) = \mu^1(A) - \mu^2(A)$.

[4] See Shiryaev (1995), p. 360.

for all $A \in \mathscr{B}$, where

$$\begin{aligned}\sigma(x, Q) &:= \int_S f(x, y)Q(dy) - \int_S \int_S f(x, y)Q(dy)Q(dx) \\ &= E(\delta_x, Q) - E(Q, Q) \end{aligned} \tag{3}$$

is the success of strategy x if the population is Q, δ_x denoting a Dirac measure with mass 1 on x.

The choice of the replicator dynamics is justified by virtue of its status as the leading biological selection dynamic. In the context of the evolution of preferences, such a biological dynamic seems more appropriate than a behavioral dynamic such as best response. The choice of the variational norm, meanwhile, revolves around the question of what is meant by the "distance" between populations in the infinite-strategy-space setting. Arguments can be made in favor of the variational norm (Oechssler and Riedel 2001) or the topology of weak convergence (Oechssler and Riedel 2002) as the appropriate measure, but here the former is adopted, partly for the sake of convenience.

The Global Payoff Game Players are assumed (publicly) to observe the underlying fitness game f that they are playing at any given time t, along with the current population distribution of preference biases P. Each player then *privately* observes a noisy signal $\tilde{x}^i = (\tilde{x}^i_R, \tilde{x}^i_C) = (x_R, x_C) + \psi E_i$ of the true preference biases (x_R, x_C) in his game, where $\psi > 0$ is a scale parameter, the vector (E_R, E_C) has a continuous density, and the support of each E_i is contained in a ball with radius 1 around 0 in $\mathbb{R}^2$.[5] (E_R, E_C) is assumed independent of $x = (x_R, x_C)$, so that the players' signals are independent conditional on the true preference biases. However, the signals are unconditionally

[5]Note that the E_i's can be such that each player is better informed about his own bias than about his opponent's.

correlated, and hence the players can draw inferences about their opponents' signals based on their own signals. But given the public observation of P, information about x_i is uninformative about x_j, $j \neq i$ (assuming N is large). This setup describes a global payoff game Γ^{ψ} (Carlsson and van Damme 1993), parameterized by ψ.

3 Analysis

So, what will be the equilibrium preferences to emerge from this schema? A dynamic analysis of this question requires the existence of a well-defined expected fitness function f in equation (1). Assuming that the players' signals of (x_R, x_C) are sufficiently accurate, the global payoff game delivers this well-defined function through its equilibrium selection.

3.1 The expected fitness function

Lemma 1 *For ψ sufficiently small, the outcome of Γ^{ψ} is the unique strict equilibrium or the risk-dominant equilibrium.*

This is a direct consequence of Carlsson and van Damme's (1993) (henceforth CvD's) theorem. To see the familiar "infection" argument in this setting—with diffuse priors for convenience—observe that player i will play strategy 1 iff

$$f_{11}\rho_j^i + f_{12}(1-\rho_j^i) + \mathrm{E}(x_i \mid \tilde{x}^i) > f_{21}\rho_j^i + f_{22}(1-\rho_j^i),$$

where $\rho_j^i = \Pr(j \text{ plays } 1 \mid \tilde{x}^i)$. Assuming $f_1 + f_2 > 0$, this holds iff

$$\rho_j^i > \frac{f_2 - \tilde{x}_i^i}{f_1 + f_2}.$$

Since $\rho_j^i \in [0,1]$, strategy 1 is strictly dominant for player i if $\tilde{x}_i^i > f_2$, whilst strategy 2 is strictly dominant if $\tilde{x}_i^i < -f_1$. It then follows from CvD's Lemma 4.2 that, if $(1,1)$ is the iterative dominance solution for given (x_R, x_C), then playing strategy 1 is conditionally dominant at (x_R, x_C) in Γ^ψ when observations are only slightly noisy. Now, consider player i observing signal $\tilde{x}^i = (f_2, f_2)$. Since ρ_j^i is bounded below by the probability that strategy 1 is strictly dominant for player j, it must be the case that

$$\rho_j^i \geq \Pr(\tilde{x}_j^j > f_2 \mid \tilde{x}^i) > 0 = \frac{f_2 - \tilde{x}_i^i}{f_1 + f_2},$$

so that strategy 1 is strictly dominant for player i conditional on the signal $\tilde{x}^i = (f_2, f_2)$. But this must also be true for $\tilde{x}^i$ within an ξ-ball of (f_1, f_2), with ξ sufficiently small. And it then follows that

$$\begin{aligned}\rho_j^i \geq \Pr(\tilde{x}_j^j > f_2 \mid \tilde{x}^i) + \Pr(\tilde{x}_j^j \leq f_2 \mid \tilde{x}^i).\Pr(\tilde{x}^j \in B((f_1, f_2), \xi) \mid \tilde{x}^i) \\ > \Pr(\tilde{x}_j^j > f_2 \mid \tilde{x}^i) \geq \frac{f_2 - \tilde{x}_i^i}{f_1 + f_2}\end{aligned}$$

for $\tilde{x}^i$ on the boundary of $B((f_1, f_2), \xi)$. But then this in turn must also be true for $\tilde{x}^i$ within a ζ-ball of signals on the boundary of $B((f_1, f_2), \xi)$, with ζ sufficiently small, and so on.

This infection argument can be repeatedly applied to give a region D_{11} of iterative dominance of strategy 1. The boundary of this region is then defined by the condition

$$\rho_j^i \geq \Pr(\text{Strategy 1 dominant for } j \mid \tilde{x}^i = \tilde{x}_*) \geq \Pr(p\tilde{x}^j \geq p\tilde{x}_* \mid \tilde{x}^i = \tilde{x}_*) = \frac{f_2 - \tilde{x}_i^i}{f_1 + f_2}$$

for each player i, where $\tilde{x}_*$ is a particular boundary point of D_{11}, $p\tilde{x}^j \geq p\tilde{x}_*$, $p \in \mathbb{R}^2$, is the halfspace above the line $p\tilde{x}^j = p\tilde{x}_*$ that is tangent to some open

ball $B' \subset D_{11}$ with $\tilde{x}_*$ on its boundary. The second inequality here follows from the fact that $\Pr(\tilde{x}^j \in B' \mid \tilde{x}^i = \tilde{x}_*) \to \Pr(p\tilde{x}^j \geq p\tilde{x}_* \mid \tilde{x}^i = \tilde{x}_*)$ as $\psi \to 0$, as demonstrated in Appendix A of CvD. Combining the conditions for each player, and noting that $\tilde{x}^i = \tilde{x}^j = \tilde{x}_* = (\tilde{x}_R, \tilde{x}_C)$,

$$\Pr(p\tilde{x}^j \geq p\tilde{x}_* \mid \tilde{x}^i = \tilde{x}_*) + \Pr(p\tilde{x}^i \geq p\tilde{x}_* \mid \tilde{x}^j = \tilde{x}_*) = \frac{2f_2 - \tilde{x}_R - \tilde{x}_C}{f_1 + f_2}.$$

It then follows from CvD's Lemma 4.1 that

$$\begin{aligned} 1 &= \frac{2f_2 - \tilde{x}_R - \tilde{x}_C}{f_1 + f_2} \\ \Leftrightarrow \tilde{x}_R + \tilde{x}_C &= f_2 - f_1 \end{aligned}$$

as $\psi \to 0$. Note that this is precisely the threshold where the risk-dominant equilibrium shifts between $(1,1)$ and $(2,2)$:

$$\begin{aligned} (f_1 + x_R)(f_1 + x_C) &= (f_2 - x_R)(f_2 - x_C) \\ \Leftrightarrow x_R + x_C &= f_2 - f_1. \end{aligned}$$

A similar argument establishes the iterative dominance of strategy 2 in the complementary region.

The resulting cutoff strategies employed by the players in Γ^ψ are depicted in Figure 1(a), where a unique equilibrium is selected as a function of the true preference biases (x_R, x_C) as $\psi \to 0$. The expected fitness function $f(x_i, x_j)$ is thus (well) defined (and continuous for $\psi > 0$).

