Subgame Perfection in Evolutionary Dynamics with Recurrent Perturbations

Herbert Gintis
Ross Cressman
Thijs Ruijgrok*

April 27, 2007

Abstract

We consider finite noncooperative games with a unique subgame perfect Nash equilibrium. Assuming the game strategies are subject to recurrent mutations, we investigate the nature of the equilbria of the resulting evolutionary game, subject to a monotonic dynamic, as the perturbation rate goes to zero. We show by example that limiting equilibria need not be near the subgame perfect equilibrium, or even the connected set of Nash equilibria containing the Nash equilibrium. In particular, game payoffs need not converge to the subgame perfect payoffs. In the n-player Centipede game, however, we show that payoffs converge to the subgame perfect payoff.

1 Introduction

A fundamental property of evolutionary systems governed by the replicator dynamic (Taylor and Jonker 1978) is that every stable equilibrium is a Nash equilibrium of the underlying stage game (Nachbar 1990, Samuelson and Zhang 1992). Since it is well-known that some Nash equilibria are dynamically unstable (Samuelson 1997), stability with respect to the replicator dynamic serves as an equilibrium selection technique. However, this technique is not very effective when the game is given in extensive form. For instance, for every perfect information game (i.e.

*Herbert Gintis: Santa Fe Institute and Central European University, hgintis@comcast.net, http://www-unix.oit.umass.edu/˜gintis; Ross Cressman: Department of Mathematics, Wilfrid Laurier University, email: rcressman@wlu.ca; Thijs Ruijgrok: Mathematics Department, Utrecht University, email: ruijgrok@math.uu.nl.We thank Ken Binmore, Drew Fudenberg, and Sergiu Hart for helpful comments.

an extensive form where each player decides what action to take with complete knowledge of the previous decisions by all players), Cressman and Schlag (1998) (see also Cressman (2003)) show that every pure Nash equilibrium is Lyapunov stable in the replicator dynamic. They also show that, if this pure Nash equilibrium is not subgame perfect, then its Nash equilibrium component is never stable in the replicator dynamic since the subgame perfect equilibrium is contained in any interior asymptotically stable set. This latter result suggests that subgame perfection can be justified on dynamic grounds. Unfortunately, the subgame perfect Nash equilibrium need not be stable either, as elementary examples exist (Cressman and Schlag 1998) where the subgame perfect component is not even locally attracting in the replicator dynamic.

A heuristic argument why the replicator dynamic fails to select subgame perfection in perfect information games is that this is a monotonic selection dynamic (Samuelson and Zhang 1992) and so is characterized by negligible selection pressure towards higher payoff actions at decision points off the subgame perfect equilibrium path.

This intuition led to dynamic models that combine a discrete-time deterministic selection process with a small stochastic effect whereby players sometimes make mistakes and do not choose better strategies. Informally, a stochastically stable state is then a strategy for each player that persists as the mutation rate goes to zero (Young 1993). A particularly relevant model for us is the one developed by Hart (2002) who showed that, with a specific non-monotonic selection dynamic, there is always convergence to the subgame perfect equilibrium as population size increases and the mutation rate goes to zero in such a way that the per-period number of mutations is bounded away from zero. That is, the subgame perfect equilibrium is the unique stochastically stable state for Hart's model.[1] Hart speculated that a similar result would not hold for an underlying monotonic dynamic, though he provided no counterexample. In this paper, we provide analytical counterexamples based on the (continuous-time) replicator dynamic and with mutations modeled as recurrent perturbations. Our analysis shows that these games have a unique dynamically stable Nash equilibrium that does not converge to the subgame perfect equilibrium of the underlying game as the perturbation rates tend to zero. Moreover, our numerical simulations for a wide variety of simple games with perfect information suggest this same result holds quite generally.

We assume throughout that Γ is a finite n-player extensive form game of perfect information with a unique subgame perfect Nash equilibrium, and there are no

[1] Hart (2002) used "evolutionarily stable" in place of "stochastically stable." We avoid this former phrase since it is potentially misleading if it is confused with the static concept of an evolutionarily stable strategy (ESS) that was introduced by Maynard Smith and Price (1973) and forms a cornerstone of evolutionary game theory.

moves by Nature. Evolutionary dynamics are defined through the corresponding n-player normal form of Γ (also called the stage game) where player i has a finite number of pure strategies $s_{i1} \ldots s_{ik_i} \in S_i$. Let α_{kl} be the fraction of population k using pure strategy $s_{kl} \in S_k$. Then the payoff to strategy j of player i, assuming one member of each of the other $n-1$ populations is chosen at random to play the stage game, is given by

$$\pi_{ij} = \sum_{s_{kl_k} \in S_k, k \neq i} \left(\prod_{k \neq i} \alpha_{kl_k} \right) \pi_i(s_{1l_1}, \ldots, s_{i-1l_{i-1}}, s_{ij}, s_{i+1l_{i+1}}, \ldots, , s_{nl_n}).$$

The average payoff to a member of population i is then

$$\overline{\pi}_i = \sum_{j=1,\ldots,k_i} \alpha_{ij}\pi_{ij}.$$

The standard replicator equations governing the dynamics of the population are given by

$$\dot{\alpha}_{ij} = \alpha_{ij}(\pi_{ij} - \overline{\pi}_i), \qquad i = 1, \ldots, n,\ j = 1, \ldots, k_i - 1. \tag{1}$$

The results presented in this paper were anticipated by Binmore, Gale and Samuelson (1995), who added a low level of "drift" to the replicator equations for the Ultimatum Game, and found a locally stable equilibrium near the non-subgame perfect component of the game. Ponti (2000) showed by means of simulation that if agents are forced to adopt strictly mixed strategies, however close to pure, in the three-legged Centipede game there may be no stable equilibria, and the dynamics may exhibit limit cycles. This paper is a generalization of their result to other games and analyzes the limiting behavior of the system as the perturbation rate goes to zero.

