Learning to Transfer Information

Simon M. Huttegger and Brian Skyrms

Konrad Lorenz Institute for Evolution and Cognition Research
Adolf Lorenz Gasse 2, A-3422 Altenberg, Austria; e-mail: simon.huttegger@kli.ac.at

Department of Logic and Philosophy of Science, School of Social Sciences
University of California, Irvine, CA 92697-5100; e-mail: bskyrms@uci.edu

Abstract We study a simple game theoretic model of information transfer. Some results on reinforcement learning and fictitious play in these games will be reported and discussed. The main conclusion is that reinforcement learning does in general not lead to efficient information transfer, while fictitious play does.

1 Introduction

Learning from experience becomes particularly interesting in contexts of interactive decision making. It can be studied by theories of learning in games (see [4]). We should expect to find interesting relations between the complexity of a game situation and the least demanding kind of learning behavior which leads to optimal, or close to optimal, play in such a game. Thus, asking whether some, perhaps very simple, learning behavior can lead to optimal decisions for a particular class of games raises profound epistemological issues.

In this paper we study a simple model of information transfer (thus adding another epistemological dimension to the study of learning processes). Our ultimate aim is to relate this model to game theoretic models of communication which have been studied within various disciplines such as biology, economics, AI, and philosophy. The basic features of these models can already be found in signaling games where simple signaling interactions take place between a sender and a receiver. If there is perfect incentive alignment between sender and receiver, then there always exist equilibria which correspond to perfect communication. There are other equilibria as well, however. This raises the question whether boundedly rational individuals will reach an outcome where they can communicate and, thus, transfer information. For standard evolutionary dynamics, as well as for reinforcement learning dynamics, this question has been partially answered affirmatively (see [7, 14, 15]).

This simple baseline model has been extended in a number of different directions. Our proposal aims at considering a slightly different extension where more than two agents are trying to transfer information. To simplify matters, we suppose that agents already have a common "language" (i.e. we assume that they are able to signal successfully). This allows us to use a model for transfer of goods from the economics literature that was introduced by Bala and Goyal (see [2]). This class of games is introduced in section 2. Sections 3 and 4 present some results on the behavior of different kinds of reinforcement learners in the Bala-Goyal game. The baseline result here is that reinforcement learning leads to suboptimal behavior in a non-trivial number of cases. Section 5 gives an outlook on a quite different learning algorithm, fictitious play, whose behavior in the Bala-Goyal game is significantly better.

2 The Bala-Goyal Game

The Bala-Goyal game of information transfer involves at least three agents. Each agent has a piece of information which has a value of 1, say. Agents may visit each other. If A visits B, then A gets B's information, but not the other way round. However, each visit bears a cost c. Suppose that $0 < c < 1$. An agent gets the information from agents she visits herself and from agents she is connected to via a chain of visits. For instance, if A visits B and B visits C, then A gets information from B and C although A does not visit C herself. It is quite obvious that each agent would like to have all the pieces of information from all of the other agents if the value of having several pieces of information is assumed to be the sum of the values of each piece. To get several pieces of information an agent might, for instance, visit an according number of other agents. This is not efficient, however, since every visit bears additional costs. It would be better for each agent if she just had to visit one other agent while getting all the other information indirectly due to this connection.

As it turns out, this goal is achieved for every agent if the connections of the agents form a circle. In game theoretic terms, a circle corresponds to a strict Nash equilibrium. This means that every agent gets strictly less payoff by unilaterally deviating from it. It is also efficient. An agent just has to pay the minimal cost c to get any information at all, while getting all of the available information at the same time. There are other network configurations which constitute non-strict Nash equilibria, however. Thus it is not quite obvious why agents who are choosing independently from each other should arrive, and stay, at an efficient Nash equilibrium.

A real world example for a circle of information transfer is provided by the Kula Ring of the Trobriand islands (see Malinowski [9]). Each island is connected to two neighboring islands, so that the whole network consists of two circles. There are two things of value, necklaces and bracelets. They are interchanged in different directions. Along with interchanging necklaces and bracelets, the connections have the function to facilitate the transfer of goods and information between indirectly connected island societies. Ring structures are common in many societies. They might involve hosting of feasts, exchange of brides, and so on (see, e.g., McKinnon [10]).