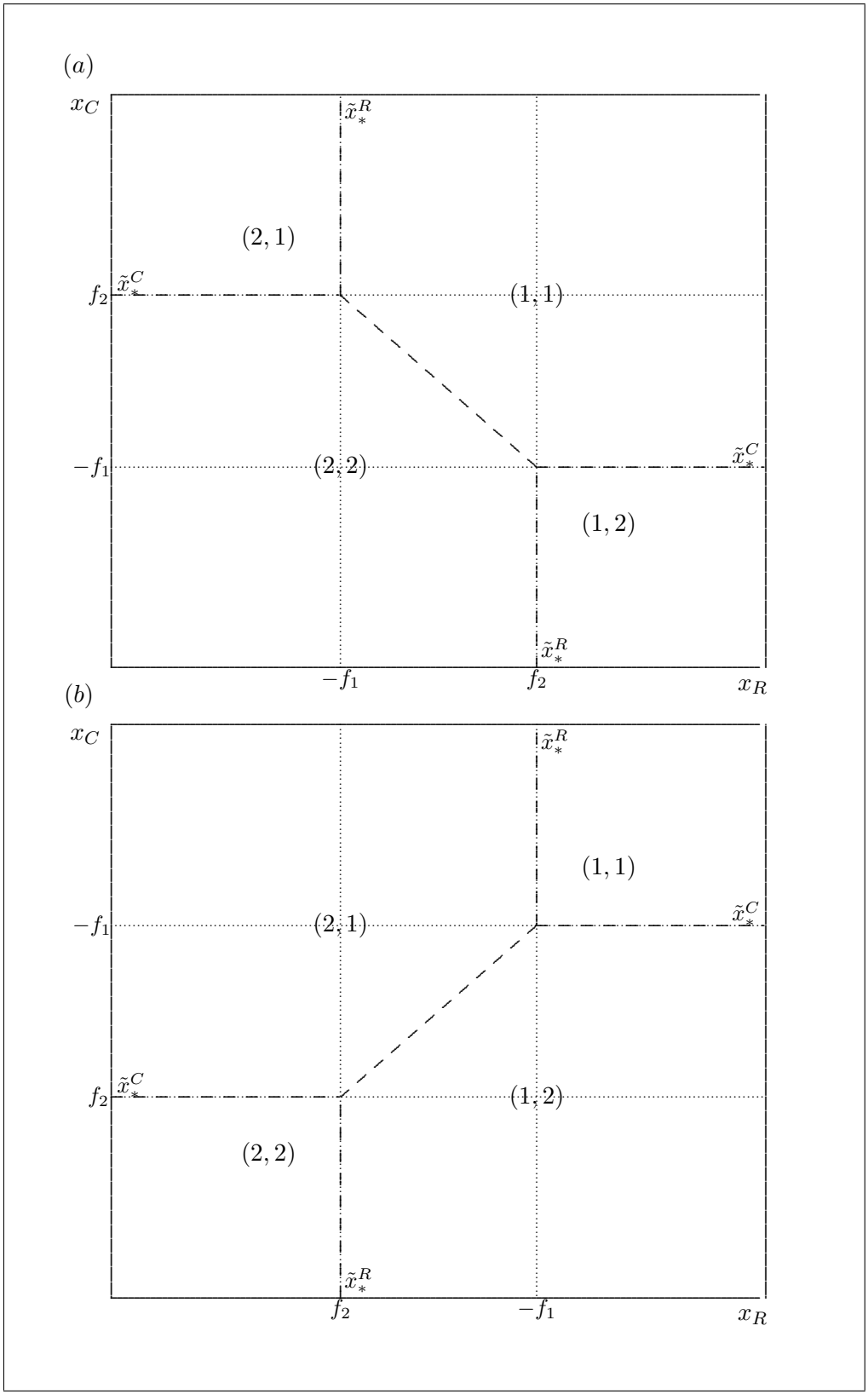

Figure 1: Γ^ψ cutoff strategies, $\psi \to 0$—(a) $f_1 + f_2 > 0$; (b) $f_1 + f_2 < 0$

For the $f_1 + f_2 < 0$ case, a similar argument yields the cutoff strategies in Figure 1(b). Here though, player i will play strategy 1 iff

$$\rho_j^i < \frac{f_2 - \tilde{x}_i^i}{f_1 + f_2},$$

so that strategy 1 is strictly dominant if $\tilde{x}_i^i > -f_1$ and strategy 2 is strictly dominant if $\tilde{x}_i^i < f_2$. The signal $\tilde{x}^i = (-f_1, f_2)$ is now the platform for the region D_{12} where the outcome $(1, 2)$ is the iteratively-dominant-strategy equilibrium, since it gives

$$\begin{aligned} 1 - \rho_C^R &\geq \Pr(\tilde{x}_C^C < f_2 \mid \tilde{x}^R) \\ \Rightarrow \rho_C^R &\leq 1 - \Pr(\tilde{x}_C^C < f_2 \mid \tilde{x}^R) < 1 = \frac{f_2 - \tilde{x}_R^R}{f_1 + f_2} \end{aligned}$$

and

$$\rho_R^C \geq \Pr(\tilde{x}_R^R > -f_1 \mid \tilde{x}^C) > 0 = \frac{f_2 - \tilde{x}_C^C}{f_1 + f_2}.$$

The infection then takes hold by noting that these conditions must both be satisfied for $\tilde{x}^i$ within an ξ-ball of $(-f_1, f_2)$, with ξ sufficiently small, and so on.

The boundary of the D_{12} region is then defined by the conditions

$$\begin{aligned} \rho_C^R &\leq 1 - \Pr(\text{Strategy 2 dominant for } C \mid \tilde{x}^R = \tilde{x}_*) \\ &\leq 1 - \Pr(p\tilde{x}^C \geq p\tilde{x}_* \mid \tilde{x}^R = \tilde{x}_*) = \frac{f_2 - \tilde{x}_R^R}{f_1 + f_2} \end{aligned}$$

and

$$\rho_R^C \geq \Pr(\text{Strategy 1 dominant for } R \mid \tilde{x}^C = \tilde{x}_*)$$
$$\geq \Pr(p\tilde{x}^R \geq p\tilde{x}_* \mid \tilde{x}^C = \tilde{x}_*) = \frac{f_2 - \tilde{x}_C^C}{f_1 + f_2}.$$

Combining the two conditions for each player, and noting that $\tilde{x}^R = \tilde{x}^C = \tilde{x}_* = (\tilde{x}_R, \tilde{x}_C)$,

$$\Pr(p\tilde{x}^C \geq p\tilde{x}_* \mid \tilde{x}^R = \tilde{x}_*) + \Pr(p\tilde{x}^R \geq p\tilde{x}_* \mid \tilde{x}^C = \tilde{x}_*) = 1 - \frac{f_2 - \tilde{x}_R}{f_1 + f_2} + \frac{f_2 - \tilde{x}_C}{f_1 + f_2},$$

which—by CvD's Lemma 4.1—gives

$$1 = 1 - \frac{\tilde{x}_R - \tilde{x}_C}{f_1 + f_2}$$
$$\Leftrightarrow \tilde{x}_R = \tilde{x}_C$$

as $\psi \to 0$. This is again precisely the threshold where the risk-dominant equilibrium shifts between $(1, 2)$ and $(2, 1)$:

$$(f_2 - x_R)(f_1 + x_C) = (f_1 + x_R)(f_2 - x_C)$$
$$\Leftrightarrow x_R = x_C.$$

A similar argument establishes the iterative dominance of the $(2, 1)$ equilibrium in the complementary region.

The resulting cutoff strategies employed by the players in Γ^ψ as $\psi \to 0$ are depicted in Figure 1(b), the expected fitness function $f(x_i, x_j)$ thus again (well) defined (and continuous for $\psi > 0$).

Proposition 1 *For ψ sufficiently small, the replicator dynamics (2) on $(S, \mathscr{B})$*

are well defined.

Proof. By Lemma 1, the expected fitness function $f(x_i, x_j)$ is well defined for sufficiently small ψ. The range of f is the convex hull of $\{f_{11}, f_{12}, f_{21}, f_{22}\}$. Hence, f is bounded, and the result follows from Oechssler and Riedel's (2001) Theorem 2. ■

3.2 The bias game

Before analyzing the evolution of preference biases, consider first a scenario where each player has an "evolutionary principal" who *chooses* that player's preference bias (prior to the playing of the global payoff game Γ^ψ) with the aim of maximizing the player's *expected fitness*. In this case, there is a meta-game Θ^ψ of preference bias choice—the *bias game*—amongst the evolutionary principals, which is a one-shot simultaneous-move game of complete information. The outcome of this bias game will vary according to the underlying fitness game.

Coordination fitness games Assuming without loss of generality that $f_{11} > f_{22}$, the Nash equilibria of the bias game Θ^ψ for coordination fitness games occupy the shaded regions of Figure 2(a) as $\psi \to 0$. Evidently, there is a set Λ_{11} of bias equilibria—containing an uncountable number of pure- and mixed-strategy equilibria—that lead to efficient coordination on strategy 1, either by making strategy 1 dominant for one or both players, or by ensuring that Pareto dominance entails risk dominance in the coordination game setting. However, there is also a set Λ_{22} of equilibria giving inefficient coordination on strategy 2, due to the familiar Prisoners' Dilemma problem of the suboptimal strategy being dominant for both players in this region. For non-vanishing ψ in Figure 2(b), Λ_{11} has its boundary ψ inside the cutoff functions

