EVOLUTIONARY STABILITY OF GAMES WITH COSTLY SIGNALING

GERHARD JÄGER

Faculty of Linguistics and Literature
University of Bielefeld
Gerhard.Jaeger@uni-bielefeld.de

The paper investigates evolutionary stability conditions of the class of signaling games with the following properties: (a) the interests of sender and receiver coincide, (b) different signals incur differential costs, and (c) different events (meanings/types) have different probabilities.

The main finding is that a profile belongs to some evolutionarily stable set if and only if a maximal number of events can be reliably communicated. Furthermore, it is shown that under the replicator dynamics, a positive measure of the state space is attracted to "sub-optimal" equilibria that do not belong to any asymptotically stable set.

1. Introduction

In his book *Convention*, David Lewis gave a game theoretic formalization of strategic communication (Lewis 1969). Lewis showed that a convention which guarantees successful communication can be self-reinforcing provided the interests of the communicators are sufficiently aligned. In game theoretic parlance, communication conventions are Nash equilibria. As the phenomenon of communication is of high relevance for many scientific disciplines, Lewis style *signaling games* and similar game theoretic models of communication received a great deal of attention since then (see for instance Spence 1973 and Crawford and Sobel 1982 in economics, Grafen 1990 and Hurd 1995 in biology, Skyrms 1996 in philosophy, Hurford 1989 and van Rooij 2004 in linguistics and much subsequent work in all mentioned disciplines). The common theme of all these models can be summarized as follows:

- There are two players, the sender and the receiver.
- The sender has private information about an event that is unknown to the receiver. The event is chosen by nature according to a certain fixed probability distribution.
- The sender emits a signal which is revealed to the receiver.

- The receiver performs an action, and the choice of action may depend on the observed signal.
- The utilities of sender and receiver may depend on the event, the signal and the receiver's action.

Depending on the precise parameters, signaling games may have a multitude of equilibria. Therefore the question arises how a stable communication convention can be established. A promising route is to assume that such equilibria are the result of biological or cultural evolution. Under this perspective, communication conventions should be evolutionarily stable in the sense of evolutionary game theory.

Trapa and Nowak 2000 consider the class of signaling games where signaling is *costless* (i.e. the utility of sender and receiver does not depend on the emitted signal) and the interests of sender and receiver completely coincide. Also, they assume that the actions of the receiver are isomorphic to the set of events. So the task of the receiver is essentially to guess the correct event. Furthermore, they assume a uniform probability distribution over events. Under these conditions it turns out that the evolutionarily stable states (in the sense of Maynard Smith 1982) are exactly those states where the sender strategy is a bijection from events to signals, and the receiver strategy is the inverse of the sender's strategy. This means that in an evolutionarily stable state, the receiver is always able to reliably infer the private information of the sender.[1]

Pawlowitsch 2006 investigates the same class of games, with the additional restriction that the number of events and signals must be identical. She shows that each such game has an infinite number of neutrally stable strategies (again in the sense of Maynard Smith 1982) that are not evolutionarily stable. In these states, communication is not optimal because certain events cannot be reliably communicated. Perhaps surprisingly, these sub-optimal equilibria attract a positive measure of the state space under the replicator dynamics. Natural selection alone thus does not necessarily lead to perfect communication.

In many naturally occuring signaling scenarios emitting a signal may incur a cost to the sender. Games with *costly signaling* have been studied extensively by economists (like Spence 1973) and biologists (as Grafen 1990) because costs may help to establish credibility in situations where the interests of sender and receiver are not completely aligned (an effect that is related to Zahavi's 1975 famous *handicap principle*).

2. Matrix representation of games with costly signaling

A sender strategy is a function from the set $\mathcal{E}$ of events into the set of signals $\mathcal{F}$, and vice versa for the receiver strategy. It is convenient to represent these functions as

[1] Similar results have also been obtained by Wärneryd 1993. Since he only considers pure strategies though, his results are perhaps less general.

matrices containing exactly one 1 per row and only zeros otherwise. A symmetrized strategy is a pair of such matrices (S, R). The probabilities of the events $\mathcal{E}$ are represented as a vector $\vec{e}$ of length $n = |\mathcal{E}|$. We can safely assume that $\forall i : e_i > 0$. The costs of the signals from $\mathcal{F}$ are represented as a vector $\vec{c}$ of length $m = |\mathcal{F}|$. Costs are negative utilities, so it is reasonable to assume $\forall i : c_i \leq 0$. If $c_i - c_j > 1$, the use of the j-th signal would die out under evolution. Therefore we can assume that $\forall i : c_i > -1$. Finally, I will only consider games that are *structurally stable*, i.e. there are no pairs of events that have identical probabilities, and no pairs of signals that incur identical costs. Almost all games of the class considered have this property, so this does not seriously restrict the generality of the setting.

