EVOLUTIONARY STABILITY AND LEXICOGRAPHIC PREFERENCES*

Larry Samuelson
Department of Economics
University of Wisconsin
Madison, Wisconsin 53706-1320
LarrySam@ssc.wisc.edu

Jeroen M. Swinkels
Olin School of Business
Washington University in St. Louis
St. Louis, Missouri 63130
Swinkels@mail.olin.wustl.edu

February 20, 2002

Abstract We explore the interaction between evolutionary stability and lexicographic preferences. To do so, we define a *limit Nash equilibrium* for a lexicographic game as the limit of Nash equilibria of nearby games with continuous preferences. Nash equilibria of lexicographic games are limit Nash equilibria, but not conversely. Modified evolutionarily stable strategies (Binmore and Samuelson [2]) are limit Nash equilibria. Modified evolutionary stability differs from "lexicographic evolutionarily stability" (defined by extending the common characterization of evolutionary stability to lexicographic preferences) in the order in which limits in the payoff space and the space of invasion barriers are taken.

*Financial support from the National Science Foundation is gratefully acknowledged. We thank David Cooper, Bill Sandholm and Oscar Volij for comments.

Evolutionary Stability and Lexicographic Preferences

by Larry Samuelson and Jeroen M. Swinkels

1 Introduction

Lexicographic games have been used to model situations in which players face two (or more) objectives, one arbitrarily more important than the other. For example, concerns about complexity have been addressed by models of repeated games in which the players' first objective is their monetary repeated-game payoff and their lexicographically second objective is the simplicity of their strategy.

Nash equilibria may not exist in lexicographic games. Further, different plausible extensions of evolutionary stability to lexicographic games have conflicting implications.

In this paper we take the view that lexicographic preferences are a limiting case for situations where preferences are allow continuous trade-offs over various objectives, such as monetary payoffs and complexity, but some objectives are much more important than others. We explore the implications of this view for the appropriate formulation of equilibrium and of evolutionary stability in lexicographic games.

We first define a *limit Nash equilibrium* for a lexicographic game as the limit of Nash equilibria of a converging sequence of games with continuous preferences. Limit Nash equilibria exist for finite normal form lexicographic games. Nash equilibria are limit Nash equilibria but the converse relationship does not hold, failing most obviously (though not only) in cases in which Nash equilibria fail to exist.

Binmore and Samuelson [2] introduce the concept of a modified evolutionarily stable strategy (MESS) for lexicographic games. One could also simply apply the conventional ESS criterion to define a *lexicographic* ESS (henceforth LESS) in such games. We show that a MESS (LESS) is a limit Nash equilibrium (Nash equilibrium), with both converses failing. Neither MESS nor LESS implies the other, with the two concepts corresponding to differences in the order in which two limits are taken. MESS corresponds to a situation in which the complexity cost is small compared to the mutation barrier, while LESS corresponds to a situation in which the mutation barrier is small compared to complexity costs. The appropriate stability concept thus reduces to the relative magnitude of complexity costs and the invasion barrier.

Section 2 presents the model. Section 3 examines the concept of a limit Nash equilibrium. Section 4 explores evolutionary stability.

2 Lexicographic Games

Consider a game G with N players having finite pure strategy sets $S_1, \ldots, S_N$. A typical pure strategy profile is denoted s, while σ denotes a strategy profile that may be either pure or mixed. To capture the idea that players may have primary and secondary (tertiary...) objectives, assume there exist M functions

$$\pi^m(s) : \prod_{i=1}^{N} S_i \to I\!R^N, \qquad m = 1, \ldots, M,$$

where π_i^m is the outcome for player i of objective m given s. Each π^m is extended to mixed strategies in the standard way. In a *lexicographic game*, each player i has lexicographic preferences, denoted by $\Pi^L(\sigma)$, over $(\pi_i^1(\sigma), \ldots, \pi_i^M(\sigma))$. In a repeated game with complexity concerns, for example, $M = 2$ with π^1 identifying monetary payoffs and π^2 reflecting complexity.