2 The Two-move Centipede Game

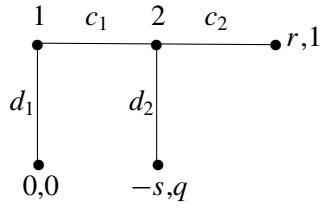

Figure 1: The Two-move Centipede Game.

Figure 1 depicts the extensive form of a two-move Centipede game. We assume $q > 1$ and $r, s > 0$. Player 2 has a weakly dominant strategy d_2 and the Nash equilibria consist of the move d_1 by player 1 and any mixed strategy of player 2 that places at least probability $r/(r+s)$ on d_2. The unique subgame perfect Nash equilibrium is (d_1, d_2).

Let α be the frequency of d_1, let β be the frequency of d_2, let μ_1 be the probability per unit time that player 1 spontaneously changes strategy, and let μ_2 be the corresponding probability for player 2. The replicator equations (1) for this game are then given by:

$$\begin{aligned} \dot{\alpha} &= (1-\mu_1)(1-\alpha)\alpha((s+r)\beta - r) + \mu_1(1-2\alpha) \\ \dot{\beta} &= (1-\mu_2)(q-1)(1-\alpha)\beta(1-\beta) + \mu_2(1-2\beta). \end{aligned} \tag{2}$$

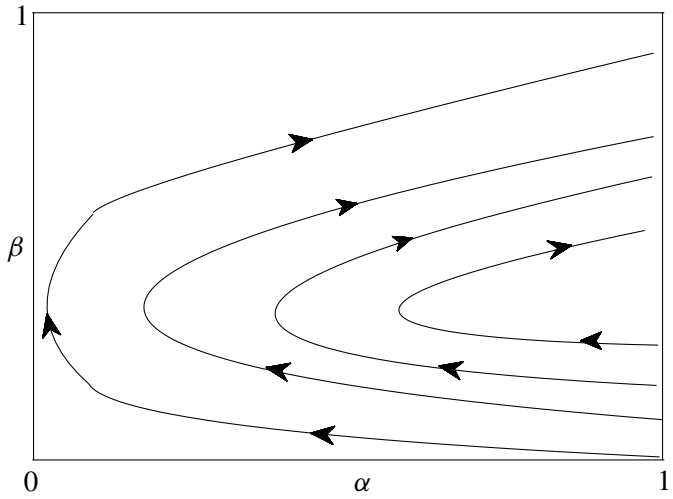

Figure 2: Phase-diagram for equation (2), with $r = 1$, $s = 2$, and $q = 2$ and $\mu_1 = \mu_2 = 0$.

Figure 2 shows the phase diagram for equations (2), with $r = 1$, $s = 2$, $q = 2$ and $\mu_1 = \mu_2 = 0$. With these parameter values, system (2) is not structurally stable. In particular, a generic perturbation will destroy the continuum of fixed points on the line $\alpha = 1$, leaving a finite number of fixed points and possibly limit cycles. For convenience, we will take $r = 1$, $s = 2$, $q = 2$ in the following, although the results can be easily generalized.

Theorem 1. *Consider the dynamical system given by equations (2) for $\mu_1, \mu_2 > 0$.*

a. The attracting set is contained in A (see Figure 3).

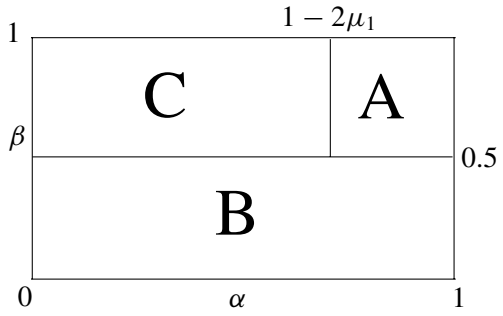

Figure 3: Partitioning of the phase-space.

b. The set A contains an attracting, 1-dimensional invariant manifold.

c. The flow in the invariant manifold contains one attracting fixed point. This fixed point is the limit point for all orbits starting in $[0, 1] \times [0, 1]$.

d. The α*-component of the unique fixed point is equal to one when the mutation terms go to zero. The* β*-component of the unique fixed point can take on any value between* $1/2$ *and* 1 *when the mutation terms go to zero. The final value depends on the relative rates with which* μ_1 *and* μ_2 *go to zero. For instance, if* $\mu_1 = \mu_2$*, then* β *tends approximately to 0.773.*

Proof: Assume $\mu_1, \mu_2 > 0$.

a. It is easy to check that $[0, 1] \times [0, 1]$ is invariant for the flow of (2). When $\beta < 1/2$, we see that $\dot{\beta} = (1-\alpha)\beta(1-\beta)(1-\mu_2)+\mu_2(1-2\beta) > 0$, so there can be no fixed points or limit cycles in B. When $\alpha < 1-2\mu_1$, then $(1-\alpha)\alpha/2 > \mu_1$. So, for $(\alpha, \beta) \in C$, we have that $\dot{\alpha} = (1-\alpha)\alpha(3\beta-1)(1-\mu_1)+\mu_1(1-2\alpha) > (1-\alpha)\alpha(1-\mu_1)/2+\mu_1(1-2\alpha) > \mu_1(1-\mu_1)+\mu_1(1-2\alpha) = 2\mu_1(1-\alpha)-\mu_1^2 > 0$, for sufficiently small μ_1. It follows that there can be no fixed points or limit cycles in C for sufficiently small μ_1. Therefore, the attracting set of the flow of (2) must lie in A for sufficiently small μ_1.