We can construct matrices P and Q:

Definition 1

$$\begin{aligned} p^S_{ij} &\doteq s_{ij} \times e_i \\ q^R_{ij} &\doteq r_{ij} + c_i \end{aligned}$$

(S, R) and (P, Q) stand in a 1-1 correspondence. Therefore we can identify the strategies of sender and receiver with P and Q respectively.

3. Utilities

If the receiver correctly guesses the signal that the sender wants to communicate, both parties score a point. Additionally, the costs of the signal that the sender emits are added to the the utility of both players. This (asymmetric) utility function can be expressed succinctly as

$$u(P, Q) = \mathrm{tr}(PQ) \tag{1}$$

A mixed strategy x corresponds to a pair of stochastic matrices

$$(P^x, Q^x) = (\sum_{(P,Q)} x(P, Q) \times P, \sum_{(P,Q)} x(P, Q) \times Q).$$

The symmetrized expected utility function turns out to be

$$u(x, y) = \frac{1}{2}(\mathrm{tr}(P^x Q^y) + \mathrm{tr}(P^y Q^x)) \tag{2}$$

4. Strong evolutionary stability

The set of evolutionarily stable states (ESS) can be characterized as

Theorem 1 *x is an ESS if and only if $m \leq n$, the first column of P^x has $n - m + 1$ positive entries, each other column of P^x has exactly one positive entry, and $q^x_{ji} = 1 + c_j$ iff $i = \min(\{i' : p^x_{i'j} > 0\})$, otherwise $q^x_{ji} = c_j$.*[2]

In many cases there are also non-trivial evolutionarily stable sets (ESSet) in the sense of Thomas 1985, that can be characterized by the following theorem:

Theorem 2 *A set of strategies A is an ESSet iff for each $x \in A$, x is an ESS or $m > n$, the restriction of P^x to the first n columns and the restriction of Q^x to the first n rows form an ESS, and for each y such that $P^x = P^y$, and Q^x and Q^y agree on the first n rows: $y \in A$.*

5. Weak evolutionary stability

Next to the notion of (strong) evolutionary stability, there is also the concept of weak evolutionary stability (or neutral stability) that characterizes states where the incumbent strategy cannot be replaced by a mutant due to natural selection, but where it is not required that the incumbent is necessarily able to drive any mutant to extinction. The necessary and sufficient condition for neutral stability are

Theorem 3 *x is a neutrally stable state (NSS) if and only if it is a Nash equilibrium and Q^x does not contain multiple column maxima.*

In a NSS, non-determinism thus can only occur in the receiver strategy. It is quite restricted insofar as it can only occur as response to some zero column in P:

Observation 1 *If x is a NSS and there are some i, i', j with $c_j < q^x_{ji}, q^x_{ji'} < 1 + c_j$, then $\forall i' : p^x_{i'j} = 0$.*

This follows directly from the facts that non-determinism can only occur in response to multiple column maxima, which, due to structural stability, can only occur in a zero-column of P if P is pure.

Note that neutral stability without evolutionary stability is quite a pervasive phenomenon.

Observation 2 *If $m, n \geq 2$, there is always at least one NSS that is not element of an ESSet.*

For instance, putting all probability mass into the first column both on the sender side and the receiver side leads to an NSS that is obviously not contained in any ESSet.

[2] Proofs are omitted in this abstract for reasons of space. They can be found in the full paper, which is available online from http://wwwhomes.uni-bielefeld.de/gjaeger/publications/signalspiele.pdf.

6. Dynamic stability and basins of attraction

The games considered in this paper are symmetrized asymmetric (or bimatrix) games. As developed in detail in Cressman 2003, there is a tight connection between static stability and dynamic stability under the replicator dynamics for this class of game. Most notably, a set of strategies is asymptotically stable under the replicator dynamics if and only if it is an ESSet. As a corollary, it follows that the asymptotically stable states are exactly the ESSs.