A sequence of (non-lexicographic) games G_τ with players $1, \ldots, N$ and strategy sets $S_1, \ldots, S_N$ *approaches* G if the payoff function for each G_τ is given by

$$\Pi(s, \tau) = \sum_{m=1}^{M} \lambda_m(\tau) \pi^m(s),$$

where for $m = 2, \ldots, M$,

$$\begin{aligned} \lim_{\tau \to \infty} \lambda_m(\tau) &= 0 \\ \lim_{\tau \to \infty} \frac{\lambda_{m-1}(\tau)}{\lambda_m(\tau)} &= \infty, \end{aligned}$$

with $\lambda_1(\tau) = 1$ throughout.

3 Limit Nash Equilibria

Fix a lexicographic game G. Then, σ_i is a *best response* to σ if for each σ_i' there is an $m \in \{0, \ldots, M\}$ such that $\pi_i^k(\sigma) = \pi_i^k(\sigma_i', \sigma_{-i})$ for all $k \leq m$, while $\pi_i^{m+1}(\sigma) > \pi_i^{m+1}(\sigma_i', \sigma_{-i})$ if $m < M$. That is, σ_i is a best response if it does better than any other σ_i' at the first level (if any) at which they

$$\begin{array}{cc|c|c|} & & \multicolumn{2}{c}{2} \\ & & X & Y \\ \cline{3-4} \multirow{2}{*}{1} & X & 2,2 & 0,2 \\ \cline{3-4} & Y & 1,2 & 1,1 \\ \cline{3-4} \end{array} \qquad \qquad \begin{array}{cc|c|c|} & & \multicolumn{2}{c}{2} \\ & & X & Y \\ \cline{3-4} \multirow{2}{*}{1} & X & 0,0 & 0,1 \\ \cline{3-4} & Y & 1,0 & 1,1 \\ \cline{3-4} \end{array}$$

$$\pi^1 \qquad \qquad \qquad \qquad \pi^2$$

Figure 1: Lexicographic payoffs for a game with no Nash equilibrium.

disagree. A Nash equilibrium of a lexicographic game is a profile where each player is using a best response (Abreu and Rubinstein [1], Rubinstein [5]).

We first note that a lexicographic game may fail to have a Nash equilibrium. Consider the payoff matrices shown in Figure 1. Consider the first-level payoffs. Any strategy for player 1 which does *not* put probability one on X induces player 2 to choose X, to which X is a strict best response for 1. Hence, the only possible Nash equilibrium is (X, X), in which both players choose X with probability one. But the second-level payoffs ensure that X is not a best response for player 2 to player 1's choice of X, ensuring that (X, X) is not a Nash equilibrium.

This result is no surprise. Once payoffs are lexicographic, the best response correspondence ceases to be upper hemicontinuous, precluding fixed point arguments. For example, X is a (strict) best response for player 2 to any player-1 strategy that attaches positive probability to Y, but is not a best response to a strategy that plays Y with zero probability.

Our motivation throughout will be that lexicographic constructions are interesting to the extent that they provide convenient approximations for cases in which preferences are continuous, but some considerations are very much more important than others. This prompts:

Definition 1 *The strategy profile σ^* of a lexicographic game G is a* limit Nash equilibrium *if σ^* is the limit of Nash equilibria of an approximating sequence $\{G_\tau\}_{\tau=1}^{\infty}$.*

In Figure 1, the unique limit Nash equilibrium calls for player 1 to choose X and player 2 to mix equally over X and Y. To see this, note that for each τ, player 2 prefers Y to X if player 1 chooses X. But player 1's best response to Y is Y, to which player 2's best response is X. Hence, player 1 must mix in equilibrium. To get player 1 to mix in game G_τ, player 2 must play X with probability $(1 + \lambda_2(\tau))/2$, converging to $\frac{1}{2}$ in the limit.

Proposition 1

(1.1) A Nash equilibrium may fail to exist in a lexicographic game.

(1.2) A limit Nash equilibrium exists.

(1.3) Every Nash equilibrium is a limit Nash equilibrium, but the converse fails.