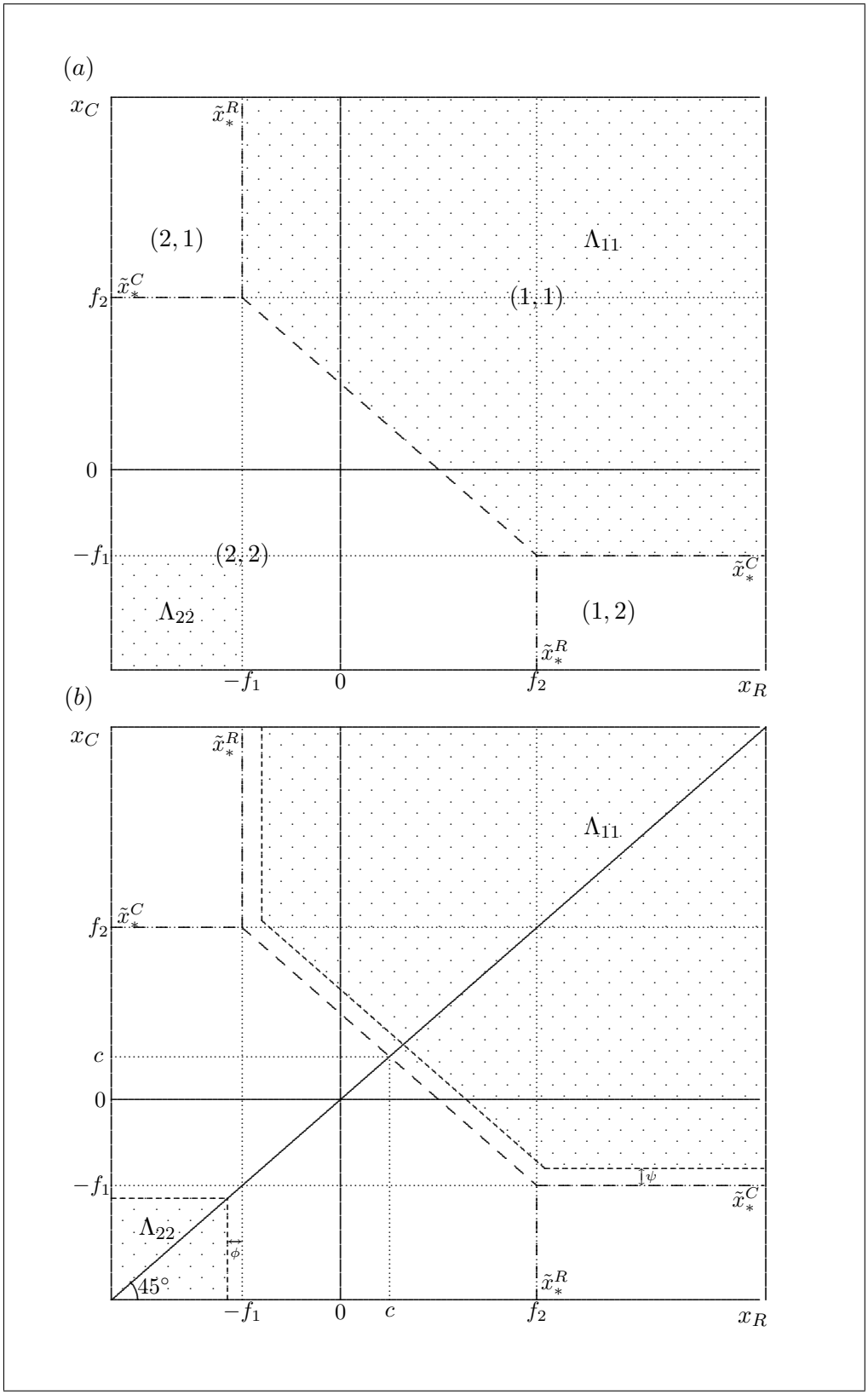

Figure 2: Coordination fitness game, $f_{11} > f_{22}$—(a) $\psi \to 0$; (b) $\psi > 0$

$\tilde{x}^i_*$, in order to guarantee that the $(1,1)$ outcome is played with probability 1. Similarly, Λ_{22} has its boundary $\phi \leq \psi$ inside the $-f_1$ lines, representing the point beyond which one of the players would deviate from a certain f_{22} to some available lottery over outcomes. The symmetric equilibria of Θ^ψ are those that coincide with the 45°-line in Figure 2(b)—i.e. the sets $\Delta_{[c+(\psi/\sqrt{2}),b]}$ and $\Delta_{[a,-f_1-\phi]}$ of all measures on the given intervals, where $c = (f_2 - f_1)/2$.

Hawk-Dove fitness games Assuming without loss of generality that $f_{12} > f_{21}$, the shaded regions of Figure 3(a) contain (uncountably many) Nash equilibria of Θ^ψ for Hawk-Dove fitness games as $\psi \to 0$. Clearly, there is a set Λ_{12} of bias equilibria giving the Row player's preferred outcome $(1,2)$ by making strategy 1 dominant for him. Similarly, there is another set Λ_{21} giving Column's preferred outcome $(2,1)$ by making strategy 1 dominant for him. Preference bias choice here then is essentially an "arms race" to commit to playing strategy 1. For nonvanishing ψ in Figure 3(b), the boundaries of Λ_{12} and Λ_{21} again retreat by ψ or $\phi \leq \psi$. Moreover, depending on the model's parameters, there may exist equilibria in the region $\lambda = (-f_1 - \psi, -f_1 + \phi) \times (-f_1 - \psi, -f_1 + \phi)$, and if so, possibly also a set Λ_ϱ of mixed-strategy equilibria randomizing over the interval $[-f_1 + \phi, b]$ and a subset(s) of $(-f_1 - \psi, -f_1 + \phi)$ with appropriate probabilities ϱ. These are the only possible symmetric equilibria of Θ^ψ.

Prisoners' Dilemma fitness games Assuming without loss of generality that strategy 2 is dominant (i.e. $f_1 < 0, f_2 > 0$), if $f_{11} > f_{22}$ then the dominant-strategy fitness game is inefficient, or a "Prisoners' Dilemma." The equilibria of Θ^ψ for such fitness games differ according to the sign of $f_1 + f_2$, as illustrated in Figures 4 and 5. In either case, there is a set Λ_{22} of bias equilibria that give the usual inefficient $(2,2)$ outcome. However, if—as in Figure 5—$f_1 + f_2 > 0$,

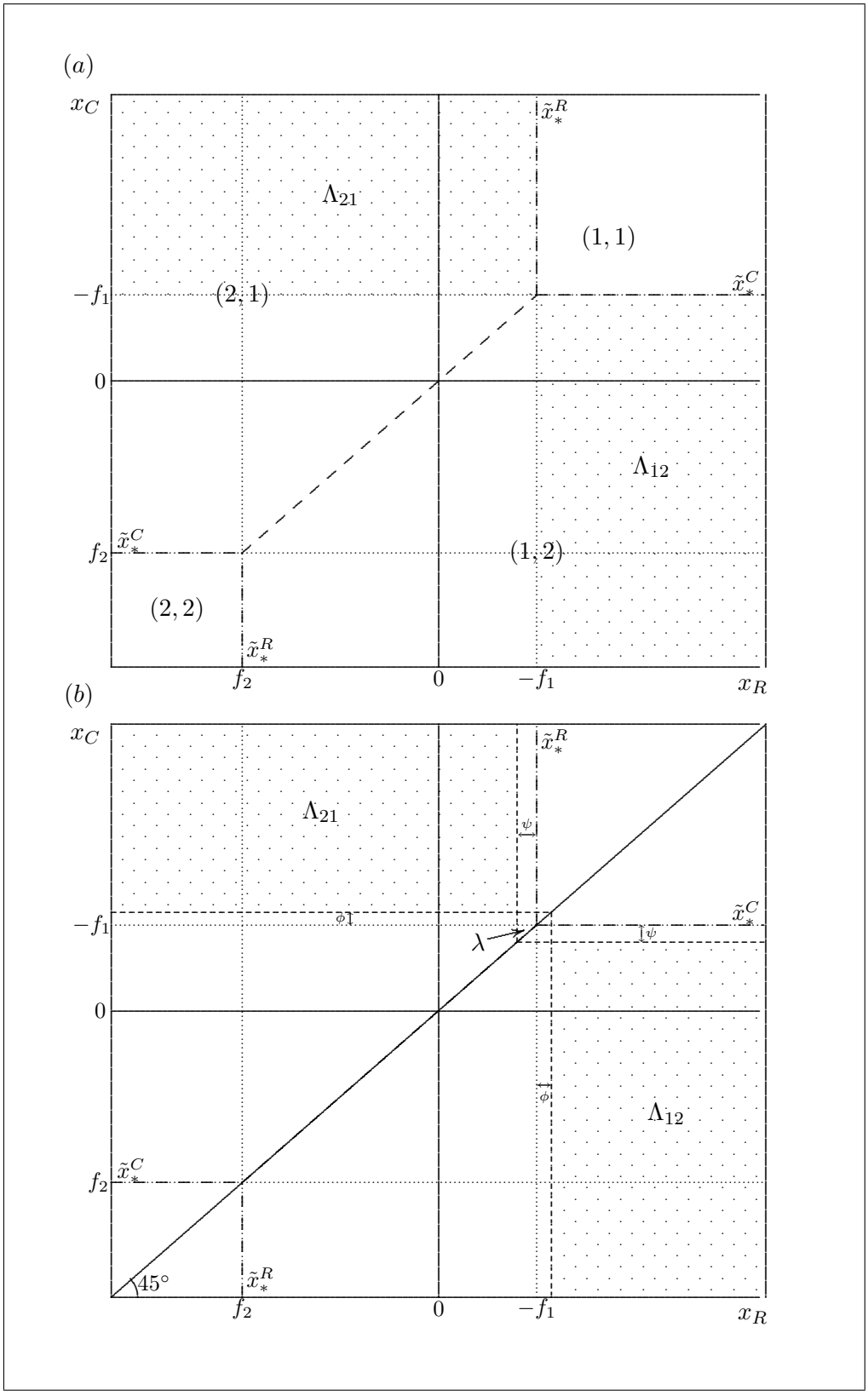

Figure 3: Hawk-Dove fitness game, $f_{12} > f_{21}$—(a) $\psi \to 0$; (b) $\psi > 0$

