# Co-Learning and the Evolution of Social Activity

Yoav Shoham and Moshe Tennenholtz
Robotics Laboratory
Department of Computer Science
Stanford University
Stanford, CA 94305

**Abstract**

We introduce the notion of *co-learning*, which refers to a process in which several agents simultaneously try to adapt to one another's behavior so as to produce desirable global system properties. Of particular interest are two specific co-learning settings, which relate to the emergence of conventions and the evolution of cooperation in societies, respectively. We define a basic co-learning rule, called Highest Cumulative Reward (HCR), and show that it gives rise to quite non-trivial system dynamics. In general, we are interested in the eventual convergence of the co-learning system to desirable states, as well as in the efficiency with which this convergence is attained. Our results on eventual convergence are analytic; the results on efficiency properties include analytic lower bounds as well as empirical upper bounds derived from rigorous computer simulations.

# 1 Introduction

In multi-agent AI systems, such as multi-planner systems, it is crucial that the agents agree on certain rules, in order to decrease conflicts among them and promote cooperative behavior. Without such rules even the simplest goals might become unattainable by any individual agent, or at least not efficiently attainable (just imagine driving in the absence of traffic rules). These rules strike a balance between on the one hand allowing agents sufficient freedom to achieve their goals, and on the other hand restricting them so that they do not interfere too much with one another. Some of these rules are social laws, designed and imposed ahead of time; traffic laws are an example. Previous work [8, 11] investigated some aspects of this off-line design of social law. However, not all rules can be legislated in advance. This is either because the characteristics of the society are unknown, or because they change over time. In such cases, it is often important that the society converge on a convention in a dynamic fashion. In human societies this is common; this is how (e.g., software) standards emerge long before they are enshrined in official regulations.

How do such conventions emerge? The usual process, which will motivate our formal framework, is one in which individual agents occasionally interact with one another, and as a result gain some new information. Based on its personal accumulated information, each agent updates its behavior over time. Since all agents are simultaneously engaged in this activity, we call the process *co-learning*.

The complexity of co-learning derives from its concurrent nature: as one agent adapts to the behavior of the agents it has encountered, these agents update their behavior in a similar fashion. This tends to result in highly non-linear system dynamics. The typical question we will be interested in is how different update rules and other system characteristics affect the eventual emergence of desirable global system characteristics (for example, a convention), and how efficiently these desirable states are achieved. As it turns out, our results on eventual convergence will be primarily analytic, whereas the results on efficiency include both analytic lower bounds and empirical upper bounds derived from extensive computer simulations.

Although our initial motivation for this research lay in our interest in multi-agent systems, co-learning is obviously a natural extension of work in Machine Learning (ML). Research in ML is typically concerned with an agent that tries to adapt to an environment. In different areas of research in ML the environment has different structures: it might be a stochastic process that generates examples [3], a teacher[14], a source of reinforcement feedback [13], and so on. In all work in ML of which we are aware, however, the environment is passive: the agent tries to adapt to the environment, but in no sense does the environment try to adapt to the agent.[1] This is an appropriate model in many applications, but not in all.

This having been said, it is the case that the formal setting of co-learning to be presented is similar to that of reinforcement learning. At any point in time, each agent selects one of several possible *actions*, and as a result receives a particular *payoff* or *feedback*. On the basis of the feedback, the agent might *update* its behavior and decide on a new action. If the agent chooses a satisfactory update rule, then the group of agents will eventually exhibit some satisfactory emergent behavior. Again, the critical property of co-learning is that the update performed by one agent might well affect the future feedback of other agents.

In order to study co-learning, we must be more precise about the criteria of successful learning, as well as the nature of the feedback agents get. Starting from the former, standard ML enjoys natural measures of the success of the learning process. These include an increase in the ability of an agent (student) to identify instances of a concept, an improvement in its ability to predict the next symbol generated by a stochastic source, and so on. Co-learning, on the other hand, does not come equipped with such obvious criteria. In order to measure the success of co-learning we will have to introduce external measures; these will measure the success of the society, as opposed to success of individuals. Much of the paper will concentrate on two particular criteria of success, called *convention* and *cooperation*, which will be defined later.