b. For the unperturbed equation, the manifold S_0 consisting of the points with $\alpha = 1, 0 < \beta < 1/2$ is invariant, 1-dimensional, locally attracting and normally hyperbolic (all solutions approach it non-tangentially). According to a theorem by Fenichel (Fenichel 1971), for sufficiently small perturbations, there will be a manifold S_μ in the perturbed system with the characteristics mentioned above

and close to S_0, although no longer consisting of fixed points. The manifold S_μ can be described by $\alpha = 1 + h(\beta, \mu_1, \mu_2)$ with $h(\beta, \mu_1, \mu_2)$ a C^∞ function such that $h(\beta, 0, 0) = 0$.

For $(\alpha, \beta) \in S_\mu$, we must have $\frac{\partial h}{\partial \alpha}\dot{\alpha} + \frac{\partial h}{\partial \beta}\dot{\beta} = 0$. This leads to the equation (we have abbreviated $h(\beta, \mu_1, \mu_2) = h(\beta)$):

$$\begin{aligned} &-(1-\alpha)\alpha(3\beta-1)(1-\mu_1) + \mu_1(1-2\alpha) \\ &\quad + h'(\beta)[(1-\alpha)\beta(1-\beta)(1-\mu_2) + \mu_2(1-2\beta)] = 0. \end{aligned}$$

Substituting $\alpha = 1 + h(\beta)$ yields the equation for $h(\beta)$:

$$\begin{aligned} &-h(\beta)\,(1+h(\beta))(3\beta-1)(1-\mu_1) \\ &\qquad - \mu_1(1+2h(\beta)) + h'(\beta)[h(\beta)\beta(1-\beta)(1-\mu_2) + \mu_2(1-2\beta)] = 0 \end{aligned}$$

Since $h(\beta, \mu_1, \mu_1)$ is small for small μ_1 and μ_2, the first order approximation gives $-h(\beta)(3\beta - 1) - \mu_1 = 0$, so $h(\beta, \mu_1, \mu_2) = -\mu_1/(3\beta - 1)$, so the invariant manifold S_μ is given to first order by $\alpha = 1 - \mu_1/(3\beta - 1)$.

c. To first order, the flow in the invariant manifold S_μ is given by

$$\begin{aligned} \dot{\beta} &= (1-\alpha)\beta(1-\beta) + \mu_2(1-2\beta) \\ &= -h(\beta)\beta(1-\beta) + \mu_2(1-2\beta) \\ &= \mu_1 \frac{\beta(1-\beta)}{3\beta-1} + \mu_2(1-2\beta) \end{aligned} \tag{3}$$

The flow of (2) on $[0, 1] \times [0, 1]$ is given in Figure 4. The flow in A is shown in Figure 5.

d. Equation (3) has only one fixed point when $\beta > 1/2$, and it is easy to check that it is an attractor in S_μ, and therefore an attractor for the full equation (2). Let $\beta^*(\mu_1, \mu_2)$ be the fixed point. Let $\mu_1, \mu_2 \to 0$ in such a way that $\lambda = \mu_1/\mu_2$ remains constant. Then $\beta^*(\mu_1, \mu_2)$ converges to $\beta^*(\lambda)$, where $\beta^*(\lambda)$ is the solution of

$$\lambda \frac{\beta(1-\beta)}{3\beta-1} + (1-2\beta) = 0$$

In Figure 6 $\beta(\lambda)$ is plotted, showing that every limiting value of β is possible, when the mutation terms go to zero. Solving this equation when $\lambda = 1$ yields $\beta = 0.773$.

The behavior of system (2) can be explained as follows. Initially, the logic of backward induction applies, whereby player 2 consistently reduces his probability of

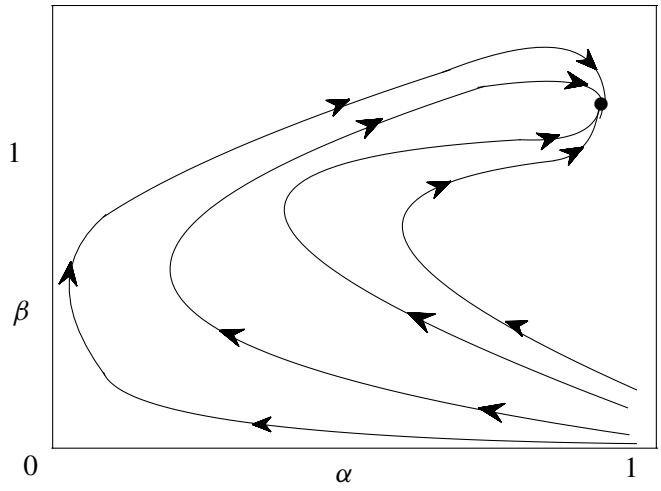

Figure 4: Phase-diagram for equation (2), with $r = 1$, $s = 2$, and $q = 2$ and $\mu_1 = \mu_2 = 0.01$.

playing c_2 and consequently player 1 will more and more use strategy d_1. However, when the system is in a state in set A, a different mechanism takes over. Suppose α and β are both close to 1. Then it pays for player 2 to play c_2 more often, as it entices player 1 to start playing c_1 more. This situation corresponds to a trajectory in Figure 4, where α initially increases, but close to the line $\alpha = 1$ starts to decrease again as β starts to decrease also. However, player 2 needs to be helped by random perturbations, or mistakes, to discover this mechanism. Without mutations, he would simply not discover that if player 1 plays d_1 almost always, it becomes profitable for him to increase the frequency of c_2.