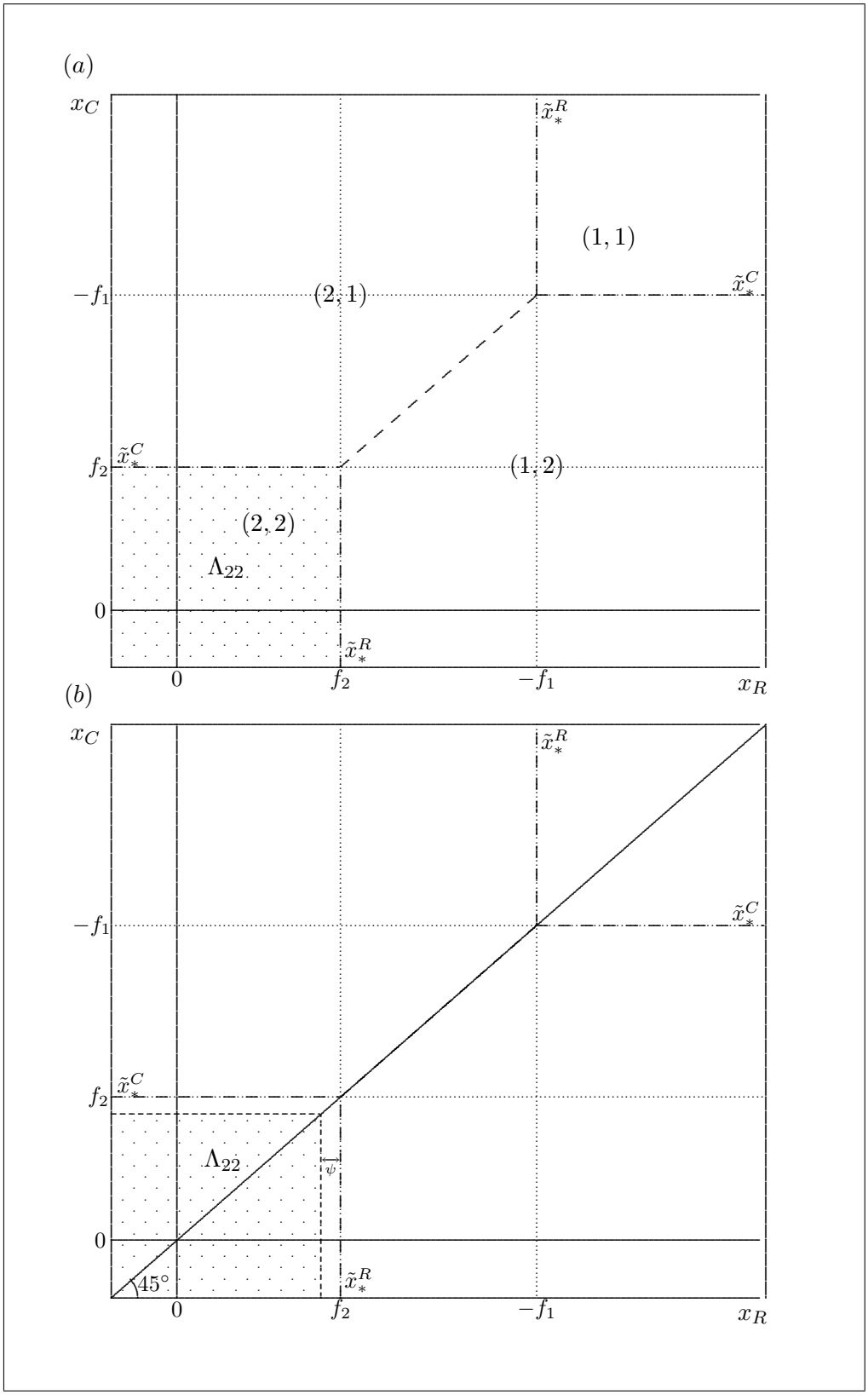

Figure 4: PD fitness game, $f_1 + f_2 < 0$—(a) $\psi \to 0$; (b) $\psi > 0$

then there is also a set Λ_{11} of equilibria supporting efficient coordination on strategy 1, and this set grows with the size of $f_1 + f_2$. Moreover, there is a set Λ_ϱ of mixed-strategy equilibria randomizing over the interval $[c+(\psi/\sqrt{2}), f_2 - \phi']$ and subsets of $(f_2 - \phi', b]$ with appropriate probabilities ϱ, depending on the model's parameters; these equilibria also support efficient coordination on strategy 1. Positive $f_1 + f_2$ corresponds to satisfaction of Kandori and Rob's (1998) "marginal bandwagon property" (MBP), which here captures the feature that the disadvantage of strategy 1 over strategy 2 is minimized when the other player is using strategy 1. This can be interpreted as meaning that the Prisoners' Dilemma is less "bad," in the sense that the efficient outcome $(1, 1)$ is less vulnerable to deviation. The effect of the MBP on the bias game Θ^ψ is to allow the preference biases to transform the Prisoners' Dilemma *fitness* game into a coordination *payoff* game—which is stable when the risk-dominant equilibrium is $(1, 1)$, since the usual unilateral deviations are not available.[6] Thus, preference biases allow the possibility of cooperation in the Prisoners' Dilemma.

3.3 The evolutionary game

Turning now to the *evolution* of preference biases, the "evolutionary principals" of the above bias game are replaced by the replicator dynamics (2) for a population playing that game.

Now, it is well known that symmetric Nash equilibria are (amongst other states) stationary under the replicator dynamics. In particular, if a symmetric Nash equilibrium of the bias game Θ^ψ puts positive probability on a preference bias x, then the pure strategy x must achieve the same (maximal) expected

[6]Notice that the usual unilateral deviations *are* available from the regions where strategy 1 is *dominant* for one or more of the players, so that these regions are unstable.

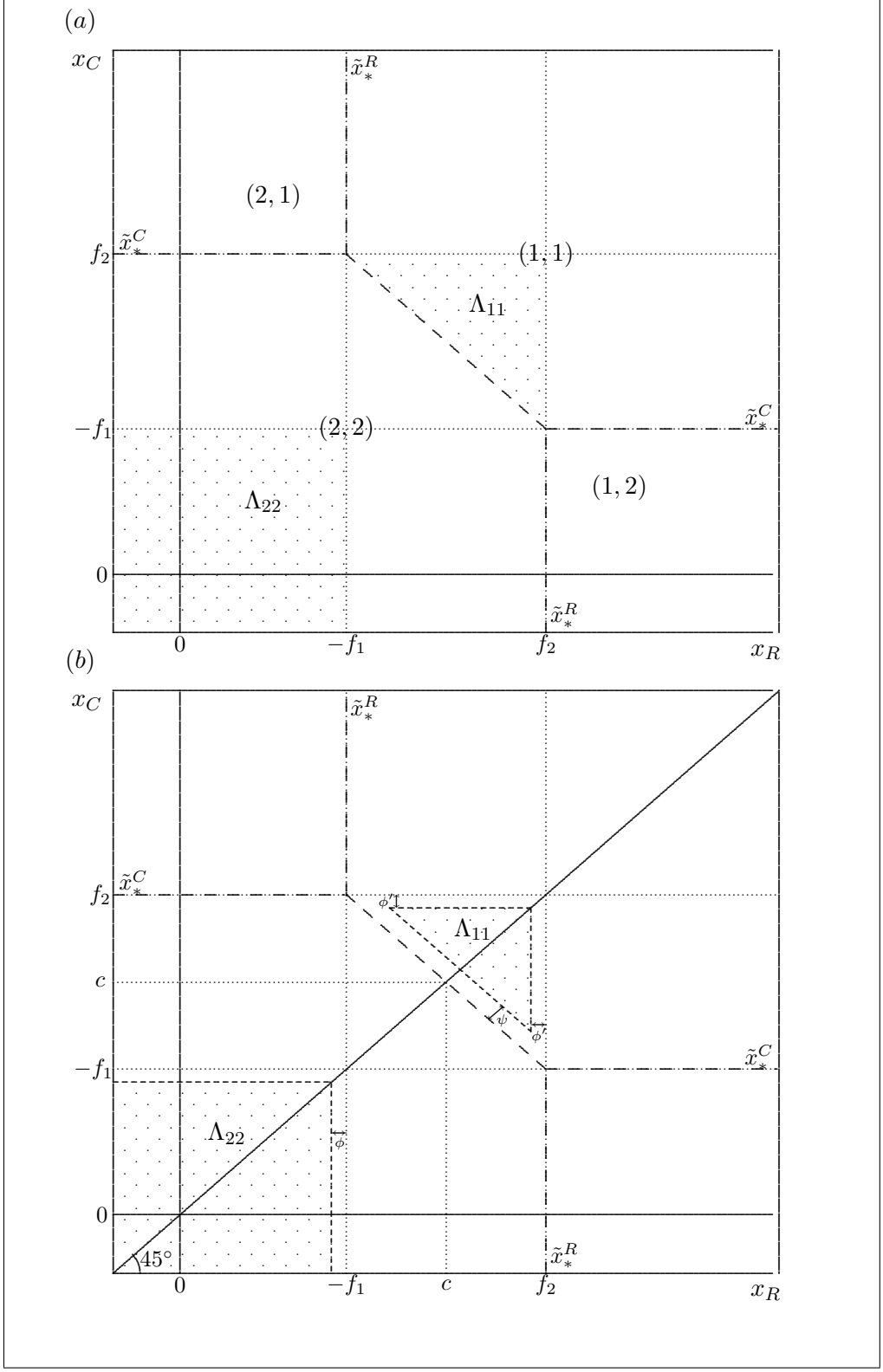

Figure 5: PD fitness game, $f_1 + f_2 > 0$—(a) $\psi \to 0$; (b) $\psi > 0$

fitness as any other pure strategy played with positive probability in that equilibrium. Hence, all pure strategies (preference biases) played with positive probability in a symmetric Nash equilibrium of Θ^{ψ} earn the population average fitness $E(Q,Q)$, and their success (3) is thus 0, giving stationarity by (2).

To arbitrate between the large number of stationary states of the replicator dynamics, it is common to compare their dynamic stability properties in the implicit presence of local mutations. Such an analysis can be conducted in infinite strategy spaces after an appropriate redefinition of "local."