Proof. (1.1) is shown by Figure 1. (1.2) follows from the fact that the space of mixed strategy profiles is sequentially compact. The limit Nash equilibrium of Figure 1 shows that not every limit Nash equilibrium is a Nash equilibrium, giving the first part of (1.3). To see that a Nash equilibrium is a limit Nash equilibrium, we note that for σ^* to be a Nash equilibrium, it must be that, for each player i and strategy σ_i,

$$\begin{array}{rcl}
\pi_i^1(\sigma^*) & = & \pi_i^1(\sigma_i, \sigma^* - i) \\
& \vdots & \\
\pi_i^m(\sigma^*) & = & \pi_i^m(\sigma_i, \sigma^* - i) \\
\pi_i^{m+1}(\sigma^*) & > & \pi_i^{m+1}(\sigma_i, \sigma^*_{-i}) \\
& \vdots &
\end{array}$$

for some $m \in \{0, \dots, M\}$. But then whenever τ and hence $\lambda_{m+1}/\lambda_{m+2}$ is sufficiently large, the weighted payoff inequality $\lambda_{m+1}\pi_i^{m+1}(\sigma^*) > \lambda_{m+1}\pi_i^{m+1}(\sigma_i, \sigma^* - i)$ overwhelms any payoff differences that appear for larger values of m, ensuring that σ^* is a best response in G_τ. ||

Limit Nash equilibria thus capture the equilibria of the game the lexicographic game was constructed to approximate.

It is interesting to note that different approximating sequences may generate different limit Nash equilibria. Consider a game with five players, 1, $1'$, 2, $2'$ and 3, for whom $M = 3$. Figure 2 shows parts of the payoff function for the first four players. Players 1 and 2 have payoffs that only depend upon their actions, with $\pi^3 \equiv 0$ for these players. Player $1'$ and $2'$ have payoffs that only depend upon their actions, with $\pi^1 \equiv 0$ for these players. Player 3's payoff function is such that player 3 prefers X (Y) if player 1 is more (less) likely to choose X than is player $1'$. It is now straightforward to calculate that along an approximating sequence, there is a unique mixed Nash equilibrium in which player 1 plays X with probability

$$p_1(X) = \frac{1 - 2\lambda_2(\tau) - \lambda_3(\tau)}{1 - \lambda_2(\tau) - \lambda_3(\tau)}$$

$$\begin{array}{cc|c|c|} & & \multicolumn{2}{c}{2} \\ & & X & Y \\ \cline{3-4} 1 & X & 1,1 & 1,1 \\ \cline{3-4} & Y & 1,1 & 0,0 \\ \cline{3-4} \end{array} \qquad \begin{array}{cc|c|c|} & & \multicolumn{2}{c}{2} \\ & & X & Y \\ \cline{3-4} 1 & X & 0,0 & 0,1 \\ \cline{3-4} & Y & 1,0 & 1,1 \\ \cline{3-4} \end{array}$$

$$\pi^1_1, \pi^1_2 \qquad\qquad \pi^2_1, \pi^2_2$$

$$\begin{array}{cc|c|c|} & & \multicolumn{2}{c}{2'} \\ & & X & Y \\ \cline{3-4} 1' & X & 1,1 & 1,1 \\ \cline{3-4} & Y & 1,1 & 0,0 \\ \cline{3-4} \end{array} \qquad \begin{array}{cc|c|c|} & & \multicolumn{2}{c}{2'} \\ & & X & Y \\ \cline{3-4} 1' & X & 0,0 & 0,1 \\ \cline{3-4} & Y & 1,0 & 1,1 \\ \cline{3-4} \end{array}$$

$$\pi^2_{1'}, \pi^2_{2'} \qquad\qquad \pi^3_{1'}, \pi^3_{2'}$$

Figure 2: Partial payoffs for game in which limit Nash equilibrium varies with approximating sequence.

while player $1'$ plays X with probability

$$p_{1'}(X) = \frac{\lambda_2(\tau) - \lambda_3(\tau)}{\lambda_2(\tau)}.$$

Let $\lambda_3(\tau) = \lambda_2(\tau)^2$. Then for sufficiently large τ, $p_1(X) < p_{1'}(X)$, and hence in the limit player 3 chooses Y. But if $\lambda_3(\tau) = \lambda_2(\tau)^{\frac{3}{2}}$, then for sufficiently large τ, $p_1(X) > p_{1'}(X)$, and hence in the limit player 3 chooses X.

4 Evolutionary Stability

Abreu and Rubinstein [1] and Rubinstein [5] examined two-player symmetric lexicographic games with two levels in the lexicographic payoff hierarchy, where the second-level payoff depends only upon one's own strategy. A strategy was interpreted as the choice of an automaton to play a repeated game. The first payoff level was the limit-of-the-means monetary payoff in the repeated game, while the second was a measure of the simplicity of the player's automaton. In particular, second-level payoffs depend only on one's own strategy.