At a certain point, however, the tendency of increasing c_1 and c_2 stops again, namely when it becomes more profitable for player 2 to play d_2 again. This mixed strategy by player 2, teasing and punishing in a certain proportion, is only successful because the mutations in player 1's strategy guarantees a steady trickle of hopeful c_1 plays. We note that the final outcome only benefits player 2, and these benefits are only of order μ.

The only possibility that the subgame perfect solution can be stable is when the mutation rate μ_2 of player 2 is negligible compared to that of player 1. In that case, player 2 simply takes advantage of the small number of c_1 plays by player 1, by eventually playing only d_2. Only when his mutation rate increases, will player 2 discover that by being less greedy (in the form of playing c_2 more often) he can increase the frequency of c_1 by player 1, and on balance profit from the increase in non-trivial interactions.

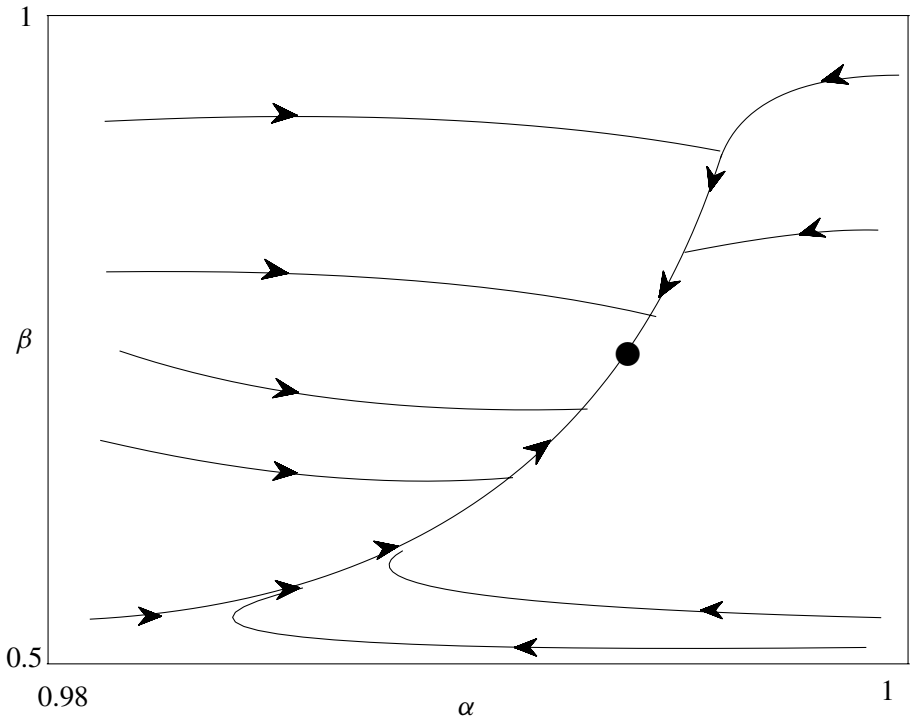

Figure 5: Flow within the set A and the invariant manifold S_μ.

3 Three-move Centipede Game: Agent Extensive Form

Figure 7 depicts the agent extensive form of a three-move Centipede game (i.e., players 1 and 3 have the same payoffs and information, so are in effect one player), assuming r,s,$q > 0$. The unique subgame perfect Nash equilibrium is (d_1, d_2, d_3), although there are many other Nash equilibria.

The replicator equations with recurrent perturbations, with equal perturbation constant μ for all players, are given by:

$$\begin{aligned}
\dot{\alpha} &= \alpha(1-\alpha)(1-q(1-\beta)\gamma)(1-2\mu) + \mu(1-2\alpha) \\
\dot{\beta} &= \beta(1-\alpha)(1-\beta)((r+s)\gamma - s)(1-2\mu) + \mu(1-2\beta) \\
\dot{\gamma} &= q\gamma(1-\alpha)(1-\beta)(1-\gamma)(1-2\mu) + \mu(1-2\gamma),
\end{aligned} \tag{4}$$

where α is the frequency of d_1, β is the frequency of d_2 and γ is the frequency of d_3.

For $\mu = 0$, the equations have a two-dimensional invariant set

$$I = \{(\alpha, \beta, \gamma) | \alpha = 1, 0 \leq \beta, \gamma \leq 1\},$$

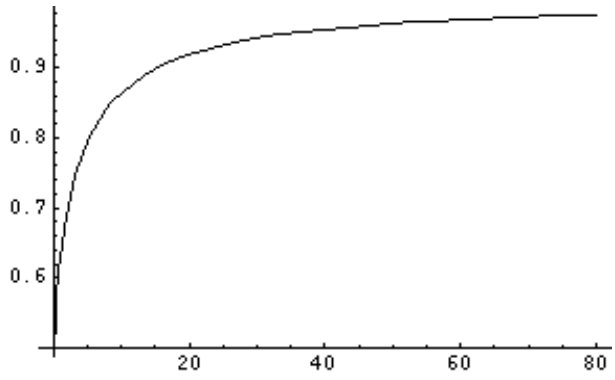

Figure 6: Value of $\beta^*(\lambda)$.