Definition 2 *Let Q^* be a rest point of the replicator dynamics,*

$$\sigma(\cdot, Q^*) = 0 \quad Q^* - a.e.$$

Then

- *Q^* is called* Lyapunov stable *if for all $\varepsilon > 0$ there exists an $\eta > 0$ such that $\|Q(0) - Q^*\| < \eta \Rightarrow \|Q(t) - Q^*\| < \varepsilon$ for all $t > 0$.*
- *Q^* is called* strongly attracting *if there exists $\varepsilon > 0$ such that $\|Q(0) - Q^*\| < \varepsilon \Rightarrow \|Q(t) - Q^*\| \to 0$.*
- *Q^* is called* weakly attracting *if there exists $\varepsilon > 0$ such that $\|Q(0) - Q^*\| < \varepsilon \Rightarrow Q(t) \to Q^*$ in distribution.*

The last two concepts coincide for finite strategy spaces. Together with Lyapunov stability they are called *asymptotic stability.*

Now, whilst there are stationary states of the replicator dynamics that are not symmetric Nash equilibria, no such state satisfies even the weak requirement of Lyapunov stability. This result, due to Bomze (1986) for finite strategy spaces, readily carries over to infinite strategy spaces if f is bounded (so that the bilinear functional $E(P,Q)$ given by (1) is continuous in the variational

topology (Oechssler and Riedel 2001)). Hence, if a population profile of preference biases is Lyapunov stable under the replicator dynamics, then it must be one of the symmetric Nash equilibria of Θ^{ψ} identified in Subsection 3.2. But which one? Indeed, it is unclear whether such a Lyapunov stable population even exists.

With finite strategy spaces, the static criteria of evolutionary and neutral stability are helpful in this regard.

Definition 3 (Maynard-Smith 1974) *A population P is called an* evolutionarily stable state *(ESS) if for every "mutation" Q, there is an invasion barrier $\varepsilon(Q) > 0$ such that for all $0 < \eta \leq \varepsilon$*

$$E(P, (1-\eta)P + \eta Q) > E(Q, (1-\eta)P + \eta Q). \tag{4}$$

Definition 4 (Maynard-Smith 1982) *A population P is called a* neutrally stable state *(NSS) if for every "mutation" Q, there is an invasion barrier $\varepsilon(Q) > 0$ such that for all $0 < \eta \leq \varepsilon$*

$$E(P, (1-\eta)P + \eta Q) \geq E(Q, (1-\eta)P + \eta Q).$$

In the finite case, every ESS is asymptotically stable in the replicator dynamics (Taylor and Jonker 1978, Hofbauer, Schuster, and Sigmund 1979), and every NSS is Lyapunov stable (Thomas 1985, Bomze and Weibull 1995).[7] However, in infinite strategy spaces, Oechssler and Riedel (2001) show that the former result no longer holds in general, so that stronger concepts than ESS are necessary.

Definition 5 (Vickers and Cannings 1987) *A population P is called* un-

[7]For a review of results in finite strategy spaces, see Weibull (1995).

invadable *if there is a uniform invasion barrier, that is, an* $\varepsilon > 0$ *such that (4) holds for all* Q *and all* $0 < \eta \leq \varepsilon$.

Uninvadability implies evolutionary stability, merely requiring a *uniform* invasion barrier against mutations amongst a small fraction of the population. If mutations amongst the *whole* population to "nearby" populations are allowed, the following even stronger definition is reached.

Definition 6 (Bomze 1990) *A population* P *is called* strongly uninvadable *if there is a barrier* $\varepsilon > 0$ *such that for all populations* $R \neq P$ *with* $\|R-P\| \leq \varepsilon$, *we have*

$$E(P, R) > E(R, R).$$

Strong uninvadability implies uninvadability, and if the equilibrium population state is homogeneous, then the criteria coincide (Oechssler and Riedel 2001). Oechssler and Riedel's (2001) Theorem 3 shows that every uninvadable, homogeneous population is Lyapunov stable, and—if f is continuous—weakly attracting.

Unfortunately, this result is of limited use in analyzing the bias game Θ^{ψ} of Subsection 3.2. If there is a symmetric, uninvadable pure-strategy equilibrium in the region λ for the Hawk-Dove fitness game (Figure 3), then it is clearly Lyapunov stable and weakly attracting. But none of the symmetric Nash equilibria of the coordination and Prisoners' Dilemma fitness games are uninvadable. This is because they all belong to regions of fitness-equivalent equilibria, so that any such population P has a "nearby" population $(1-\eta)P + \eta Q$ such that $E(P, (1-\eta)P + \eta Q) = E(Q, (1-\eta)P + \eta Q)$, for η arbitrarily small. However, it seems likely that some of these regions will have some evolutionary stability property, suggesting a switch of focus to setwise generalizations of the static stability criteria.

Definition 7 (Thomas 1985) *A set of populations* $\Pi \subset \Delta$ *is an* evolutionarily stable *(ES) set if and only if it is nonempty and closed, and each* $P \in \Pi$ *has some neighborhood* U *such that* $E(P,Q) \geq E(Q,Q)$ *for all* $Q \in U$*, with strict inequality if* $Q \notin \Pi$*.*

Clearly if a population P belongs to an ES set, then it is neutrally stable.

To investigate the dynamic stability properties of such sets, it is first necessary to generalize Definition 2 to apply to sets of states.

Definition 8 *Let* $\Pi \subset \Delta$ *be a closed set of states (populations). Then*

- Π *is called* Lyapunov stable *if for all* $\varepsilon > 0$ *there exists an* $\eta > 0$ *such that* $\inf_{R\in\Pi} \|Q(0) - R\| < \eta \Rightarrow \inf_{R\in\Pi} \|Q(t) - R\| < \varepsilon$ *for all* $t > 0$*.*

- Π *is called* strongly attracting *if there exists* $\varepsilon > 0$ *such that* $\inf_{R\in\Pi} \|Q(0) - R\| < \varepsilon \Rightarrow \|Q(t) - R'\| \to 0$*, for some* $R' \in \Pi$*.*

- Π *is called* weakly attracting *if there exists* $\varepsilon > 0$ *such that* $\inf_{R\in\Pi} \|Q(0) - R\| < \varepsilon \Rightarrow Q(t) \to R'$ *in distribution, for some* $R' \in \Pi$*.*

In finite strategy spaces, every ES set is asymptotically stable (Weibull 1995), suggesting that the following concept may be of use in the infinite case.

Definition 9 *A set of populations* $\Pi \subseteq \Delta$ *is a* strongly uninvadable (SU) set *if and only if it is nonempty and closed, and there is a barrier* $\varepsilon > 0$ *such that each* $P \in \Pi$ *has* $E(P,Q) \geq E(Q,Q)$ *for all* Q *within the* ε*-neighborhood of* P*, with strict inequality if* $Q \notin \Pi$*.*

Here, of course, the ε-neighborhood of P is defined on the variational topology. The following result unsurprisingly holds.

Proposition 2 *If* $P \in \Delta$ *belongs to a strongly uninvadable set* $\Pi \subseteq \Delta$*, then* P *is a Nash equilibrium.*

Proof. Suppose otherwise. Then there exists some $R_b \in \Delta$ such that $E(R_b, P) > E(P, P)$. Letting $Q = (1-\eta)P + \eta R_b$, clearly $\|P - Q\| \leq \eta$. Equation (1) then gives

$$\begin{aligned} E(P,Q) &= \int_S \int_S f(x,y)((1-\eta)P + \eta R_b)(dy)P(dx) \\ &= (1-\eta)\int_S \int_S f(x,y)P(dy)P(dx) + \eta \int_S \int_S f(x,y)R_b(dy)P(dx) \\ &= (1-\eta)E(P,P) + \eta E(P,R_b) \end{aligned}$$

and

$$\begin{aligned} E(Q,Q) &= \int_S \int_S f(x,y)((1-\eta)P + \eta R_b)(dy)((1-\eta)P + \eta R_b)(dx) \\ &= (1-\eta)^2 \int_S \int_S f(x,y)P(dy)P(dx) \\ &\quad + \eta(1-\eta)\left(\int_S \int_S f(x,y)P(dy)R_b(dx) + \int_S \int_S f(x,y)R_b(dy)P(dx)\right) \\ &\quad + \eta^2 \int_S \int_S f(x,y)R_b(dy)R_b(dx) \\ &= (1-\eta)^2 E(P,P) + \eta(1-\eta)\left(E(R_b,P) + E(P,R_b)\right) + \eta^2 E(R_b,R_b). \end{aligned}$$

Hence,

$$\begin{aligned} E(P,Q) - E(Q,Q) &= \eta\left(E(P,P) - E(R_b,P)\right) \\ &\quad - \eta^2\left(E(P,P) - E(R_b,P) + E(R_b,R_b) - E(P,R_b)\right), \end{aligned}$$

which is negative for η sufficiently small, in violation of the strong uninvadability of Π. ∎

The concept of a strongly uninvadable set allows the following natural generalization of Oechssler and Riedel's (2001) Theorem 3.