For an informed discussion of the issues raised above we clearly have to know more about the agents' learning behavior. The central question we are interested in is whether we can characterize the minimal learning abilities our agents must have in order to arrive at the ring structure. To this end we will look at a very simple learning dynamic first:

reinforcement learning.

3 Reinforcement Learning

The principle behind all reinforcement learning algorithms states that actions which are associated to higher rewards in the past are more likely to be performed in the future. This principle can be characterized formally in many different ways. We will consider two versions of reinforcement learning. One, often called the *basic model*, is due to Roth and Erev [13]. The other one is due to Arthur [1].

Let Γ be a n-person game in normal form and let player j have $m(j)$ pure strategies. Each player j is characterized by a vector of propensities $q^j_n = (q^j_{1n}, \ldots, q^j_{m(j)n})$ at time n. Each q^j_{in} is a positive number that represents the propensity with which player j chooses her ith strategy at time n. Thus, q^j_0 represents the player's initial propensities. Let $Q^j_n = \sum_{k=1}^n q^j_{kn}$. In the basic model of Erev and Roth the probability that player j will choose her ith strategy in round n is given by

$$p^j_{in} = \frac{q^j_{in}}{Q^j_n}.$$

The vector $p^j_n = (p^j_{1n}, \ldots, p^j_{m(j)n})$ gives the probability distribution according to which player j chooses her strategies.

Learning from experience requires the agents to update according to some information they have received. In reinforcement learning, this information consists in the payoff the agents get when they have played a particular action. In the basic model the payoff, which is assumed to be non-negative, is just added to the propensity of the chosen action:

$$q^j_{in+1} = q^j_{in} + \sigma^j_{in},$$

where σ^j_{in} is the random variable which is player j's payoff when she plays her action i at time n and is zero otherwise.

According to the Arthur model each player's step size is renormalised each round. The step size of player j is $\frac{1}{Q^j_n}$. It will in general not be the same for different agents. The renormalization is achieved by multiplying the player's new propensities by a certain factor:

$$q^j_{in+1} = (q^j_{in} + \sigma^j_{in})\frac{C(n+1)}{Cn + P^j_n}$$

where $C > 0$ and P^j_n is j's realized payoff at time n. This has the effect that

$$p^j_{in} = \frac{q^j_{in}}{Cn}.$$

Thus, in the Arthur-model of reinforcement learning step sizes are deterministic sequences proportional to $\frac{1}{n}$. Moreover, they are the same for all players. (This has some technical advantages, as we shall see below.)

There is an evolutionary argument for considering reinforcement learning as an adaptive mechanism. First, observe that the informational and computational requirements for reinforcement learning are quite low. The behavior of a reinforcement learner depends solely on the history of her own payoffs and not on the other players, their payoffs or their pattern of play. Second, it has recently been shown that reinforcement learning leads to optimal behavior in stationary environments (for the, quite involved, proof of the corresponding formal results see Beggs [3] and Hopkins and Posch [6]). One way to express this is given by the following theorem. Suppose the decision maker is making choices at time steps $n = 0, 1, 2, \ldots$. She can choose from a finite set of actions $\{a_1, \ldots, a_m\}$. $u_n(a_i)$ is the payoff she gets from action a_i at time n. Moreover, she updates by the Erev-Roth rule. Let $\mathcal{F}_n$ be the σ-field generatet by the choices up to time n.

Theorem 1 (Beggs 2005) *Let $\gamma > 1$ be a constant. If for some i*

$$\mathbb{E}[u_{n+1}(a_i)|\mathcal{F}_n, a_i \textit{ is chosen at time } n+1] > \gamma\mathbb{E}[u_{n+1}(a_j)|\mathcal{F}_n, a_j \textit{ is chosen at time } n+1]$$

for all n, then the probability that the decision maker chooses j converges to zero almost surely.

Thus, a reinforcement learner almost surely learns to avoid actions which are constantly suboptimal. Thus we might expect evolution to produce organisms that are equipped with some reinforcement learning algorithm. Such an organism, once equipped with a reinforcement learning algorithm, might also employ it when interacting with other organisms. In this case, however, the environment will in general not be stationary. The utility of an action will depend on the choices of the other players. But this choices will, in general, not be constant over time (see Young [16] for more on this).