In this section, we examine the issues raised by extending the concept of evolutionary stability to such games. Binmore and Samuelson define the concept of a *modified evolutionarily stable strategy* (MESS), which we extend to a *strict* MESS:[1]

Definition 2 *Pure strategy s^* is a* MESS *if, for any mutant strategy σ,*

$$\begin{aligned} \pi^1(s^*, s^*) &\geq \pi^1(\sigma, s^*) \\ \{\pi^1(s^*, s^*) = \pi^1(\sigma, s^*)\} &\implies \{\pi^1(s^*, \sigma) \geq \pi^1(\sigma, \sigma)\} \\ \{\pi^1(s^*, s^*) = \pi^1(\sigma, s^*) \text{ and } \pi^1(s^*, \sigma) = \pi^1(\sigma, \sigma)\} &\implies \{\pi^2(s^*) \geq \pi^2(\sigma)\}. \end{aligned}$$

Strategy s^* is a strict MESS *if one of the inequalities is strict.*

A MESS is an analogue to a neutrally stable strategy (Maynard Smith [4, p. 107]), while a strict MESS is an analogue to an evolutionarily stable strategy. We focus on strict MESS and evolutionary stability, though similar results could be obtained for MESS and neutral stability.

Let Π^L denote the (lexicographic) payoff function of the lexicographic game. Then an alternative approach to incorporating evolutionary stability considerations would involve simply applying the standard definition of an ESS to the function Π^L (Volij [7]). We call the resulting concept a lexicographic ESS, since it extends the commonly-used ESS formulation to games with lexicographic preferences:

Definition 3 *Strategy s^* is a* lexicographic ESS *(*LESS*) of a lexicographic game if, for any mutant σ,*

$$\begin{aligned} \Pi^L(s^*, s^*) &\geq \Pi^L(\sigma, s^*) \\ \{\Pi^L(s^*, s^*) = \Pi^L(\sigma, s^*)\} &\implies \{\Pi^L(s^*, \sigma) > \Pi^L(\sigma, \sigma)\}. \end{aligned}$$

or, equivalently, if, for any mutant σ,

$$\begin{aligned} \pi^1(s^*, s^*) &\geq \pi^1(\sigma, s^*) \\ \{\pi^1(s^*, s^*) = \pi^1(\sigma, s^*)\} &\implies \{\pi^2(s^*) \geq \pi^2(\sigma)\} \\ \{\pi^1(s^*, s^*) = \pi^1(\sigma, s^*) \text{ and } \pi^2(s^*) = \pi^2(\sigma)\} &\implies \{\pi^1(s^*, \sigma) > \pi^1(\sigma, \sigma)\}. \end{aligned}$$

Making the final inequality weak would give a lexicographic neutrally stable strategy of a lexicographic game.[2] We have followed the standard practice

---

[1] Exploiting the symmetry of the game, we drop the player subscripts from our notation, letting S denote a player's strategy set and letting $\pi^m(\sigma, \sigma')$ be the m-level payoff to a player choosing strategy σ and facing an opponent who chooses σ'.

[2] If we relaxed the assumption that π^2 depends only upon one's own strategy, this definition would require a fourth condition of the form $\{\pi^1(s^*, s^*) = \pi^1(\sigma, s^*)$ and $\pi^2(s^*, s^*) = \pi^2(\sigma, s^*)$ and $\pi^1(s^*, s) = \pi^1(\sigma, \sigma)\} \implies \{\pi^2(s^*, \sigma) > \pi^2(\sigma, \sigma)\}$.