Theorem 1 *If $\Delta_{[\alpha,\beta]} \subseteq \Delta$ is a strongly uninvadable set, then*

- *$\Delta_{[\alpha,\beta]}$ is Lyapunov stable;*
- *if additionally f is continuous, then $\Delta_{[\alpha,\beta]}$ is weakly attracting.*

Proof. Consider $Q(0) \notin \Delta_{[\alpha,\beta]}$ such that $\inf_{R\in\Delta_{[\alpha,\beta]}} \|Q(0) - R\| \leq \varepsilon$, where ε is $\Delta_{[\alpha,\beta]}$'s strong uninvadability barrier. From equations (2) and (3),

$$\begin{aligned} Q'(0)([\alpha,\beta]) &= \int_{[\alpha,\beta]} \left(E(\delta_x, Q(0)) - E(Q(0), Q(0))\right) Q(0)(dx) \\ &= \int_{[\alpha,\beta]} E(\delta_x, Q(0)) Q(0)(dx) - \int_{[\alpha,\beta]} E(Q(0), Q(0)) Q(0)(dx) \\ &= \int_{[\alpha,\beta]} \int_S f(x,y) Q(0)(dy) Q(0)(dx) - E(Q(0), Q(0)) \int_{[\alpha,\beta]} Q(0)(dx), \end{aligned}$$

since $E(\delta_x, Q(0)) = \int_S f(x,y) Q(0)(dy)$ by equation (1). Construct the measure R^* given by $Q(0)(\{x\})/Q(0)([\alpha,\beta])$ for $x \in [\alpha,\beta]$ and 0 otherwise. Evidently, $R^* \in \Delta_{[\alpha,\beta]}$, and

$$\begin{aligned} \|Q(0) - R^*\| &= 2 \sup_{A\in\mathscr{B}} |Q(0)(A) - R^*(A)| \\ &= 2\, Q(0)(S \backslash [\alpha,\beta]) \\ &= \inf_{R\in\Delta_{[\alpha,\beta]}} \|Q(0) - R\|. \end{aligned}$$

Hence $\|Q(0) - R^*\| \leq \varepsilon$, and

$$\begin{aligned} Q'(0)([\alpha,\beta]) &= Q(0)([\alpha,\beta]) \int_S \int_S f(x,y) Q(0)(dy) R^*(dx) \\ &\qquad - Q(0)([\alpha,\beta]) E(Q(0), Q(0)) \\ &= Q(0)([\alpha,\beta]) \left(E(R^*, Q(0)) - E(Q(0), Q(0))\right) \\ &> 0 \end{aligned}$$

by strong uninvadability of $\Delta_{[\alpha,\beta]}$. The same argument applies to $Q(t) \notin \Delta_{[\alpha,\beta]}$ such that $\inf_{R\in\Delta_{[\alpha,\beta]}} \|Q(t)-R\| \leq \varepsilon$, and hence $Q(t)([\alpha,\beta])$ is strictly increasing until $Q(t) \in \Delta_{[\alpha,\beta]}$, given continuity of f.[8]

If $Q(t) \in \Delta_{[\alpha,\beta]}$, then $Q(t)$ is Nash by Proposition 2. It follows that $E(\delta_x, Q(t)) = E(Q(t), Q(t))$ for all x in the support of $Q(t)$, so that $\sigma(x, Q(t)) = 0$ and hence $Q'(t)(\{x\}) = 0$ from equation (2). For x not in the support of $Q(t)$ meanwhile, $Q(t)(\{x\}) = 0$, so that again $Q'(t)(\{x\}) = 0$.

■

This result may be of general interest, but it is also useful in arbitrating between the symmetric Nash equilibria of the bias game Θ^ψ in Subsection 3.2—at least for coordination fitness games.

Corollary 1 *For coordination fitness games, the set of preference biases supporting coordination on the Pareto-dominant equilibrium is Lyapunov stable and weakly attracting.*

Proof. $\Delta_{[c+(\psi/\sqrt{2}),b]}$ is nonempty for $\psi \leq \sqrt{2}(b-c)$.

Evidently, $\inf_{R\in\Delta_{[c+(\psi/\sqrt{2}),b]}} \|P-R\| > 0$ for all $P \in \Delta\backslash\Delta_{[c+(\psi/\sqrt{2}),b]}$. Letting $\xi(P) = (\inf_{R\in\Delta_{[c+(\psi/\sqrt{2}),b]}} \|P-R\|)/2$, it is clear that $B_{\xi(P)}(P) \subset \Delta\backslash\Delta_{[c+(\psi/\sqrt{2}),b]}$ for all $P \in \Delta\backslash\Delta_{[c+(\psi/\sqrt{2}),b]}$. Hence, $\Delta\backslash\Delta_{[c+(\psi/\sqrt{2}),b]}$ is open in the variational topology, and $\Delta_{[c+(\psi/\sqrt{2}),b]}$ is closed.

Now, for any $P \in \Delta_{[c+(\psi/\sqrt{2}),b]}$, and any $Q \notin \Delta_{[c+(\psi/\sqrt{2}),b]}$ such that $2|P([a,c+(\psi/\sqrt{2})))-Q([a,c+(\psi/\sqrt{2})))| = \|P-Q\| = \varepsilon$, $E(P,Q) > E(Q,Q)$

[8] See Oechssler and Riedel (2001), pp. 149–150, for the necessity of continuous f for weak attractiveness.

if and only if

$$\int_{[a,b]} \left(\int_{[c+(\psi/\sqrt{2}),b]} f(x,y)Q(dy) + \int_{[a,c+(\psi/\sqrt{2}))} f(x,y)Q(dy) \right) P(dx) > \int_{[a,b]} \left(\int_{[c+(\psi/\sqrt{2}),b]} f(x,y)Q(dy) + \int_{[a,c+(\psi/\sqrt{2}))} f(x,y)Q(dy) \right) Q(dx). \quad (5)$$

Note that

$$\int_{[c+(\psi/\sqrt{2}),b]} f(x,y)Q(dy) = (1-k) \int_{[a,b]} f(x,y)R_1(dy),$$

where $k = \varepsilon/2$, and $R_1(\{x\}) = Q(\{x\})/(1-k)$ for $x \in [c + (\psi/\sqrt{2}), b]$ and 0 otherwise. Similarly,

$$\int_{[a,c+(\psi/\sqrt{2}))} f(x,y)Q(dy) = k \int_{[a,b]} f(x,y)R_2(dy),$$

where $R_2(\{x\}) = Q(\{x\})/k$ for $x \notin [c + (\psi/\sqrt{2}), b]$ and 0 otherwise. Hence, (5) holds if and only if

$$\begin{aligned}(1-k) \int_{[a,b]} \int_{[a,b]} f(x,y)R_1(dy)P(dx) + k \int_{[a,b]} \int_{[a,b]} f(x,y)R_2(dy)P(dx) > \\ (1-k) \int_{[c+(\psi/\sqrt{2}),b]} \int_{[a,b]} f(x,y)R_1(dy)Q(dx) \\ + (1-k) \int_{[a,c+(\psi/\sqrt{2}))} \int_{[a,b]} f(x,y)R_1(dy)Q(dx) \\ + k \int_{[c+(\psi/\sqrt{2}),b]} \int_{[a,b]} f(x,y)R_2(dy)Q(dx) \\ + k \int_{[a,c+(\psi/\sqrt{2}))} \int_{[a,b]} f(x,y)R_2(dy)Q(dx)\end{aligned}$$

$$\Leftrightarrow (1-k)f_{11} + kE(P,R_2) >$$
$$(1-k)^2 \int_{[a,b]} \int_{[a,b]} f(x,y)R_1(dy)R_1(dx)$$
$$+ k(1-k) \int_{[a,b]} \int_{[a,b]} f(x,y)R_1(dy)R_2(dx)$$
$$+ k(1-k) \int_{[a,b]} \int_{[a,b]} f(x,y)R_2(dy)R_1(dx)$$
$$+ k^2 \int_{[a,b]} \int_{[a,b]} f(x,y)R_2(dy)R_2(dx)$$

$$\Leftrightarrow (1-k)f_{11} + kE(P,R_2) >$$
$$(1-k)^2 f_{11} + k(1-k)(E(R_2,R_1) + E(R_1,R_2)) + k^2 E(R_2,R_2)$$

$$\Leftrightarrow (1-k)(f_{11} - E(R_2,R_1)) >$$
$$k(E(R_2,R_2) - E(R_1,R_2)) + (E(R_1,R_2) - E(P,R_2)).$$

Since $f_{11} > E(R_2,R_1)$ and $E(R_1,R_2) \to E(P,R_2)$ as $\varepsilon \to 0$, this must hold for ε sufficiently small. This is true *a fortiori* if $2|P([a,c+(\psi/\sqrt{2}))) - Q([a,c+(\psi/\sqrt{2})))| < \|P-Q\| = \varepsilon$, since $k < \varepsilon/2$ in that case. For $Q \in \Delta_{[c+(\psi/\sqrt{2}),b]}$, meanwhile, $E(P,Q) = E(Q,Q) = f_{11}$.

$\Delta_{[c+(\psi/\sqrt{2}),b]}$ is thus a strongly uninvadable set, and is hence Lyapunov stable and weakly attracting (given continuity of the expected fitness function for $\psi > 0$) by Theorem 1. Any population belonging to this set coordinates on strategy 1, giving the Pareto-dominant equilibrium of the underlying fitness game. ■

The replicator dynamics thus support preference biases in favor of the efficient strategy in coordination fitness games, making it risk dominant or dominant

and hence giving efficient coordination. Moreover, biases in favor of the *inefficient* strategy fail to receive such support.