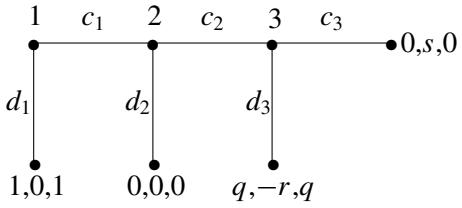

Figure 7: Three-Move Centipede Game: Agent Extensive Form

consisting of fixed points.

The fixed points in $S_0 \subset I$ for which additionally $1-q(1-\beta)\gamma > 0$ are locally attracting. There are no other attracting fixed points.

The manifold S_0 is in fact *globally* attracting. This can be seen by noting that a conserved quantity can be derived from equations (4). Let

$$Q(\beta, \gamma) = \beta^q (1-\gamma)^r \gamma^s, \tag{5}$$

then

$$\frac{d}{dt} Q(\beta, \gamma) = \frac{\partial Q}{\partial \beta}\dot{\beta} + \frac{\partial Q}{\partial \gamma}\dot{\gamma} = 0,$$

as can easily be checked.

The phase space is therefore foliated by invariant two-dimensional manifolds $Q(\beta, \gamma) =$ constant, as shown in Figure 8. These invariant manifolds do not contain any fixed points, so the only possible dynamics within such a manifold is that all solutions flow to a fixed point in S_0, as depicted in Figure 9. This shows that S_0 is globally attracting, or, to put it in another way, that the stable manifold of S_0 is the whole of the phase space.

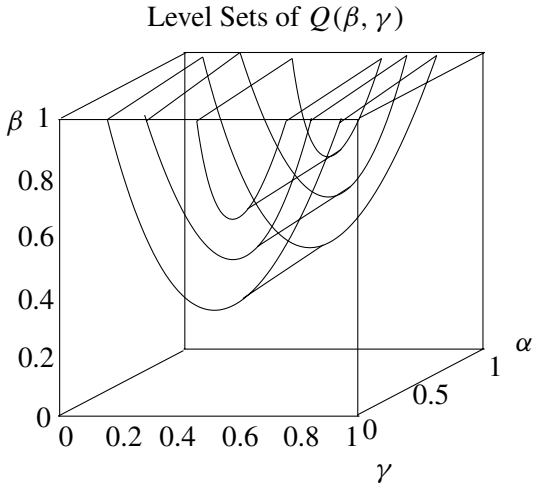

Figure 8: Foliation of Two-Dimensional Manifolds without Fixed Points: Three Dimensional View

For sufficiently small $\mu > 0$, Fenichel's theorem guarantees the existence of a locally attracting two-dimensional invariant manifold S_μ, close to S_0, but not consisting of fixed points. The same theorem also states that locally, the stable manifold of S_μ is close to, and of the same dimension as, the stable manifold of S_0. The small perturbation does not affect the topology of the flow outside a neighborhood of S_0. Therefore, in the perturbed system all solutions will tend to a neighborhood of S_μ. From there, all solutions will be attracted to S_μ itself, since all solution starting in its neighborhood will converge to S_μ.

As in the previous example, S_μ can be defined through $\alpha = 1 + h(\beta, \gamma, \mu)$. Inserting this expression in (4) leads, to first order, to

$$h(\beta, \gamma, \mu) = \frac{-\mu}{1 - q(1-\beta)\gamma} \tag{6}$$

The dynamics within S_μ can be derived by inserting (6) in (4). This yields, to first order,

$$\begin{aligned} \dot{\beta} &= \frac{\beta(1-\beta)((r+s)\gamma - s)}{1 - q(1-\beta)\gamma} + (1 - 2\beta) \\ \dot{\gamma} &= \frac{q\gamma(1-\beta)(1-\gamma)}{1 - q(1-\beta)\gamma} + (1 - 2\gamma) \end{aligned} \tag{7}$$

The phase space of (7) is shown in Figure 10. From this diagram, it follows that there is one attracting fixed point within S_μ. From the above considerations we conclude that this fixed point is the unique global attractor.

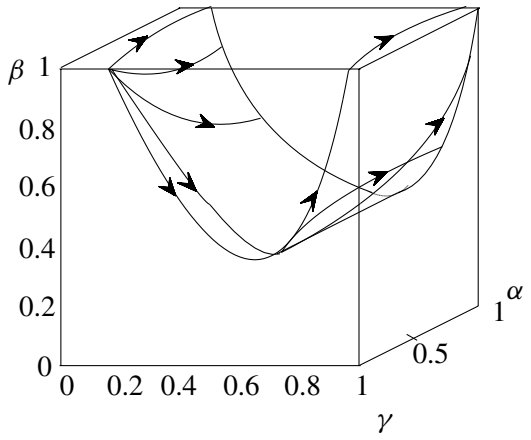

Figure 9: Flow within an invariant manifold $Q(\beta, \gamma) = c$

The replicator equations without perturbations for the fixed point (β^*, γ^*) are

$$(1 - 2\beta^*)(1 - q(1 - \beta^*)\gamma^*) + \beta^*(1 - \beta^*)(r\gamma^* - s(1 - \gamma^*)) = 0 \qquad (8)$$

$$(1 - 2\gamma^*)(1 - q(1 - \beta^*)\gamma^*) + q(1 - \beta^*)\gamma^*(1 - \gamma^*) = 0. \qquad (9)$$

These equations can be solved in closed analytical form, but the result is uninteresting, comprising many hundreds of lines of Mathematica symbols. The more interesting result is that the subgame perfect equilibrium cannot be among the solutions, since both left-hand sides evaluate to -1 when $\beta^* = \gamma^* = 1$. Evaluating the solutions at $r = 1$, $s = 2$, $q = 2$ gave the approximations $\beta^* = 0.566588$ and $\gamma^* = 0.732637$. Clearly, our limit $(1, \beta^*, \gamma^*)$ is far from the subgame perfect equilibrium. The subgame perfect Nash payoffs obtain, however, as $\alpha^* = 1$ since, in fact, $(1, \beta^*, \gamma^*)$ is a mixed strategy Nash equilibrium of Figure 7.