The feedback that an agent gets, given the action it has chosen and the state of the environment (that is, the action of other agent(s)), can be represented in various forms. Works within ML diverges on this form: it may be a

[1] To the reviewer: We believe this statement is essentially correct, but welcome pointers to exceptions; those will not invalidate the contribution of the paper, hopefully.

positive/negative answer to a query, a numerical value, etc. In the framework of co-learning there is a natural candidate for the feedback representation. The actions that the agents perform at a particular point can be treated as strategies in a *game* in the sense of economics (see [9], and definitions in next section), with the agent's payoff at each point being interpreted as its feedback. This game-theoretic representation is quite general, and can be specialized by defining particular types of payoff functions; each restriction on the payoff function defines a particular game type. Despite these game-theoretic matrices, however, the similarity between our framework and game theory will be limited. In particular, we emphasize that the payoff function will determine the feedback of the agents, but not their update rule nor the criterion of success; those will have to be defined independently.

A major part of this paper is devoted to the study of a simple but fundamental update rule, and its effects on co-learning. Roughly speaking, this update rule states that the agent should adopt the action that has yielded the highest cumulative reward to date. We refer to this update rule as the Highest Cumulative Reward (HCR) update rule. This is perhaps the simplest update rule that comes to mind; certainly it is simpler than update rules in the reinforcement-learning literature, such as Q-learning [15]. However, the simultaneous adaptation of the various agents will lead to highly nontrivial behaviors even with this relatively simple rule. We will be able to show that this simple rule leads to eventual success in a large class of games, including two basic games associated with the notions of convention and cooperation respectively. We will also demonstrate that such convergence properties do not tell the whole story; through computer simulations, we will show that in one type of game HCR leads to very rapid convergence, and in another it is quite hopeless as a practical method. We will also demonstrate additional surprising and illuminating phenomena associated with the HCR rule.

It should be emphasized that co-learning is a novel framework. As we have discussed, it is a clear generalization of learning. It is also related to, but different from, dynamics of other systems, of the sort that arise in physics, biology, and economics. We will discuss connections with related work further in Section 5.

This paper is organized as follows. Section 2 formally defines the co-learning setting, the particular setting of convention evolution and cooperation evolution, and proves general convergence results of HCR in these settings. We then start to address the efficiency of convergence: Section 3 concentrates on the convention evolution setting, while Section 4 concentrates on the cooperation evolution setting. Section 5 is devoted to a discussion about related frameworks. Finally, Section 6 summarizes the main message of the article.

# 2 Social Games and the HCR Update Rule

The basic framework we present shares some features with recent work in game theory (e.g., [5]), which in turn was inspired by work in theoretical biology. In spite of this similarity, there are significant differences between the approach we take and work in game theory. We will discuss the differences in a later section; here we will develop the material in a self-contained fashion.

## 2.1 Social Games

We start by defining the standard notion of a (one-shot) *game*. Intuitively, a game involves a number of players, each of which has available to it a number of possible actions.[2] Depending on the actions selected by each agent, they each receive a certain payoff, or reward. The payoffs of the different agents are in general independent of one another, and are captured in a *payoff matrix*. Formally:

**Definition 1 [k-person game]:** A k-person game is defined by a k-dimensional matrix M, the entries of which are each a k-long vector of real numbers.

Intuitively, each dimension of the matrix represents the possible actions of the k players of the game. The $j$'th element of the vector residing in the $(i_1, i_2, \ldots, i_k)$ cell of M represents the feedback to the $j$'th player if the actions taken by all the players are $i_1, i_2, \ldots, i_k$, respectively.

In this article we will be concerned mostly with 2-persons-2-choices games (i.e., k=2 and M is a $2 \times 2$ matrix). More specifically, we will be concerned with *homogeneous games* that are defined below. In these games the role

[2] In economics the term 'strategy' is used rather than the term 'action'; in our context the latter term seems preferable.

of the two players is identical. We will discuss the intuition behind this homogeneity requirement later, in connection with the definition of action-update functions.

**Definition 2 [homogeneous game]:** Let $g$ be a 2-persons-2-choices game. Let $u_i(x, y)$ be the payoff for agent $i$ when the first agent performs action $x$ and the second agent performs action $y$. The game $g$ will be called an *homogeneous* game iff the following hold:

1. The same two actions are available to each of the two agents.
2. $u_i(x, y) = u_j(y, x)$ for all $x, y$, and $i \neq j$. Technically, this requirement causes the payoffs of one agent to coincide with the payoffs of the other agent on the major diagonal of the matrix, and to be a mirror view of the other agent payoffs on the other diagonal. Intuitively, this requirement says that the payoff for agents do not depend on the agents' names.