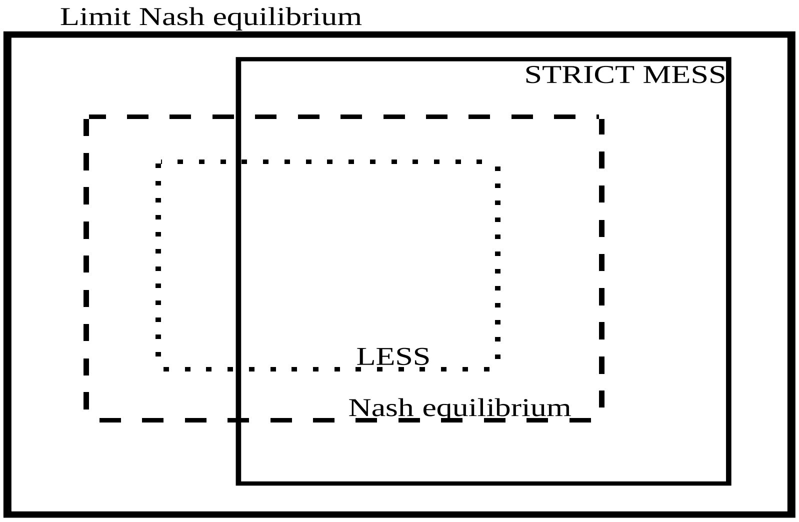

Figure 3: Illustration of Proposition 2. A limit Nash equilibrium always exists, while a Nash equilibrium, LESS and strict MESS need not exist. A Nash equilibrium may fail to exist while a strict MESS exists, and a strict MESS may fail to exist while a LESS exists.

in the complexity-in-repeated-games literature of focusing attention on pure strategies in the automaton choice game as candidates for a MESS or LESS (though retaining the standard equilibrium condition that mixed mutants must also be repelled).

Proposition 2

(2.1) A LESS *is a Nash equilibrium. A strict* MESS *is a limit Nash equilibrium. Each of the converses fails.*

(2.2) A LESS *or strict* MESS *may fail to exist.*

(2.3) A LESS *need not be a strict* MESS*, nor need a strict* MESS *be a* LESS *or Nash equilibrium.*

(2.4) A MESS *may exist while a* LESS *does not, and a* LESS *may exist while a strict* MESS *does not.*

We summarize Proposition 2 in Figure 3.

Proof. (2.1)–(2.2) That a LESS is a Nash equilibrium follows from the first requirement for a LESS given in Definition 3.

Next, letting strategy s^* be a strict MESS, we show that it is a limit Nash equilibrium. Let $\hat{S}$ be the set of all strategies σ (including s^*) for which

$$\pi^1(\sigma, s^*) = \pi^1(s^*, s^*).$$

Fix an approximating sequence of games and, for each τ, let $\sigma(\tau)$ be an equilibrium of the game obtained from G_τ by restricting strategies to be in $\hat{S}$.

We first claim that in the limit, $\sigma(\tau)$ allocates unitary probability to s^*. To show this, suppose that it is not true and note that because $\sigma(\tau)$ is contained in $\hat{S}$,

$$\pi^1(\sigma(\tau), s^*) = \pi^1(s^*, s^*). \tag{1}$$

Because $\sigma(\tau)$ is a Nash equilibrium, $\pi^1(\sigma(\tau), \sigma(\tau)) \geq \pi^1(s^*, \sigma(\tau))$. The opposite inequality is implied by the fact that s^* is a strict MESS, giving

$$\pi^1(\sigma(\tau), \sigma(\tau)) = \pi^1(s^*, \sigma(\tau)).$$

In light of the previous two equalities, s^* can be a strict MESS only if

$$\pi^2(s^*) > \pi^2(\sigma(\tau)). \tag{2}$$

Combining (1) and (2), s^* is a strict best response to $\sigma(\tau)$ among the strategies in $\hat{S}$, a contradiction which allows us to conclude that $\sigma(\tau) = s^*$ for all τ. Next, because s^* is a strict best response to strategies outside $\hat{S}$, s^* must be an unconstrained equilibrium for sufficiently large τ, ensuring that a MESS is a limit Nash equilibrium.[3]

To show that each converse in (2.1) fails, and that a LESS and MESS might fail to exist, consider a game in which π^2 is degenerate. The LESS and strict MESS concepts then coincide with that of an ordinary ESS. Thus, even if we allow a MESS or LESS to be mixed, we need only let π^1 be such that the game has no ESS (e.g., Figure 9.2.1 in van Damme [6, p. 219]). Because the game involves only one level in its lexicographic payoff hierarchy, it has a Nash equilibrium that is also a limit Nash equilibrium, but is neither a LESS nor a MESS.