4 Multi-Stage Centipede Game

The examples above can be generalized to the multi-move centipede game in Figure 11. The equations for the multi-move centipede game with n moves, general payoffs and $\mu = 0$, are:

$$\dot{\alpha} = \alpha(1 - \alpha)\Big(a_1 - (a_2 x_2 + a_3(1 - x_2)x_3 + \ldots a_{n+1}(1 - x_2)\ldots(1 - x_n))\Big)$$

$$\dot{x}_j = x_j(1 - \alpha)\ldots(1 - x_j)\Big(c_j - (c_{j+1}x_{j+1} + c_{j+2}(1 - x_{j+2})x_{j+2} + \ldots$$

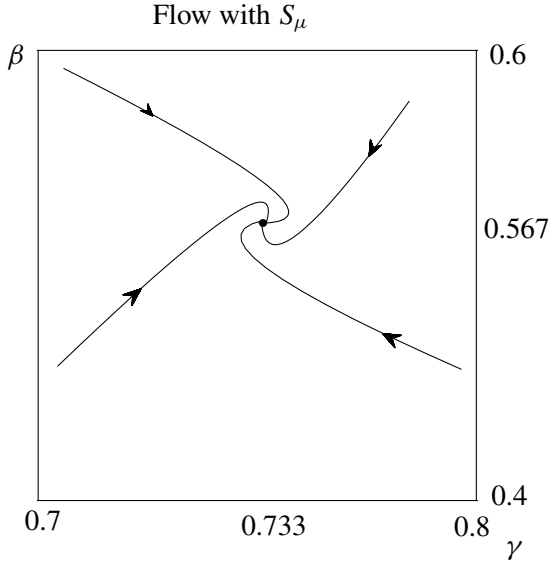

Figure 10: Phase space of (8), (9) in the invariant manifold near $\alpha = 1$

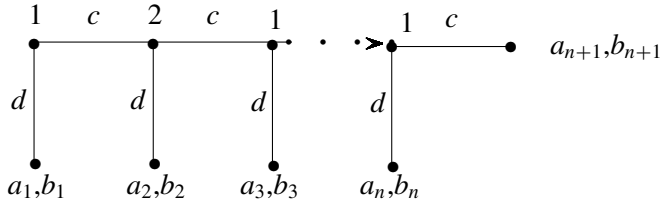

Figure 11: Centipede Game of Length n.

$$\ldots + c_{n+1}(1 - x_{j+1}) \ldots (1 - x_n))\Big) \quad j = 2, \ldots n - 1.$$

$$\dot{x}_n = x_n(1 - \alpha) \ldots (1 - x_n)\Big(a_n - a_{n+1}(1 - x_n)\Big), \tag{10}$$

where $c_j = a_j$ if j is odd and $c_j = b_j$ if j is even. The above equation is for the case that n is odd. When n is even, the letter a in the last equation must be replaced with a b. In these equations, α is the frequency of d_1, x_j is the frequency of d_j, $j = 2, \ldots n$. To ensure that the subgame perfect equilibrium of this game is the solution whereby all players play d, we must impose the condition

$$a_1 > a_2, b_2 > b_3, a_3 > a_4, \ldots a_n > a_{n+1} \tag{11}$$

(see Cressman, 2003, p. 255). Equations (10) have a hierarchy of invariant manifolds, all consisting of fixed points, and all fixed points are contained in one of

these manifolds. V_{n-1} is the $(n-1)$-dimensional manifold of points for which $\alpha = 1$, V_{n-2} is the $(n-2)$-dimensional manifold of points for which $\alpha = 0$ and $x_2 = 1$, and in general V_{n-k} is the $(n-k)$-dimensional manifold of points for which $\alpha = 0, x_2 = 0, \ldots, x_{k-1} = 0, x_k = 1$.

V_{n-1} has a subset of fixed points that are stable, namely those points $(x_1, \ldots, x_n)$ for which

$$a_1 - (a_2 x_2 + a_3(1-x_2)x_3 + \ldots + a_{n+1}(1-x_2)\ldots(1-x_n)) > 0. \qquad (12)$$

All fixed points in the other invariant manifolds are unstable. For V_{n-2}, this can be seen by linearizing around such a fixed point. If we write $\alpha = y_1$ and $(x_2 = 1 + y_2$, with y_1 and y_2 small, then to first order the equation for y_1 is given by: $\dot{y}_1 = (a_1 - a_2)y_1$. It follows from condition (11) that this solution is unstable. Similarly, it follows from (11) that every point in V_{n-k} is unstable in the x_{k-1} direction.