In the remainder of the article, a game will be understood to be homogeneous, unless specified otherwise.

A game defines the feedback that the agents will get when they *choose* (and perform) certain actions. We refer to the collection of actions chosen and performed by all agents as their *joint action*. Next we define an evaluation criteria of joint actions. Unlike in economics (see later discussion), in our setting the designer of the system has the freedom to declare the evaluation criteria. The designer will define some of the joint actions to be *successful joint actions*. Formally:

**Definition 2 [k-person social game]:** A k-person social game is defined by (i) a k-person game, and (ii) a subset of the joint actions in the game (called the *successful* joint actions).

Given an homogeneous game, a corresponding social game will be referred to as homogeneous social game. Again, in the remainder of the article we will concentrate on 2-person social games. We should remark that the fact that we have both positive-valued and negative-valued payoffs will turn out to be quite important; this stands in contrast with work in economics and genetics, in which system properties tend to be insensitive to offsetting payoffs by a constant. Here are two examples of social games featuring both positive and negative payoffs. These two games, which are well-known in the literature and which will figure prominently later in the article, capture coordination and cooperation among agents, respectively. Intuitively, the first game describes a situation in which the goal is to reach homogeneity in the society, a goal that is reflected in both the evaluation criteria and the payoff structure; it is also a a basic game in Lewis' study of conventions [7].

**Definition 3: [the convention game]**: Denote the payoff function for a particular agent by $u$.

Payoff: $u(x) = 1$ if the other agent performs $x$, and $-1$ otherwise.

Success: Joint actions in which all agents select the same action are successful; the others are not.

The second game we will consider corresponds to the well known *prisoners' dilemma* setting, of the sort studied for example in Axelrod's [2]. This game is a basic game for the study of cooperation.

**Definition 4: [the cooperation game, aka prisoners' dilemma]**: Denote the actions available to the agents by $c$ (for 'cooperate') and $d$ (for 'defect'), and the payoff function for a particular agent by $u$.

Payoff: $u$ is defined by the parameters $w, v, x$, such that $w > v > x > 0$. $u(c) = x$ if the other agent performs c; $u(c) = -w$ if the other agent performs $d$; $u(d) = w$ if the other agent performs $c$; $u(d) = -v$ if the other agent performs $d$.

Success: The joint action in which all the agents cooperate (i.e., adopt $c$) is successful; the others are not.

In fact, the term 'prisoners' dilemma' is slightly misleading here. Although the payoff matrix is the same as that encountered in economics, in fact in our setting there is no dilemma associated with it, which brings us to the last point associated with the general framework: How do agents select their actions? In economics, this question is answered to a large extent by the payoff matrix itself. Much work in economics assumes that the players have access to this matrix, and, even more strongly, that this matrix is common knowledge to the players (that is, they each know that they each know it, and so on). Based on this assumption it is then argued that certain actions are *irrational* for a player, since the player can reason that his payoff would be higher were he to take another action, no matter what other actions are taken by the other players. This gives rise to interesting static notions such as dominance, and participates in the definition of other notions such as Nash equilibrium and Pareto-optimality [9], and of dynamic notions such as evolutionary stable strategies (ess's) [12].

We do not follow this route, and instead stay closer to the spirit of reinforcement learning. In our framework, the payoff matrix is the designer's way of encoding the feedback given to the agents, but it is not accessible to the agents, nor does it uniquely determine their behavior. In particular, the matrix does not place any restriction on the way in which agents select their actions; the feedback the agents accumulate serves as input to a function which computes their next action, but that function is a new degree of freedom. We call this function the *action-update function*, or simply the *update function*. The basic question is how different update rules and other system characteristics affect the emergence of successful joint actions.

First, however, we need to define the process by which agents accumulate feedback; we do this through the following definition.