In the Bala-Goyal game each agent has four possible strategies: visit no one, visit one of the other two agents, visit the other one, or visit both. Suppose each agent is initially equally likely to choose one of these strategies. Reinforcement learning may then be represented by an urn scheme as follows. Each agent has an urn which initially contains four balls of different colors. The agent then chooses one of the balls, plays the corresponding strategy and adds balls of the same color according to the payoff her strategy has yielded. This process is repeated with the new distribution of balls.

This urn scheme is equivalent to Erev-Roth reinforcement learning. We conducted simulations of the Bala-Goyal game for three and four players with the rationale that if it is not possible to get convergence to the circle with a small number of players, then it will not be possible with more players as well. All players were assumed to be reinforcement learners of the Erev-Roth type. For the simulations we varied the cost c. The simulations led to the same qualitative results. The players did in general not converge to a circle. That is to say, they converged to configurations different from the circle a non-trivial proportion of time. These results turned out to be stable even if we let the simulation go on for a very long period of time (1.000.000 rounds). It should be emphasized, however, that the simulation results do not disprove the claim that Erev-Roth type reinforcement learners do converge to the circle with probability one in the long run. It might be the case that the players sometimes spend a very long time in a neighborhood of a suboptimal state, but eventually tend away from it. Nevertheless, if we want to explain the emergence of efficient information transfer in the Bala-Goyal game, reinforcement learning doesn't seem to be a good candidate. Even if there is almost sure convergence to the circle, the time it would take to converge will exceed the lifetime of any simulation, and perhaps also the lifetime of any relevant physical system.

The following result shows that particular kinds of reinforcement learning algorithms may indeed converge to suboptimal states in the Bala-Goyal game. (The proof mimics the argument in the proof of Proposition 5 in Posch [12]. But notice that Posch requires $C > 0$ and that C is less than every payoff in the game. Such a C is not possible for the Bala-Goyal game as we have specified it. Notice also that the crucial step in the proof is that a player gets a fixed payoff if she visits all of the other players herself.)

Theorem 2 *In the Bala-Goyal game for three players, if* $0 < C < 2(1-c)$ *the Arthur model of reinforcement learning converges with positive probability to a state where one agent visits the other two while the other two agents visit the first one.*

Proof. Suppose, without loss of generality, that we are looking at the case where player 1 plays her 4th strategy (which means that she visits the other two agents), and the other two visit 1. We also assume that every strategy has positive initial propensity. If

$$\prod_{n=1}^{\infty} p^1_{4n} > 0$$

then the event that player 1 adds only balls of type 4 to her urn has positive probability in the limit. The above condition holds if

$$\sum_{n=1}^{\infty} (1 - p^1_{4n}) < \infty$$

since $1 > p^j_{in} > 0$ for all n. Suppose 1 chooses action 4 at time n. Then her payoff is $\alpha = 2(1-c)$ regardless of the other players. Hence

$$p^1_{4n+1} - p^1_{4n} = \frac{q^1_{4n} + \alpha}{Q^1_n + \alpha} - \frac{q^1_{4n}}{Q^1_n} = \frac{\alpha(1 - p^1_{4n})}{Q^1_n + \alpha}.$$

By setting $d_n = 1 - p^1_{4n}$ and rearranging we arrive at

$$\frac{d_{n+1}}{d_n} = 1 - \frac{\alpha}{Q^1_n + \alpha} = 1 - \frac{\alpha}{Q^1_n} + O\left((Q^1_n)^{-2}\right).$$

For the Arthur model $Q^1_n = Cn$. Since $C < \alpha$, $\frac{\alpha}{Cn} > \frac{\gamma}{n}$ for some $\gamma > 1$. Thus there exists an N such that for all $n > N$

$$\frac{d_{n+1}}{d_n} < 1 - \frac{\gamma}{n}.$$

By the criterion of Raabe the series $\sum d_n$ converges. ■

This result should not surprise us too much, however. The Arthur model seems in general to be less robust than the Erev-Roth model. Thus, the more interesting question is whether Erev-Roth reinforcement learning does indeed converge to suboptimal states. The best general result concerning the convergence of Erev-Roth reinforcement learning to date can be found in Hopkins and Posch (2005). Their main result relates the behavior of Erev-Roth reinforcement learners to the behavior of multi-population versions of the evolutionary replicator dynamics (see [5]). The standard n-population replicator dynamics is given by

$$\dot{\mathbf{x}}_i^j = \mathbf{x}_i^j \left(u(s_i, \mathbf{x}^{-j}) - u(\mathbf{x}^j, \mathbf{x}^{-j}) \right) \quad \text{for } i = 1, \ldots, n.$$

Here s_i is one of player j's pure strategies. (Each player position of the underlying game is associated with a population.) $\mathbf{x}_i^j$ is the frequency of population j playing strategy s_i. $\dot{\mathbf{x}}_i^j$ is the rate of change of this frequency. The rate of change is determined by the current frequency, the payoff s_j gets when the state in the other populations is given by $\mathbf{x}^{-j}$, and the average payoff in population j given $\mathbf{x}^{-j}$.

