EVOLUTIONARY STABILITY, COOPERATION, AND HAMILTON'S RULE

Ingela Alger and Jörgen Weibull

Carleton University and Stockholm School of Economics

1 Introduction

- A series of joint studies in an ongoing research project

- All drawing upon a general notion of

  local evolutionary stability

- Today's study (#3) is focused on evolutionarily stable altruism, with

  - implications for cooperation and

  - relations to Hamilton's rule and inclusive-fitness maximization

1.1 Related literatures

- Strategy evolution in populations with assortative matching:
  - Hamilton (1964), Grafen (1979)

  - Bergstrom (1995, 2003), Day and Taylor (1998)

  - Nowak (2006)

- Preference evolution in populations with uniform random matching:
  - Güth and Yaari (1992), Bester and Güth (1998), Ely and Yilankaya (2001)

  - Heifetz, Shannon and Spiegel (2006a,b), Dekel, Ely and Yilankaya (2007)

2 The model

- A large population of individuals
- Individuals have some (abstract) heritable *trait* $\theta \in T$
- They are randomly, but not necessarily uniformly randomly, matched into pairs
- Each pair has a symmetric two-person interaction
- Average *increase of (personal) fitness* $f(\theta, \theta')$ to an individual with trait $\theta \in T$ when interacting with an individual with trait $\theta' \in T$

2.1 The matching process

- Allow for non-uniform random matching, use "algebra of assortative encounters" in Bergstrom (2003)

1. Two traits, θ and θ', present in the population

2. $1-\varepsilon$ the population share with trait θ, ε that with trait θ'

3. Let $\sigma(\varepsilon) = \Pr[\theta|\theta,\varepsilon] - \Pr[\theta|\theta',\varepsilon]$

   This defines an *assortment function* $\sigma : (0,1) \longrightarrow [-1,1]$ which to each population share ε assigns an *index of assortativity,* $\sigma(\varepsilon)$

4. Expected fitness increase for a θ-individual:

$$F\left(\theta, \theta', \varepsilon\right) = \left[\sigma\left(\varepsilon\right) + \left(1 - \varepsilon\right)\left(1 - \sigma\left(\varepsilon\right)\right)\right] \cdot f\left(\theta, \theta\right) + \varepsilon\left(1 - \sigma\left(\varepsilon\right)\right) \cdot f\left(\theta, \theta'\right)$$

and likewise for a θ'-individual

- Assume matching process such that $\sigma\left(\varepsilon\right)$ continuous in ε with limit σ_0 as $\varepsilon \longrightarrow 0$

- *Uniform* random matching in *infinite* population: $\sigma(\varepsilon) \equiv 0$

- *Uniform* random matching in *finite* population:

  $\sigma(\varepsilon) \equiv -1/(N-1) < 0$

- *Sibling interaction* in sexually reproducing diploid species:

  $\sigma(\varepsilon) \longrightarrow r = 1/2$ as $\varepsilon \longrightarrow 0$ [r being Wright's coefficient of relatedness]

2.2 Evolutionary stability

Definition 2.1 *A trait $\theta \in T$ is* **evolutionarily stable** *if for every trait $\theta' \neq \theta$ there exists some $\bar{\varepsilon} \in (0,1)$ such that for all $\varepsilon \in (0, \bar{\varepsilon})$:*

$$F\left(\theta, \theta', \varepsilon\right) > F\left(\theta', \theta, 1-\varepsilon\right)$$

- Special case: Uniform random matching in an asexually reproducing infinite population. The interaction a finite and symmetric two-player game. The set T of traits is the mixed-strategy simplex. Payoff functions being bilinear, the definition $\Leftrightarrow$ ESS

- A *sufficient* condition for evolutionary stability of a trait:

$$f(\theta, \theta) > (1-\sigma_0) \cdot f\left(\theta', \theta\right) + \sigma_0 \cdot f\left(\theta', \theta'\right) \tag{1}$$

  $\forall\ \theta' \neq \theta$.