Proposition 3 *For coordination fitness games, the set of preference biases supporting coordination on the Pareto-dominated equilibrium is not Lyapunov stable.*

Proof. Consider a population Q such that $\inf_{R\in\Delta_{[a,-f_1-\phi]}} \|Q-R\| = \varepsilon$ and $Q([c+(\psi/\sqrt{2}), f_2-\psi]) = k = \varepsilon/2$. Clearly

$$\begin{aligned} E(\delta_x, Q) &= f_{22} \\ &< (1-k^2)f_{22} + k^2 f_{11} \\ &= E(Q,Q) \end{aligned}$$

for all $x \in [a, -f_1-\phi]$ and $\varepsilon > 0$. Hence, $\Delta_{[a,-f_1-\phi]}$ is not strongly uninvadable, and moreover, $\sigma(x,Q) < 0$ for all $x \in [a, -f_1-\phi]$, so that if $Q(t) = Q$, $Q'(t)([a,-f_1-\phi]) < 0$. Thus, $\Delta_{[a,-f_1-\phi]}$ is not Lyapunov stable. ■

Interestingly, the equilibrium selection in the global payoff game Γ^ψ plays a key role in the instability of the populations supporting the Pareto-dominated equilibrium, the efficient coordination engendered by preference biases in the region $[c+(\psi/\sqrt{2}), f_2-\psi] \times [c+(\psi/\sqrt{2}), f_2-\psi]$ providing the destabilizing mutation.

Turning to Prisoners' Dilemma fitness games, the analysis is a little more involved. If the MBP fails, $\Delta_{[a,f_2-\psi]}$ is strongly uninvadable and hence Lyapunov stable and weakly attracting by Theorem 1, with the usual inefficient $(2,2)$ outcome resulting. But if the MBP is satisfied, then there are no longer any strongly uninvadable sets, and a weaker setwise stability criterion is required.

Definition 10 *A set of populations* $\Pi \subset \Delta$ *is a* neutrally uninvadable (NU) set *if and only if it is nonempty and closed, and there is a barrier* $\varepsilon > 0$ *such that each* $P \in \Pi$ *has* $E(P,Q) \geq E(Q,Q)$ *for all* Q *within the* ε*-neighborhood of* P*.*

Clearly Proposition 2 carries over to neutrally uninvadable sets, but Theorem 1 is weakened in the following manner.

Theorem 2 *If* $\Delta_{[\alpha,\beta]} \subseteq \Delta$ *is a neutrally uninvadable set, then* $\Delta_{[\alpha,\beta]}$ *is Lyapunov stable.*

Proof. Consider $Q(t) \in \Delta$ such that $\inf_{R \in \Delta_{[\alpha,\beta]}} \|Q(t) - R\| \leq \varepsilon$, where ε is $\Delta_{[\alpha,\beta]}$'s neutral uninvadability barrier. Proceeding in the manner of the proof of Theorem 1,

$$Q'(0)([\alpha,\beta]) = Q(0)([\alpha,\beta])\left(E(R^*, Q(0)) - E(Q(0), Q(0))\right),$$

which is nonnegative by neutral uninvadability of $\Delta_{[\alpha,\beta]}$. The same argument applies to $Q(t) \in \Delta$ such that $\inf_{R \in \Delta_{[\alpha,\beta]}} \|Q(t) - R\| \leq \varepsilon$, and hence $Q(t)([\alpha,\beta])$ is nondecreasing in t. ■

Corollary 2 *For Prisoners' Dilemma fitness games satisfying the MBP, a set of preference biases supporting efficient cooperation is Lyapunov stable.*

Proof. $\Delta_{[c+(\psi/\sqrt{2}),f_2-\phi']}$ is nonempty for $\psi \leq \sqrt{2}(f_2 - \phi' - c)$.

Evidently, $\inf_{R \in \Delta_{[c+(\psi/\sqrt{2}),f_2-\phi']}} \|P - R\| > 0$ for all $P \in \Delta \backslash \Delta_{[c+(\psi/\sqrt{2}),f_2-\phi']}$. Letting $\xi(P) = (\inf_{R \in \Delta_{[c+(\psi/\sqrt{2}),f_2-\phi']}} \|P - R\|)/2$, it is clear that $B_{\xi(P)}(P) \subset \Delta \backslash \Delta_{[c+(\psi/\sqrt{2}),f_2-\phi']}$ for all $P \in \Delta \backslash \Delta_{[c+(\psi/\sqrt{2}),f_2-\phi']}$. Hence, $\Delta \backslash \Delta_{[c+(\psi/\sqrt{2}),f_2-\phi']}$ is open in the variational topology, and $\Delta_{[c+(\psi/\sqrt{2}),f_2-\phi']}$ is closed.

Now, proceeding as in the proof of Corollary 1, for any $P \in \Delta_{[c+(\psi/\sqrt{2}),f_2-\phi']}$, and any $Q \notin \Delta_{[c+(\psi/\sqrt{2}),f_2-\phi']}$ such that $2|P([a,b] \backslash [c + (\psi/\sqrt{2}), f_2 - \phi']) -$

$Q([a,b]\backslash[c+(\psi/\sqrt{2}), f_2-\phi'])| = \|P-Q\| = \varepsilon$, $E(P,Q) \geq E(Q,Q)$ if and only if

$$(1-k)(f_{11} - E(R_2,R_1)) \geq$$
$$k(E(R_2,R_2) - E(R_1,R_2)) + (E(R_1,R_2) - E(P,R_2)). \quad (6)$$

Clearly $f_{11} \geq E(R_2,R_1)$, with equality if and only if R_2 has support confined to $(f_2-\phi',b]$, in which case $E(R_2,R_1) = E(R_2,R_2) = E(R_1,R_2)) = E(P,R_2) = f_{11}$, giving 0 on the RHS of (6). With $f_{11} > E(R_2,R_1)$, and given that $E(R_1,R_2) \to E(P,R_2)$ as $\varepsilon \to 0$, (6) must hold for ε sufficiently small. These arguments are true *a fortiori* if $2|P([a,b]\backslash[c+(\psi/\sqrt{2}), f_2-\phi']) - Q([a,b]\backslash[c+(\psi/\sqrt{2}), f_2-\phi'])| < \|P-Q\| = \varepsilon$, since $k < \varepsilon/2$ in that case. For $Q \in \Delta_{[c+(\psi/\sqrt{2}),f_2-\phi']}$, meanwhile, $E(P,Q) = E(Q,Q) = f_{11}$.

$\Delta_{[c+(\psi/\sqrt{2}),f_2-\phi']}$ is thus a neutrally uninvadable set, and is hence Lyapunov stable by Theorem 2. Any population belonging to this set coordinates on strategy 1, giving efficient cooperation in the underlying fitness game. ■

Moreover, the usual inefficient Prisoners' Dilemma outcome is *unstable* in this case.

Proposition 4 *For Prisoners' Dilemma fitness games satisfying the MBP, the set of preference biases supporting the inefficient outcome is not Lyapunov stable.*

Proof. Consider a population Q such that $\inf_{R\in\Delta_{[a,-f_1-\phi]}} \|Q-R\| = \varepsilon$ and

$Q([c + (\psi/\sqrt{2}), f_2 - \phi']) = k = \varepsilon/2$. Clearly

$$\begin{aligned} E(\delta_x, Q) &= f_{22} \\ &< (1-k^2)f_{22} + k^2 f_{11} \\ &= E(Q, Q) \end{aligned}$$

for all $x \in [a, -f_1 - \phi]$ and $\varepsilon > 0$. Hence, $\Delta_{[a,-f_1-\phi]}$ is not neutrally uninvadable, and moreover, $\sigma(x, Q) < 0$ for all $x \in [a, -f_1 - \phi]$, so that if $Q(t) = Q$, $Q'(t)([a, -f_1 - \phi]) < 0$. Thus, $\Delta_{[a,-f_1-\phi]}$ is not Lyapunov stable. ■

Remark Note that some of the mixed-strategy equilibria Λ_ϱ may also belong to a Lyapunov stable set, but that they too support efficient cooperation.

In a random matching setting, the marginal bandwagon property $f_1 + f_2 > 0$ implies that the marginal loss of switching to strategy 1 is a decreasing function of the number of strategy-1 users in the population. When Prisoners' Dilemmas are less "bad" in this sense then, Corollary 2 establishes the stability of preferences supporting cooperation. And this result has less of a "0–1" character than it might first appear, since the region of preferences supporting cooperation grows with $f_1 + f_2$—i.e. as the MBP is more strongly satisfied.

Remark A straightforward application of Theorem 2 is that every neutrally uninvadable, homogeneous population is Lyapunov stable. This result is a natural counterpart to Oechssler and Riedel's (2001) result that every uninvadable, homogeneous population is Lyapunov stable and—if f is continuous—weakly attracting.

As for Hawk-Dove games, if there is no symmetric, uninvadable pure-strategy equilibrium, it is difficult to say much in general about stability. De-

pending on the model's parameters, there may exist a (nontrivial) Lyapunov stable set of populations, the elements of which may or may not be stationary states.

Remark Note that it is quite possible for the measure on which the system comes to rest to give zero weight to the dominant-strategy biases required to begin the global game iterative dominance arguments of Subsection 3.1, so that Lemma 1 would fail. To resolve this difficulty, some noise in the players' observation of the measure could be introduced, or some continuous (small) rate of mutation away from the deterministic dynamic (as is in any case implicit in the dynamic stability analysis).

4 Conclusion

The observability of preferences is a key factor in models of indirect evolution. By building observability explicitly into the model in a global-game-theoretic manner, the multiple equilibria obstacle to a full dynamic analysis is removed. And by employing the replicator dynamics for infinite strategy spaces, the evolution of the necessarily *continuous* preferences can be examined. With the aid of new setwise stability criteria, the stable preference biases for coordination and certain Prisoners' Dilemma fitness games are seen to be gratifyingly efficient. Hawk-Dove fitness games, meanwhile, are a potentially more unstable "arms race" in biases.