The adjusted version of the n-population replicator dynamics is given by

$$\dot{\mathbf{x}}_i^j = \frac{\mathbf{x}_i^j \left(u(s_i, \mathbf{x}^{-j}) - u(\mathbf{x}^j, \mathbf{x}^{-j}) \right)}{u(\mathbf{x}^j, \mathbf{x}^{-j})} \quad \text{for } i = 1, \ldots, n.$$

Thus, the adjusted version involves normalization by average payoff in addition to the function used by the standard version of the replicator dynamics. Both versions have the same rest points. Their stability properties might be different, however.

Hopkins and Posch show that the expected motion of the Erev-Roth model as well as the Arthur model is a version of the standard replicator dynamics. Convergence to particular states is related to the stability properties of the corresponding adjusted version of the replicator dynamics. To be more specific, Hopkins and Posch prove that with probability 1 Erev-Roth reinforcement learning does not converge to a state $\hat{\mathbf{x}}$ if (i) $\hat{\mathbf{x}}$ is not a Nash equilibrium of the underlying game or (ii) $\hat{\mathbf{x}}$ is a Nash equilibrium that is linearly unstable under the adjusted replicator dynamics. (A state is linearly unstable if the Jacobian of the differential equation evaluated at this point has at least one eigenvalue with positive real part.)

A computation of the eigenvalues of the linearization of the adjusted replicator dynamics at a suboptimal state like the one from Theorem 2 yields four negative eigenvalues and two zero eigenvalues. This is in accordance with the fact that pure states cannot be linearly unstable. Thus, standard tools do not tell us whether the Erev-Roth model converges to suboptimal states with positive probability or not.

4 Discounted Reinforcement Learning

One unrealistic feature of Erev-Roth type reinforcement learning is that past memories have the same weight as recent experiences. (Mathematically, this is expressed by the fact that every action of each player is chosen infinitely often in the Erev-Roth model.) This assumption is psychologically implausible. Recent experiences do have more influence on decisions than experiences that were made in the remote past. Moreover, as we have already noted, games are usually non-stationary environments, since each agent's decision may change the environment from one round of play to the next. Thus, it may be reasonable not to assume stationarity from the start by giving all experiences equal weight.

One way to address these worries is to introduce a discounting parameter $\phi \in (0,1)$ into the basic model. After each round the agents multiply the weights of the previous rounds by ϕ and add the undiscounted payoffs of the current round to arrive at their new propensities. Apart from being more plausible, discounting also has the advantage that it may speed up the process.

The way a player computes probabilities from propensities is left untouched by discounting. The rule by which updating of propensities takes place is modified to

$$q^j_{in+1} = (1-\phi)q^j_{in} + \sigma^j_{in}$$

for each player j and each of her actions i. It is well known that discounting can cause trapping in many settings (see [11]). A process is trapping if, with positive probability, all players always choose from a proper subset of actions. The next theorem shows that this also happens in the Bala-Goyal game.

Theorem 3 *In the Bala-Goyal game for three players, the discounted basic model of reinforcement learning converges with positive probability to a state where one agent visits the other two while the other two agents visit the first one.*

Proof. Suppose there are three players, 1, 2, 3. Denote the matrix consisting of their probabilities for choosing strategies by $\mathbf{P}$. For definiteness, assume that $\mathbf{P}$ is a 3×4-matrix, where the first column gives the probability of visiting nobody, the second column gives the probability of visiting agent i with $i < j$ (where i, j are the agents which might be visited), the third gives the same for the other agent, and the fourth is for visiting both. For each such $\mathbf{P}$ we associate a directed graph $G(\mathbf{P})$ as follows. A directed edge (i,j) is in in $G(\mathbf{P})$ if $p_{ij} > \epsilon$ (where p_{ij} is the probability of i visiting j) or if $p_{i4} > \epsilon$ (or if both), where $j \neq i$ and $\epsilon < \frac{1}{8}$ is some fixed positive number. For definiteness, we assume that p_{i1} is the edge in G where an agent visits herself. We will show that, with probability at least $\delta > 0$, $G(\mathbf{P})$ will consist solely of the edges $(1,2), (1,3), (2,1), (3,1)$ from some time onward. Denote the graph having those edges by G^*.