Definition 2.2 *A trait $\theta \in T$ is* **locally strictly evolutionarily stable (LSES)** *if (1) holds for all $\theta' \neq \theta$ near θ.*

Remark 2.1 *A trait θ is LSES iff the RHS of (1), viewed as a function of the mutant type θ', has a strict local maximum at $\theta' = \theta$.*

- This simple observation turns out to give lots of analytical power:

1. Let $T \subset \mathbb{R}^k$ for some $k > 0$

2. Let $f : T^2 \to R$ be continuously differentiable

3. Take the derivative of the RHS w.r.t. θ':

$$D\left(\theta, \theta'\right) = (1 - \sigma_0) \cdot \nabla f_1\left(\theta', \theta\right) + \sigma_0 \cdot \left[\nabla f_1\left(\theta', \theta'\right) + \nabla f_2\left(\theta', \theta'\right)\right],$$

4. Call $D : T^2 \to \mathbb{R}$ the *drift function*

Proposition 2.1 *Condition (i) below is necessary for θ to be LSES. Conditions (i) and (ii) are together sufficient for θ to be LSES.*

(i) $D(\theta, \theta) = 0$

(ii) $(\theta' - \theta) \cdot D(\theta, \theta') < 0$ *for all* $\theta' \neq \theta$ *near* θ

- Call (i) the *no-drift* condition

2.3 The pairwise interaction

- The pairwise interactions as a symmetric two-player games $G = (X, \pi)$

  - X the set of strategies (available to each player)

- $\pi(x, y)$ payoff to a player who uses strategy $x \in X$ against strategy $y \in X$

- payoff $=$ the increase in one's (personal) fitness

- Special case 1: X the unit simplex of mixed strategies in a finite game

- Special case 2: X an interval of pure strategies on the real line

- Call the game G the *personal-fitness game*

3 Strategy evolution

- Trait = mixed strategy in the game $G = (X, \pi)$

- The no-drift condition:

$$\nabla \pi_1(x, x) + \sigma_0 \cdot \nabla \pi_2(x, x) = 0$$

- Identical with the necessary FOC for symmetric NE in a symmetric game with payoff function

$$u(x, y) = \pi(x, y) + \sigma_0 \cdot \pi(y, x)$$

- As if individuals were altruistic/spiteful of degree σ_0

- Special case: asexual reproduction in infinite population: $\sigma_0 = 0$

- Special case: sexual reproduction in infinite population:

  $\sigma_0 = r$ [Wright's coefficient of relatedness]

  $\Leftrightarrow$ Hamilton's rule

4 Preference evolution

- Traits as "behavioral inclinations", represented by a parameter that guides the individual's choice of strategy, in interactions where individuals know each others' behavioral inclinations

- More specifically: an individual with trait $\theta = \alpha$ chooses a strategy $x \in X$ so as to maximize

$$u^{\alpha}(x, y) = \pi(x, y) + \alpha \cdot \pi(y, x)$$

- $\alpha > 0$ expresses *altruism*, $\alpha = 0$ *selfishness*, and $\alpha < 0$ *spite*

- Now $T = A = (-1, 1)$

- The notion that each individual takes an action $x \in X$ that maximizes her (altruistic, selfish or spiteful) goal function amounts to the requirement that the strategy pair (x, y) be a NE the "derived" game $G^{\alpha,\beta} = \left(X, u^{\alpha}, u^{\beta}\right)$

- Focus on derived games $G^{\alpha,\beta}$ that have a unique, interior and regular Nash equilibrium

- The no-drift condition:

$$v_1(\alpha, \alpha) + \sigma_0 \cdot v_2(\alpha, \alpha) = 0$$

  where $v(\alpha', \alpha)$ is the fitness increase, in equilibrium play, for an α'-altruist when playing an α-altruist

4.1 A general result

Definition 4.1 *The strategies in* $G = (X, \pi)$ *are*

(a) strategically independent *if* $\pi_{12}(x, y) = 0$ *for all* $x, y \in X$

(b) strategic substitutes *if* $\pi_{12}(x, y) < 0$ *for all* $x, y \in X$ *and*

(c) strategic complements *if* $\pi_{12}(x, y) > 0$ *for all* $x, y \in X$

Proposition 4.1 *If the matching process has index of assortativity σ_0 and a degree α of altruism/spite is locally strictly evolutionarily stable, then*

(i) $\alpha < \sigma_0$ if the strategies in $G = (X, \pi)$ are strategic substitutes