**Definition 5 [n-k-g stochastic social game]:** An n-k-g stochastic social game consists of a set of n agents, a k-person social game g, and an unbounded sequence of ordered k agents selected from a uniform distribution over the n given agents.[3]

Intuitively, a stochastic social game describes a process in which, repeatedly, random k agents meet and play the particular game. Given that agent $i$ is selected to play the role of player $j$ in the game $g$ in one of the rounds of n-k-g, $i$ must choose an action from among the actions available for player $j$ in the game $g$, *given its* (i.e., agent $i$'s) *previous history of actions and payoffs.* A well-chosen action-update function will guarantee that the agents eventually settle on a successful joint action, as defined by the social game $g$, and hopefully do so fast.

[3]The uniform-distribution assumption is made to simplify the discussion, but it can be relaxed and the results in the paper can be generalized suitably.

## 2.2 The Highest Cumulative Reward Rule

As we have said, the update function (also called the 'update rule') determines how an agent updates its behavior, based on its history of action and feedback. We would like to understand how different update rules affect the emergence of successful joint actions.

Before we begin to explore this question, we should explain an important and subtle assumption that we will be making, namely that the update rules cannot make use of specific names, either of agents or of actions. Mathematically speaking, we will enforce this condition by requiring that if names of agents or actions are changed (or, in particular, exchanged), then the names of the actions chosen during the stochastic game will be changed in a corresponding fashion. For example, if we have an update rule guaranteeing that in the stochastic cooperation game agents eventually all settle on the c ('cooperate') action, then in the modified game in which we exchange the names of the actions c and d (and otherwise leave then payoff matrix unchanged) the same rule should guarantee that the agents eventually all settle on the d ('defect') action.

The intuition behind this assumption is more important that its mathematical definition, however. We are interested in emergent successful joint action precisely in cases in which we cannot anticipate in advance the games that will be played. For example, if we know that the coordination problem will be that of deciding whether to drive on the left of the road or on the right, we can very well use the names 'left' and 'right' in the update rule; in particular, we can admit the trivial update rule which has all agents drive on the right immediately. Instead, the type of coordination problem we are concerned with is better typified by the following example. Consider a collection of manufacturing robots that have been operating at a plant for five years, at which time a new collection of parts arrive that must be assembled. The assembly requires using one of the three available attachment widgets, which were introduced three years ago (and hence were unknown to the designer of the robots five years ago). Any of the three widgets will do, but if two robots use different ones then they incur the high cost of conversion when it is time for them to mate their respective parts. Our goal is that eventually, and hopefully even rapidly, the robots will learn to use the same kind of widget. The point to emphasize about this example is that five years ago the designer could have state rules of the general form "if in the future you have several choices, each of which has been tried this many times and has yielded this much payoff, then next time make the following choice"; the designer could not, however, have referred to the specific choices of widget, since those were only invented two years later.

This explains why we do not want the update rules to rely on *action* names. The prohibition on using *agent* names in the rules (e.g., "if you see Robot 17 use a widget of a certain type then do the same, but if you see Robot 5 do it then never mind") is similarly motivated by the dynamic nature of the society; agents drop in and out of the society, denying the designer the ability to anticipate membership in advance. We definitely acknowledge that it is often useful to single out certain agents (such as Head Robot), and have them be treated in a special manner. We are very interested in the role of agents with special identities (and in particular in the role of organization structure), but even with those it is still the case in a rich setting most of the agents will not be distinguishable in this fashion. In this article we investigate the emergence of successful joint actions only in such 'faceless masses,' and completely ignore the role of personal identities.

We are now ready to start investigating useful action-update rules. In [10] we reported on preliminary results of experiments with a number of such rules. Here, however, we will concentrate on one particular update rule, called *Highest Cumulative Reward*. There are a few reasons we concentrate on this rule. First, it is a very natural one, perhaps the most basic rule one can imagine. Second, past experiments have shown it to be particularly effective in stochastic settings. Finally, we will see that, despite its simplicity, this rule gives rise to highly nontrivial phenomena. (In the following definition, recall that in this article games are by default 2-persons-2-choices games.)

**Definition 6 [HCR]:** According to the *Highest Cumulative Reward* update rule (or HCR), an agent changes it current action iff the total payoff obtained from the other action in the latest $l$ iterations is greater than the payoff obtained from the currently-chosen action in that time period.

The parameter $l$ in the above definition denotes a finite bound, but the bound may vary. In the experimental studies reported later $l$ is greater than the number of iterations (i.e., agents refer to their full history), unless stated otherwise.