Our first step consists in showing that with probability at least $\delta_1 > 0$ we can get from $G(\mathbf{P})$ to G^*. Notice that there exists a number N such that if 1 chooses her forth strategy and 2 and 3 their second strategies for N consecutive rounds, then $G(\mathbf{P}(N)) = G^*$. For each round, this probability is at least ϵ^3. Take $0 < \delta_1 < \epsilon^{3N}$. Then our first step is completed.

The next, and last, step consists in showing that with probability at least δ_2, $G(\mathbf{P}(t)) = G^*$ for $t \geq N$. By definition of G, the sum of probabilities of each agent not choosing the strategies in G^* is at most 3ϵ. Hence, each sum of those probabilities is at most $\frac{1}{2}$.

After k rounds of choosing according to G^*, this sum is at most $\frac{1}{2}(1-\phi)^k$. Thus the probability of visiting only according to G^* for N rounds is at least

$$\prod_{k=0}^{N-1} \left(1 - \frac{1}{2}(1-\phi)^k\right)^3 .$$

Since $(1-\phi)^k$ is summable, letting $N \to \infty$ yields an infinite product > 0. Taking δ_2 less than the infinite product and letting $\delta = \delta_1 \delta_2$ proves the claim. ■

5 Fictitious Play: An Outlook

An agent who chooses strategies according to fictitious play knows the strategy sets of the other agents and bases her choices on the historical frequencies with which the other agents chose their actions. Formally, let $\mathbf{e}_n^j$ be the random unit vector in $\mathbb{R}^m$ which is 1 at the ith entry if player j has, at time n, chosen her ith strategy (where $i = 1, \ldots, m$) and 0 otherwise. Then

$$\mathbf{x}_n^j := \frac{1}{n} \sum_{k=1}^{n} \mathbf{e}_k^j$$

is player j's vector of historical frequencies of play up to time n. If there are k players, then j is a fictitious play learner if she best responds at time $n+1$ to

$$\mathbf{x}_n^1 \times \ldots \times \mathbf{x}_n^{j-1} \times \mathbf{x}_n^{j+1} \times \ldots \times \mathbf{x}_n^k =: \mathbf{x}_n^{-j}.$$

That is to say, at time $n+1$ player j chooses an action that maximizes her expected utility given $\mathbf{x}_n^{-j}$. It is assumed that j believes the choices of the other players to be independent. There might be more than one best reply to some $\mathbf{x}_n^{-j}$. Then we suppose that there is a suitable tie breaking rule.

It is easy to see that fictitious play is more demanding than reinforcement learning relative to computational and informational requirements. Reinforcement learning is just based on a summary statistics of one's own past payoffs. Fictitious play involves a computation of expected payoffs. Moreover, this computation assumes that the agent has a model of the other agents. This model includes information about the possible strategies of the other agents and their decisions. Thus, reinforcement learners may solely be viewed as playing against Nature while fictitious play learners differentiate at least some parts of their environment.

Suppose that there are three agents playing the Bala-Goyal game. Each one is assumed to be a fictitious play learner. Then we may associate two urns with an agent. The two urns represent the strategies of the other two players. Thus each urn contains balls of four different colors. The number of balls of color i in player j's kth urn at time n is $q_{i,k,n}^j$, where $k = 1, 2$. It represents the number of times this agent has played the corresponding strategy. Thus

$$q_{i,k,n+1}^j = q_{i,k,n}^j + \mathbf{1}_i,$$

where $\mathbf{1}_i$ is the indicator function of the event that the strategy corresponding to i has been played. The agent's probability for another agent playing a certain strategy is given by the number of balls corresponding to that strategy normalized by the total amount of balls in the urn.

$$p^j_{i,k,n} = \frac{q^j_{i,k,n}}{Q^j_{k,n}}$$

for $Q^j_{k,n} = \sum_i q^j_{i,k,n}$. The agents are assumed to choose independently. Hence an agent's probability that the other two agents will jointly play strategy combination (i, r) is just the product of the probabilities for each strategy, $p^j_{i,1,n}p^j_{r,2,n}$. Given this setup, each agent chooses an action that maximizes her expected payoffs relative to the probabilities induced by her urns. If there is more than one such strategy, we assume that the agent randomizes between them. Taken together, this defines a fictitious play process for the Bala-Goyal game.