Let us have a look at the dynamic properties of the set of neutrally stable equilibria. It is rather obvious that all ESSs are isolated points in the sense that each ESS has an environment that does not contain any other Nash equilibria. This follows from the facts that (a) all Nash equilibria are fixed points under the replicator dynamics, and (b) each ESS is asymptotically stable under the replicator dynamics.

The set of NSSs that are not ES has a richer topological structure.

Lemma 1 *Let x^* be a NSS that is not an ESS. There is some $\epsilon > 0$ such that for each Nash equilibrium y with $\|x - y\| < \epsilon$,*

1. *y is itself neutrally stable, and*
2. *for each $\alpha \in [0, 1]$, $\alpha x^* + (1 - \alpha)y$ is neutrally stable.*

We say that two NSSs x and y belong to the same continuum of NSSs if for each $\alpha \in [0, 1]$, $\alpha x^* + (1 - \alpha)y$ is neutrally stable.

Theorem 4 *Each NSS x has some non-null environment A such that each interior point in A converges to some neutrally stable equilibrium y under the replicator dynamics that belongs to the same continuum of NSSs as x.*

We get the immediate corollary:

Corollary 1 *Each NSS belongs to some continuum of NSSs that attracts a positive measure of the state space.*

It is obvious that each ESSet has a basin of attraction with a positive measure—this follows directly from the fact that each ESSet is asymptotically stable. The corollary shows though that the basins of attraction of the ESSets do not exhaust the state space. As pointed out in observation 2, In fact, there are NSSs that do not belong to any ESSet. As ESSets are asymptotically stable, each ESSet has an environment that does not contain any NSSs. Hence if an NSS x does not belong to any ESSet, the entire continuum of NSSs that x belongs to is disjoint from the ESSets. We thus get the additional corollary:

Corollary 2 *The set of Nash equilibria that do not belong to any ESSet attracts a positive measure of the state space.*

These results are of immediate relevance for game theoretic pragmatics. For instance, van Rooij 2004 claims that "signaling games select Horn strategies". This may be true for the stochastic dynamics used there, but under the deterministic replicator dynamics, even the much weaker claim that signaling games select evolutionarily stable sets turns out to be false.

Bibliography

Crawford, V. P. and Sobel, J.: 1982, Strategic Information Transmission, *Econometrica* 50(6), 1431–1451

Cressman, R.: 2003, *Evolutionary Dynamics and Extensive Form Games*, MIT Press, Cambridge (Mass.), London

Grafen, A.: 1990, Biological signals as handicaps., *Journal of Theoretical Biology* 144(4), 517–46

Hurd, P. L.: 1995, Communication in discrete action-response games, *Journal of Theoretical Biology* 174(2), 217–222

Hurford, J. R.: 1989, Biological evolution of the Saussurean sign as a component of the language acquisition device, *Lingua* 77, 187–222

Lewis, D.: 1969, *Convention*, Harvard University Press, Cambridge

Maynard Smith, J.: 1982, *Evolution and the Theory of Games*, Cambridge University Press, Cambridge

Pawlowitsch, C.: 2006, *Why evolution does not always lead to an optimal signaling system*, manuscript, University of Vienna, forthcoming in *Games and Economic Behavior*

Skyrms, B.: 1996, *Evolution of the Social Contract*, Cambridge University Press, Cambridge, UK

Spence, M.: 1973, Job Market Signaling, *The Quarterly Journal of Economics* 87(3), 355–374

Thomas, B.: 1985, On evolutionarily stable sets, *Journal of Mathematical Biology* 22, 105–115

Trapa, P. and Nowak, M.: 2000, Nash equilibria for an evolutionary language game, *Journal of Mathematical Biology* 41, 172–188

van Rooij, R.: 2004, Signalling games select Horn strategies, *Linguistics and Philosophy* 27, 493–527

Wärneryd, K.: 1993, Cheap talk, coordination and evolutionary stability, *Games and Economic Behavior* 5, 532–546

Zahavi, A.: 1975, Mate selection — a selection for a handicap, *Journal of Theoretical Biology* 53, 205–213