(ii) $\alpha > \sigma_0$ if the strategies in $G = (X, \pi)$ are strategic complements

(iii) $\alpha = \sigma_0$ if the strategies in $G = (X, \pi)$ are strategically independent

- Special case: asexual reproduction in infinite population: $\sigma_0 = 0$

- Special case: sexual reproduction in infinite population:

  $\sigma_0 = r$ [Wright's coefficient of relatedness]

  $\Leftrightarrow$ Hamilton's rule in case (iii), but not in cases (i) and(ii)

4.2 Intuition for violation of Hamilton's rule

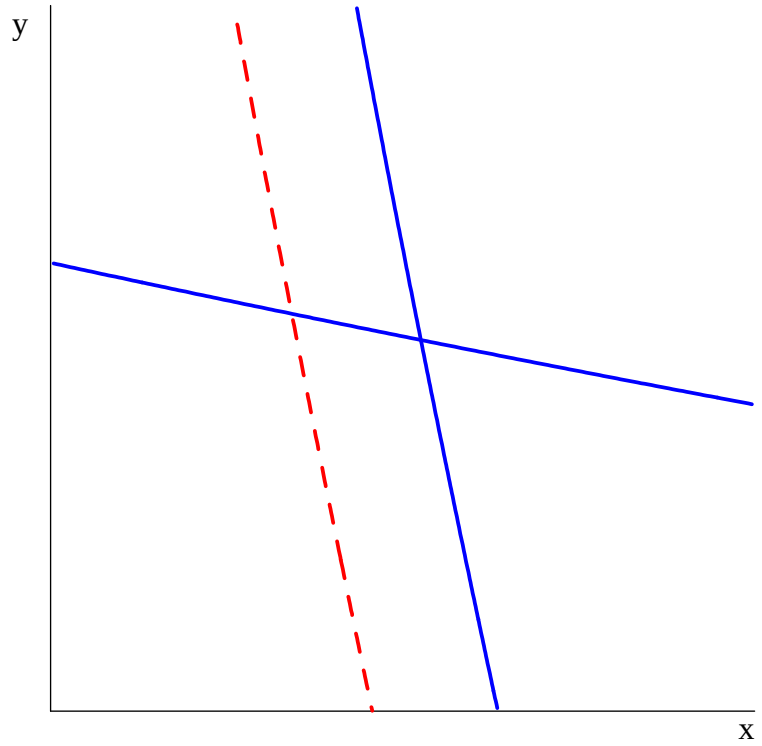
- A social dilemma

$$\pi(x, y) = (x + y)^{\tau} - c \cdot x^2$$

for $0 < \tau < 1$, $c > 0$ and $x, y \geq 0$

- $\Rightarrow$ strategic substitutes

- Two more or less selfish siblings ($\sigma_0 = r = 1/2$):



- However, Hamilton's rule holds in an abstract sense

  A meta-game in which

  - the strategies are α and β [or, equivalently, the associated best-reply correspondences]

  - the payoffs are personal fitnesses in the associated NE $\left(x^{\alpha}, y^{\beta}\right)$ [fixed point to the BR correspondences]

5 Conclusions

- In theoretical biological research on the evolution of altruistic behavior, the standard approach is to assume that evolution operates at the level of strategies, or actions.

- This amounts to assuming that individuals either have no information or knowledge about the inclinations of the individual they interact with, or they have such information or knowledge but fail to use it.

- To this standard approach we here add a model of the evolution of altruistic behavioral inclinations, in pairwise strategic interactions where both parties know each other's inclination.

- We encompass both approaches under a general paradigm for evolution of "traits"

- For this general setting, we define a notion of (local) evolutionary stability and introduce a drift function that allows the researcher to easily identify evolutionarily stable traits in a wide range of pair-wise interactions.

- We derive testable prediction for the degree of cooperation in social dilemmas and other games

- We identify classes of games in which evolutionary stability under preference evolution disagrees with standard application of Hamilton's rule and inclusive-fitness maximization

- The analysis may be extended in many directions:
  - intermediate levels of social cognitive capacity
  - repeated interactions
  - more interacting individuals than two
  - other kinds of other-regarding preferences than altruism/spitefulness