(2.3)–(2.4) First, consider the game given in Figure 4. Strategy X is a LESS that is not a MESS, while Y is both a strict MESS and a LESS. Next, consider the game shown in Figure 5, due to Volij [7, Figure 1]. The strategy

---

[3] It is straightforward that these arguments carry over to the case of a mixed MESS. We follow the standard practice (cf. Abreu and Rubinstein [1] and Rubinstein [5]) of working with pure MESS or LESS to avoid difficulties in interpreting the complexity of mixed strategies.

		2	
		X	Y
1	X	$0,0$	$0,0$
	Y	$0,0$	$2,2$

π^1

		2	
		X	Y
1	X	$1,1$	$1,0$
	Y	$0,1$	$0,0$

π^2

Figure 4: Strategy X is a LESS that is not a strict MESS.

		2	
		X	Y
1	X	$2,2$	$2,2$
	Y	$2,2$	$0,0$

π^1

		2	
		X	Y
1	X	$0,0$	$0,1$
	Y	$1,0$	$1,1$

π^2

Figure 5: The strategy X is a strict MESS, but (X, X) is not a Nash equilibrium (though (X, X) is a limit Nash equilibrium) and X is not a LESS (which does not exist).

			2	
		X	Y	Z
	X	$2,2$	$0,2$	$6,2$
1	Y	$2,0$	$3,3$	$0,4$
	Z	$0,6$	$4,0$	$5,5$

π^1

			3	
		X	Y	Z
	X	$1,1$	$1,0$	$1,0$
1	Y	$0,1$	$0,0$	$0,0$
	Z	$0,1$	$0,0$	$0,0$

π^2

Figure 6: The strategy X is a LESS, Nash equilibrium, and limit Nash equilibrium, but a strict MESS does not exist.

X is a strict MESS, but (X, X) is not a Nash equilibrium (though (X, X) is a limit Nash equilibrium) and X is not a lexicographic evolutionarily stable strategy (which does not exist). Finally, consider the game shown in Figure 6. The strategy X is a LESS, Nash equilibrium, and limit Nash equilibrium, but a strict MESS does not exist. ||

To illustrate Proposition 2, suppose that the underlying repeated game is the infinitely repeated prisoners' dilemma. Let the first place in the lexicographic preferences be the limit-of-the-means payoff in the repeated game and let the second be a function that is strictly decreasing in the number of states in the automaton. Hence, simpler (i.e., fewer state) machines are preferred. Binmore and Samuelson [2] show that this game has no strict MESS, while a MESS exists and any MESS must maximize the sum of the two players' payoffs (and hence must feature mutual cooperation). A folk theorem for payoffs of lexicographic neutrally stable strategies of the lexicographic game follows from the results of Cooper [3]. Volij [7] shows that always defecting is the unique lexicographic evolutionarily stable strategy.

How do we assess these evolutionary stability concepts? We return to the principle that lexicographic preferences are tools to capture situations in which preferences are continuous but the payoffs attached to some objectives are very small. Similarly, Maynard Smith's definition of an ESS (our Definition 3) was a tool meant to capture a situation in which

$$\Pi(\sigma^*, (1-\epsilon)\sigma^* + \epsilon\sigma') > \Pi(\sigma', (1-\epsilon)\sigma^* + \epsilon\sigma') \tag{3}$$

for all mutants σ' and for all sufficiently small ϵ. Hence, evolutionary stability is meant to ensure that the stable strategy earns a higher average payoff than any mutant who mounts a tiny invasion in a population of the stable strategy. The stipulation "for all sufficiently small ϵ" is typically captured by letting ϵ approach zero.

When working with a continuous payoff function Π, it is straightforward that Definition 3 is equivalent to condition (3) as long as interest is restricted to sufficiently small mutant proportions ϵ. When applying evolutionary stability conditions to games with lexicographic preferences, the appropriate counterpart of condition (3) is less clear. The complications arise out of the fact that the payoff function Π^L is itself an approximation of a case with nonzero but very small complexity costs. We thus have two limits to take, one as the proportion of mutants approaches zero and one as preferences become lexicographic, meaning in our case that complexity costs approach zero. Let us represent the former limit as $\epsilon \to 0$ and the latter limit as $\tau \to \infty$. The following is then a straightforward manipulation of the definitions:

Proposition 3 *Strategy σ^* is a strict* MESS *if there exists ϵ^* such that, for any $\epsilon < \epsilon^*$, there is a $\tau(\epsilon)$ such that, for all $\tau > \tau(\epsilon)$ and all mutants σ',*

$$\Pi(\sigma^*, (1-\epsilon)\sigma^* + \epsilon\sigma', \tau) > \Pi(\sigma', (1-\epsilon)\sigma^* + \epsilon\sigma', \tau).$$

Strategy σ^* is a LESS *if there exists τ^* such that, for any $\tau > \tau^*$, there is an $\epsilon(\tau)$ such that, for all $\epsilon < \epsilon(\tau)$ and all mutants σ',*

$$\Pi(\sigma^*, (1-\epsilon)\sigma^* + \epsilon\sigma', \tau) > \Pi(\sigma', (1-\epsilon)\sigma^* + \epsilon\sigma', \tau).$$

Both a strict MESS and a LESS thus reproduce the ordinary ESS requirement that the candidate for stability achieve a higher expected utility than any mutant in a population characterized by small complexity costs and a small mutant invasion. The two concepts differ in which of these, the mutant invasion or the complexity cost, is relatively small. The strict MESS concept features complexity costs that are arbitrarily small compared to the (small) size of a mutant invasion, while LESS features mutant invasions that are arbitrarily small compared to (small) complexity costs. Given that there are two possible orders for these limits, it is not surprising that we can find some discontinuities, including the potential nonexistence of a Nash equilibrium in the limiting lexicographic game.

For example, consider the game given in Figure 4. We can think of these payoffs as being derived from a choice of automata for an infinitely repeated prisoners' dilemma with limit-of-the-means payoffs. Strategy X represents a single-stage machine that always defects and strategy Y represents the two-state machine TIT-FOR-TAT. Given that mutual defection produces payoffs of $(0,0)$ and mutual cooperation $(2,2)$, π^1 gives the limit-of-the-means payoffs for these strategies. At the second level, π^2 represents complexity costs, with always defecting preferred because it is simpler.

Now consider an approximating sequence $\{G_\tau\}$ where payoffs in game G_t are given by π^1 minus λ times the number of states in the automaton. If we fix the size of a mutant invasion and let $\lambda \to 0$ and hence complexity costs go to zero, then eventually the average payoff of an invading TIT-FOR-TAT exceeds that of DEFECT, disrupting a monomorphic population of defectors. This is consistent with the observation that (Y,Y) is the unique MESS in Figure 4. If instead we fix the game G_τ and let the size of a mutant invasion go to zero, then eventually the invading TIT-FOR-TAT players fare worse then defectors. This is consistent with the fact that (X,X) is a LESS of Figure 4. Intuitively, TIT-FOR-TAT secures a payoff advantage against its fellow mutants while incurring a complexity cost, and hence fares better (worse) than DEFECT if complexity costs (numbers of mutants) are relatively small.

Finally, it is important to note that neither of the implicit limits in Proposition 3, in which either the complexity cost of size of a mutant invasion is first allowed to approach zero, is equivalent to examining simply the limit of evolutionary stable strategies of the approximating sequence. Each game in the approximating sequence for Figure 5 features a unique mixed Nash equilibrium and evolutionarily stable strategy that converges to the MESS in which X is played with probability one, but does not converge to a LESS (which does not exist). Alternatively, strategy X is an ESS in each game in the approximating sequence for Figure 4, converging to a LESS that is not a MESS.

References

[1] D. Abreu and A. Rubinstein. The structure of Nash equilibrium in repeated games with finite automata. *Econometrica*, 56:1259–1282, 1988.

[2] Ken Binmore and Larry Samuelson. Evolutionary stability in repeated games played by finite automata. *Journal of Economic Theory*, 57:278–305, 1992.

[3] David J. Cooper. Supergames played by finite automata with finite costs of complexity in an evolutionary setting. *Journal of Economic Theory*, 68:266–275, 1996.

[4] John Maynard Smith. *Evolution and the Theory of Games*. Cambridge University Press, Cambridge, 1982.

[5] Ariel Rubinstein. Finite automata play the repeated prisoners' dilemma. *Journal of Economic Theory*, 39:83–96, 1986.

[6] Eric van Damme. *Stability and Perfection of Nash Equilibria*. Springer-Verlag, Berlin, 1991.

[7] Oscar Volij. In defense of DEFECT. *Games and Economic Behavior*, 2001. Forthcoming.