The equations for x_{n-1} and x_n are:

$$\begin{aligned} \dot{x}_{n-1} &= x_{n-1}(1-\alpha)\ldots(1-x_{n-1})\Big(c_{n-1} - (c_n x_n + c_{n+1}(1-x_n))\Big) \\ \dot{x}_n &= x_n(1-\alpha)\ldots(1-x_n)\Big(a_n - a_{n+1}(1-x_n)\Big). \end{aligned} \qquad (13)$$

By 'dividing out' the common term $(1-\alpha)\ldots(1-x_{n-1})$, we can derive a conserved quantity $Q_n(x_{n-1}, x_n)$. Using this expression and the equations for x_{n-2} and x_{n-1}, a second, independent conserved quantity $Q_{n-1}(x_{n-2}, x_{n-1})$ can be derived. This process can be continued to yield $n-2$ independent integrals of motion. The existence of these integrals of motion implies that the phase-space is foliated by 2-dimensional manifolds defined by $Q_n(x_{n-1}, x_n) = c_n$, $Q_{n-1}(x_{n-2}, x_{n-1}) = c_{n-1}, \ldots Q_3(x_2, x_3) = c_3$. Since there are no interior fixed points, the flow on these invariant manifolds is straightforward: all solutions converge to the hyperplane $\alpha = 1$.

Using Fenichel's theorem, we can show, by extending the methods of Sections 2 and 3, that for $\mu > 0$, there will be a globally attracting $(n-1)$-dimensional manifold near the hyperplane $\alpha = 1$. However, it is not a priori clear that for $n > 3$, all solutions in this invariant manifold will converge to a unique fixed point, as in the previous examples.

5 Four-Stage Centipede Game: Agent Extensive Form

In the case of $n = 4$ we do find convergence to a fixed point. As in the previous examples, the subgame perfect equilibrium is never among the solutions to the set of equations determining this fixed point. For instance, for the game depicted

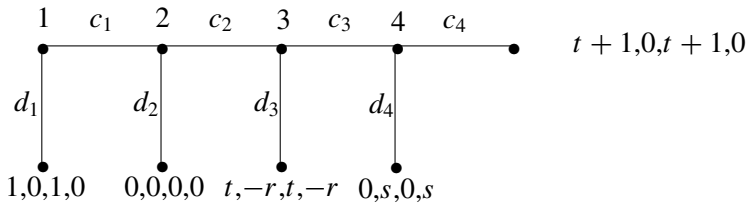

Figure 12: Four-Move Centipede Game

in Figure 12 with r=1, s=4 and q=2, the equations have one fixed point given by $\beta^* = 0.516$, $\gamma^* = 0.786$, and $\delta^* = 0.606$. Evaluation of the Jacobian of the system in a neighborhood of this equilibrium confirms that it is locally stable. Note that if player 3 defects, he gets $q = 2$, whereas if he cooperates he gets $(q+1)(1-\delta^*) = 1.182$, so he should defect. But, if he defected, player 2 should defect, to get zero as opposed to $-r = -1$. If player three cooperates, then so should player 2, so the latter can get the 1.182 also. However, three defects with probability $\gamma^* = 0.786$, so player two actually gets $1.182(1-\gamma^*) - \gamma^* = -0.533$ by cooperating. Why doesn't he just defect? Because the random walk of the perturbation leads him to cooperate with probability 1/2, and the cost of cooperation only reduces this rate by about 10%.

6 The Ultimatum Game

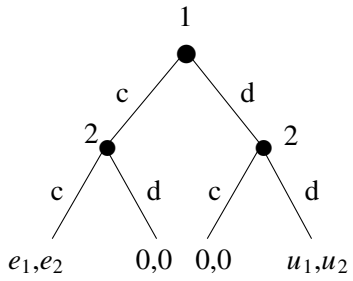

Figure 13: Ultimatum Game

Figure 13 depicts the Ultimatum Game where player 1 (the Proposer) has two strategies, c (cooperate) and d (defect). Player 2 (the Responder) has four strategies $\{cc, cd, dc, dd\}$ where xy means do x if Proposer plays d and do y if Proposer plays c. The payoffs are $\pi_i(d, dd) = \pi_i(d, dc) = u_i$, $\pi_i(d, cd) = \pi_i(c, cd) = 0$, $\pi_i(c, dc) = \pi_i(c, cc) = e_i$, $i = 1, 2$. We assume $u_i, e_i > 0$, $u_1 > \max(u_2, e_1)$, and $u_1 + u_2 = e_1 + e_2 = 10$. Cooperating thus means offering $e_2 > u_2$ for the

Proposer, and accepting e_2 while rejecting u_2 for the Responder. All other moves are defection. The unique subgame perfect Nash equilibrium is (d, dc), but any combination of dd and dc is a best response for the Responder. Another Nash equilibrium component is given by $(c,\ pdc + (1-p)cc)$ for $p \leq e_1/u_1$.

Let α be the probability the Proposer plays d. Numerical simulations show that as $\mu \to 0$, all trajectories converge and that, along a particular trajectory, either $\alpha^* \to 1$ or $\alpha^* \to 0$, and both limits are possible. In particular, unlike the result in Hart (2002), some trajectories do not converge to the subgame perfect equilibrium payoffs. When $\alpha^* \to 1$, we can write the replicator equations as

$$\dot{\alpha} = \mu(1-2\alpha) + \alpha(1-\alpha)(u_1(\beta+\gamma) - e_1(1-\beta-\delta))(1-2\mu) \quad (14)$$

$$\dot{\beta} = \mu(1-4\beta + \beta(u_2\alpha A - e_2 B(1-\beta-\delta))(1-4\mu)) \quad (15)$$

$$\dot{\gamma} = \mu(1-4\gamma + \gamma(u_2\alpha A + e_2 B(\beta+\delta))(1-4\mu)) \quad (16)$$

$$\dot{\delta} = \mu(1-4\delta) - \delta(e_2(1-\alpha)(1-\beta-\delta) + u_2\alpha(\beta+\gamma))(1-4\mu), \quad (17)$$