We conducted a series of simulations using this urn scheme. If players started with uniform beliefs over the other players actions, then they always converged to a circle in the fictitious play process described above. (Convergence here means that they played the strategies corresponding to a circle at the end of the simulations). More generally, the agents seem to converge to a circle in "normal" cases. By normal we mean that the initial distribution of balls is not too strongly biased toward a suboptimal configuration of the game. If we employ some stochastic version of the fictitious play process, then players always seem to converge to a circle. Such a stochastic version allows the agents to sometimes choose actions with only slightly suboptimal expected utility. This can be interpreted as making mistakes or as experimentation. (For more information about this kind of fictitious play see [8].)

It will be important to study the convergence properties for different kinds of fictitious play, and perhaps mixtures of reinforcement learning and fictitious play as well as other learning schemes, more thoroughly in order to understand what makes efficient information transfer in the, seemingly simple, Bala-Goyal game possible. At first sight, it seems that the lack of having a model of the other players may account for the inability of reinforcement learning to achieve this.

The Bala and Goyal model of information transfer can be extended in many interesting directions. Agents might have to learn to transfer information and to communicate simultaneously. The value of information might be decreasing in the number of pieces of information an agent gets. Models of information transfer need not be asymmetric, moreover. And lastly, there may be other learning algorithms in the spectrum between reinforcement learning and fictitious play. Consideration of these will lead to a deeper understanding of the mechanisms governing information transfer in simple adaptive systems.

References

[1] W. B. Arthur. On Designing Economic Agents that Behave like Human Agents. *Journal of Evolutionary Economics*, 57:94–107, 1993.

[2] V. Bala and S. Goyal. A Noncooperative Model of Network Formation. *Econometrica*, 68:1181–1129, 2000.

[3] A. W. Beggs. On the Convergence of Reinforcement Learning. *Journal of Economic Theory*, 122:1–36, 2005.

[4] D. Fudenberg and D. K. Levine. *The Theory of Learning in Games.* MIT Press, Cambridge, Mass., 1998.

[5] J. Hofbauer and K. Sigmund. *Evolutionary Games and Population Dynamics.* Cambridge University Press, Cambridge, 1998.

[6] E. Hopkins and M. Posch. Attainability of Boundary Points under Reinforcement Learning. *Games and Economic Behavior*, 53:110–125, 2005.

[7] S. M. Huttegger. Evolution and the Explanation of Meaning. Forthcoming in *Philosophy of Science*, 2005.

[8] D. S. Leslie and E. J. Collins. Generalized Weakened Fictitious Play. Forthcoming in *Games and Economic Behavior*, 2006.

[9] B. Malinowski. *Argonauts of the Western Pacific: An Account of Native Enterprise and Adventures in the Archipelagoes of Melanesian New Guinea.* Routledge and Kegan Paul, London, 1922.

[10] S. McKinnon. *From a Shattered Sun. Hierarchy, Gender, and Alliance in the Tanimbar Islands.* The University of Wisconsin Press, Madison, 1991.

[11] R. Pemantle and B. Skyrms. Network Formation by Reinforcement Learning: The Long and Medium Run. *Mathematical Social Sciences*, 48:315–324, 2004.

[12] M. Posch. Cycling in a Stochastic Learning Algorithm for Normal Form Games. *Journal of Evolutionary Economics*, 7:193–207, 1997.

[13] A. Roth and I. Erev. Learning in Extensive Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term. *Games and Economic Behavior*, 8:164–212, 1995.

[14] B. Skyrms and R. Pemantle. Learning to Signal. Working Paper, University of California, Irvine, 2005.

[15] R. van Rooy. Evolution of Conventional Meaning and Conversational Principles. *Synthese*, 139:331–366, 2004.

[16] H. P. Young. *Strategic Learning and its Limits.* Oxford University Press, Oxford, 2004.