These results of course require the degree of observability necessary to deliver equilibrium selection in the global payoff game, begging the question of what happens with less observable preferences. One answer is that multiple Bayesian equilibria return as usual, so that the replicator dynamics cease to be well defined. Alternatively, one could consider the scenario where underlying

fitnesses are not observed, so that an agent's own bias is informative about that of his opponent (and vice versa). As long as each agent observes his own bias with sufficient accuracy, global game cutoff strategies will be employed, and as his opponent's bias becomes unobservable, an agent's cutoff will condition solely on his own bias.[9] In this case, the structure of the bias game is essentially that of the underlying fitness game, so that stability of efficient cooperation in the Prisoners' Dilemma, for example, breaks down. Observability thus assumes its usual important role.

But the approach is not restricted to preference *biases*. The global games framework permits a quite general parameter space to be mapped into games, so that essentially any function from fitnesses to payoffs could in theory give well-defined replicator dynamics. In practice, the complexities of the analysis restrict attention to simple functions such as preference biases, but this very simplicity is perhaps appealing, in accord with the idea of emotions as shortcuts adopted by Nature allowing people to economize on reasoning resources.[10]

Indeed, the approach might even be used to adjudicate between alternative functional forms for preferences. For instance, one could posit a process of "three-speed evolution," whereby the parameter space of preferences (e.g. bias space) adjusts very slowly, the parameters themselves (e.g. biases) adjust somewhat more quickly, and behavior adjusts very quickly. The meta-meta-game in the choice of parameter space could then be studied to reveal the stable functional forms for preferences.

[9] See Carlsson and van Damme's (1993) Appendix B, and Morris and Shin's (2004) Subsection 2.3.

[10] See, for example, Samuelson (2001a). For a review of the importance of emotions in economic theory, see Elster (1998).

References

Benos, A. V. (1998): "Aggressiveness and Survival of Overconfident Traders," *Journal of Financial Markets*, 1, 353–383.

Bergman, N., and Y. Bergman (2000): "Ecologies of Preferences with Envy as an Antidote to Risk-Aversion in Bargaining," Working Paper, The Hebrew University of Jerusalem.

Bester, H., and W. Güth (1998): "Is Altruism Evolutionary Stable?," *Journal of Economic Behavior and Organization*, 34, 193–209.

Bolle, F. (2000): "Is Altruism Evolutionarily Stable? And Envy and Malevolence? - Remarks on Bester and Güth," *Journal of Economic Behavior and Organization*, 42, 131–133.

Bomze, I. (1986): "Non-Cooperative Two-Person Games in Biology: A Classification," *International Journal of Game Theory*, 15, 31–57.

——— (1990): "Dynamical Aspects of Evolutionary Stability," *Monatshefte für Mathematik*, 110, 189–206.

——— (1991): "Cross Entropy Minimization in Uninvadable States of Complex Populations," *Journal of Mathematical Biology*, 30, 73–87.

Bomze, I., and J. Weibull (1995): "Does Neutral Stability Imply Lyapunov Stability?," *Games and Economic Behavior*, 11, 173–192.

Carlsson, H., and E. van Damme (1993): "Global Games and Equilibrium Selection," *Econometrica*, 61, 989–1018.

Dekel, E., J. C. Ely, and O. Yilankaya (2004): "Evolution of Preferences," Working Paper, Northwestern University.

Elster, J. (1998): "Emotions and Economic Theory," *Journal of Economic Literature*, XXXVI, 47–74.

Ely, J. C., and O. Yilankaya (2001): "Nash Equilibrium and the Evolution of Preferences," *Journal of Economic Theory*, 97, 255–272.

Eshel, I., L. Samuelson, and A. Shaked (1988): "Altruists, Egoists and Hooligans in a Local Interaction Model," *American Economic Review*, 88, 157–179.

Fershtman, C., and Y. Weiss (1997): "Why Do We Care About What Others Think About Us?," in *Economics, Values and Organization*, ed. by A. B. Ner, and L. Putterman. Cambridge University Press, Cambridge, Massachusetts.

——— (1998): "Social Rewards, Externalities and Stable Preferences," *Journal of Public Economics*, 70, 53–73.

Frank, R. H. (1987): "If *Homo Economicus* Could Choose His Own Utility Function, Would He Want One with a Conscience?," *American Economic Review*, 77, 593–604.

Güth, W. (1995): "An Evolutionary Approach to Explaining Cooperative Behavior by Reciprocal Incentives," *International Journal of Game Theory*, 24, 323–344.

Güth, W., and B. Peleg (2001): "When Will Payoff Maximization Survive? An Indirect Evolutionary Analysis," *Journal of Evolutionary Economics*, 11, 479–499.

Güth, W., and M. E. Yaari (1992): "Explaining Reciprocal Behavior in Simple Strategic Games: An Evolutionary Approach," in *Explaining Process*

and Change, ed. by U. Witt. University of Michigan Press, Ann Arbor, Michigan.

Guttman, J. M. (2000): "On the Evolutionary Stability of Preferences for Reciprocity," *European Journal of Political Economy*, 16, 31–50.

Heifetz, A., and E. Segev (2003): "The Evolutionary Role of Toughness in Bargaining," Working Paper, Tel Aviv University, forthcoming *Games and Economic Behavior*.

Heifetz, A., C. Shannon, and Y. Spiegel (2003): "What to Maximize if You Must," Working Paper, University of California, Berkeley.

Hofbauer, J., P. Schuster, and K. Sigmund (1979): "A Note on Evolutionary Stable Strategies and Game Dynamics," *Journal of Theoretical Biology*, 81, 609–612.

Huck, S., G. Kirchsteiger, and J. Oechssler (1999): "Learning to Like What You Have: Explaining the Endowment Effect," Working Paper, Humboldt University Berlin and University of Vienna, forthcoming *Economic Journal*.

Huck, S., and J. Oechssler (1999): "The Indirect Evolutionary Approach to Explaining Fair Allocations," *Games and Economic Behavior*, 28, 13–24.

Kandori, M., and R. Rob (1998): "Bandwagon Effects and Long Run Technology Choice," *Games and Economic Behavior*, 22, 30–60.

Koçkesen, L., E. A. Ok, and R. Sethi (2000a): "Evolution of Interdependent Preferences in Aggregative Games," *Games and Economic Behavior*, 31, 303–310.

——— (2000b): "The Strategic Advantage of Negatively Interdependent Preferences," *Journal of Economic Theory*, 92, 274–299.

Kyle, A. S., and A. Wang (1997): "Speculation Duopoly with Agreement to Disagree: Can Overconfidence Survive the Market Test?," *The Journal of Finance*, LII, 2073–2090.

Maynard Smith, J. (1974): "The Theory of Games and the Evolution of Animal Conflicts," *Journal of Theoretical Biology*, 47, 209–221.

——— (1982): *Evolution and the Theory of Games*. Cambridge University Press, Cambridge.

Morris, S., and H. S. Shin (2002): "Global Games: Theory and Application," in *Advances in Economics and Econometrics: Theory and Applications*, ed. by M. Dewatripont, L. P. Hansen, and S. J. Turnovsky. Cambridge University Press, London.

——— (2004): "Heterogeneity and Uniqueness in Interaction Games," in *The Economy as an Evolving Complex System III*, ed. by L. Blume, and S. Durlauf. Oxford University Press, New York.

Oechssler, J., and F. Riedel (2001): "Evolutionary Dynamics on Infinite Strategy Spaces," *Economic Theory*, 17, 141–162.

——— (2002): "On the Dynamic Foundation of Evolutionary Stability in Continuous Models," *Journal of Economic Theory*, 107, 223–252.

Ok, E. A., and F. Vega-Redondo (2001): "On the Evolution of Individualistic Preferences: An Incomplete Information Scenario," *Journal of Economic Theory*, 97, 231–254.

Possajennikov, A. (2000): "On the Evolutionary Stability of Altruistic and Spiteful Preferences," *Journal of Economic Behavior and Organization*, 42, 125–129.

Rotemberg, J. J. (1994): "Human Relations in the Workplace," *Journal of Political Economy*, 102, 684–717.

Samuelson, L. (2001a): "Analogies, Adaptation, and Anomalies," *Journal of Economic Theory*, 97, 320–366.

——— (2001b): "Introduction to the Evolution of Preferences," *Journal of Economic Theory*, 97, 225–230.

Sandholm, W. H. (2001): "Preference Evolution, Two-Speed Dynamics, and Rapid Social Change," *Review of Economic Dynamics*, 4, 637–679.

Schelling, T. C. (1960): *The Strategy of Conflict*. The MIT Press, Harvard University Press.

Sethi, R., and E. Somanathan (2001): "Preference Evolution and Reciprocity," *Journal of Economic Theory*, 97, 273–297.

Shiryaev, A. (1995): *Probability*. Springer, Berlin, 2nd edn.

Taylor, P., and L. Jonker (1978): "Evolutionary Stable Strategies and Game Dynamics," *Mathematical Biosciences*, 40, 145–156.

Thomas, B. (1985): "On Evolutionarily Stable Sets," *Journal of Mathematical Biology*, 22, 105–15.

Vickers, G., and C. Cannings (1987): "On the Definition of an Evolutionary Stable Strategy," *Journal of Theoretical Biology*, 129, 349–353.

Weibull, J. W. (1995): *Evolutionary Game Theory*. The MIT Press, Cambridge, Massachusetts.