where β is the probability the responder plays dd, γ is the probability the responder plays dc,and δ is the probability the responder plays cd, and $A = (1-\beta-\gamma)/\mu$, $B = (1-\alpha)/\mu$. Since $\alpha^* = 1$, from (15), letting $\mu \to 0$, we find $\beta^* + \gamma^* = 1$. Now, β^* and γ^* are determined by the following equations:

$$1 - 4\beta + \beta(Au_2 - Be_2(1-\beta)) = 0 \quad (18)$$

$$1 - 4\gamma + \gamma(Au_2 + Be_2\beta) = 0 \quad (19)$$

$$\beta + \gamma = 1 \quad (20)$$

$$\beta + \gamma B(u_1 - \gamma e_1) = 1 \quad (21)$$

These equations have closed form solutions, but it is not illuminating to exhibit them. Assuming $u_1 = 9, u_2 = 1, e_1 = e_2 = 5$, we find $\beta^* = 0.4$ and $\gamma^* = 0.6$. Evaluation of the Jacobian of the system in a neighborhood of this equilibrium confirms that it is locally stable. Clearly, this is not the subgame perfect equilibrium. Indeed, it is clear that for all parameter values, the subgame perfect equilibrium does not obtain.

Suppose $\alpha^* = 0$. Then we can write the replicator equations as

$$\dot{\alpha} = \mu(1-2\alpha + A(1-\alpha)(u_1(\beta+\gamma) - e_1(1-\beta-\delta)))(1-2\mu) \quad (22)$$

$$\dot{\beta} = \mu(1-4\beta - B(u_2\alpha(-1+\beta+\gamma) + e_2(1-\alpha)(1-\beta-\delta))(1-4\mu)) \quad (23)$$

$$\dot{\gamma} = \mu(1-4\gamma + \gamma(u_2 A(-1+\beta+\gamma) - e_2(1-\alpha)((B+D)))(1-4\mu)) \quad (24)$$

$$\dot{\delta} = \mu(1-4\delta - D(u_2\alpha(\beta+\gamma) + e_2(1-\alpha)(1-\beta-\delta))(1-4\mu)), \quad (25)$$

where $A = \alpha/\mu$, $B = \beta/\mu$, and $D = \delta/\mu$. Assuming A, B and D remain bounded as $\mu \to 0$, we have $\beta^* = \delta^* = 0$ and the following four equations hold:

$$A(u_1\gamma^* - e_1) + 1 = 0 \quad (26)$$

$$1 - De_2 = 0 \tag{27}$$

$$1 - Be_2 = 0 \tag{28}$$

$$1 - 4\gamma^* - \gamma^*(u_2 A(1 - \gamma^*) + e_2(B + D)) = 0. \tag{29}$$

These equations reduce to a complicated quadratic equation for A and γ^*. Assuming $u_1 = 9, u_2 = 1, e_1 = e_2 = 5$, we have $\gamma^* = 0.1604$.

7 Conclusion

We have shown that generally, the subgame perfect equilibrium is not a solution of the corresponding replicator equation when small recurrent perturbations are taken into account. From the extensive form examples considered in this paper with non-singleton Nash equilibrium components, it seems clear that evolutionary dynamics with recurrent perturbations do not converge to the subgame perfect equilibrium as the perturbation rate approaches zero. This contrasts with the result in Hart (2002). In fact, for the Ultimatum Game, which has two Nash equilibrium components, we see that some trajectories even converge to the non-subgame perfect component. On the other hand, for Centipede Games of arbitrary length (whose only Nash equilibrium component includes the subgame perfect equilibrium), we have shown that, as the perturbation rate goes to zero, the payoffs approach that of the unique subgame perfect equilibrium, but the frequencies of strategies off the equilibrium path are quite different from the subgame perfect Nash equilibrium.

References

Binmore, Ken, John Gale, and Larry Samuelson, "Learning to be Imperfect: The Ultimatum Game," *Games and Economic Behavior* 8 (1995):56–90.

Cressman, Ross, *Evolutionary Dynamics in Extensive Form Games* (Cambridge, MA: The MIT Press, 2003).

___ and Karl Schlag, "The Dynamic (In)Stability of Backwards Induction," *Journal of Economic Theory* 83 (1998):260–285.

Fenichel, N., "Persistence and Smoothness of Invariant Manifolds for Flows," *Indiana University Mathematics Journal* 21 (1971):193–225.

Hart, Sergiu, "Evolutionary Dynamics and Backward Induction," *Games and Economic Behavior* 41 (2002):227–264.

Maynard Smith, John and G. R. Price, "The Logic of Animal Conflict," *Nature* 246 (2 November 1973):15–18.

Nachbar, John H., "Evolutionary' Selection Dynamics in Games: Convergence and Limit Properties," *International Journal of Game Theory* 19 (1990):59–89.

Ponti, Giovanni, "Cycles of Learning in the Centipede Game," *Games and Economic Behavior* 30 (2000):115–141.

Samuelson, Larry, *Evolutionary Games and Equilibrium Selection* (Cambridge, MA: MIT Press, 1997).

____ and Jianbo Zhang, "Evolutionary Stability in Asymmetric Games," *Journal of Economic Theory* 57,2 (1992):363–391.

Taylor, Peter and Leo Jonker, "Evolutionarily Stable Strategies and Game Dynamics," *Mathematical Biosciences* 40 (1978):145–156.

Young, H. Peyton, "An Evolutionary Model of Bargaining," *Journal of Economic Theory* 59,1 (February 1993):145–168.

c\Papers\Subgame Perfection\Recurrent Mutations.tex April 27, 2007