We would like to understand how HCR affects the emergence of successful social behavior, and, in particular, its effects on the evolution of convention and cooperation. In fact, we are able to show a result that applies to a broad class of social games, which include the convention and cooperation games. We call these games *social agreement games.*

We start by defining a class of ordinary (one-shot, homogeneous) games. In the following, let $u_i(x, y)$ be the payoff for agent $i$ when the first agent performs action $x$ and the second agent performs action $y$.

**Definition 7:**

A social game $g$ is called an *social agreement game* iff

1. $g$ is a homogeneous social game, whose payoff matrix

| | a | b |
|---|---|---|
| a | $x, x$ | $u, v$ |
| b | $v, u$ | $y, y$ |

   has the following properties:

   (a) $x, y, u, v \neq 0$; that is, every outcome is always considered either positive or negative from the point of view of an agent.

   (b) Either $x > 0$ or $y > 0$ or both; that is, there exists a (not necessarily unique) action that, if adopted by both agents, yields positive payoff to both.

   (c) Either $u < 0$ or $v < 0$ or $x < 0$ or $y < 0$ ; that is, there exists some negative reward for failing to agree on an action that yields positive payoffs for all agents.

2. a joint action is defined to be successful if and only if it yields positive payoffs for all agents.

It is easy to see that the cooperation game and the convention game are both instances of social agreement games. The theorems below that refer to HCR assume that the parameter (memory bound) $l$ is much larger than the entries in the payoff matrix of the game. We also assume that $l \geq n \geq 4$, and that the payoffs in $g$ have finite decimal representation. We can show:

**Theorem 1:**[4] Given an n-2-g stochastic social agreement game, placing no constraints on the initial choices of action by all agents, and assuming that all agents employ the HCR rule, the following property holds: For every $\epsilon > 0$ there exists a bounded number $M$ s.t. if the system runs for $M$ iterations then the probability that the system will arrive at a situation where only successful joint actions would be chosen is greater than $1 - \epsilon$. Once we arrived at that situation, it will never be left.

**Corollary 1:** The HCR update rule guarantees eventual emergence of a convention and of cooperation, in the respective settings.

## 2.3 The efficiency of evolution: a lower bound

The above results shed some light on the eventual emergence of social behavior, but they say nothing about the efficiency with which this behavior is attained; the remainder of this article is devoted to this question. We start by presenting a general lower bound on the efficiency of this process. This will be obtained by the following definition and theorem.

**Definition 8:**

Let g be a social agreement game. Consider the $t$ iteration of an n-2-g stochastic social game, and the $n \cdot (n-1)$ games (possible agent interactions) that might be played at that iteration. We define $X_n(t)$ to be a random variable that contains the number of games that might be played in iteration $t$ and result in an unsuccessful joint action. Let $T(n)$ be a function which associates with each $n$ a number (of iterations). Given an update rule $R$, and some distribution on the initial actions of the agents, we will say that $R$ *guarantees* the emergence of successful joint actions after $T(n)$ iterations, if $E[X_n(T(n))]$ converges to 0.

Roughly speaking, we measure how far is the system from guaranteeing successful joint action. We would like this distance to be as close to 0 as

[4] Sketch of proofs appears in the Appendix.

possible in a minimal number of iterations. Notice that Theorem 1 implies that $E[X_n(M(n))]$ converges to 0. We can also show a lower bound:

**Theorem 2:**

Let $g$ be a social agreement game, and let $R$ be an update rule. Assume there is some non-zero constant probability for starting with any particular action by any particular agent. If $R$ guarantees the emergence of successful joint actions in the related n-2-g games in $T(n)$ iterations, then $T(n) = \Omega(n \cdot log(n))$

## 2.4 Reality check: how good is HCR in practice?

At this point we seem to be converging on an understanding of the dynamics brought about by HCR; at least for social agreement games, we have a guarantee of eventual emergence to successful joint actions, as well as a cautionary lower bound on how fast we can expect to arrive at such a happy occasion. It would be natural to expect that subsequent investigations would provide finer and finer lower and upper bounds, increasing our understanding of HCR.

Unfortunately, this has not been our experience. What we found instead was that rather specific properties of the particular games being played flavor the dynamics so strongly that it appears extremely difficult to arrive at general results at the level of a particular update function. We arrived at this conclusion through extensive computer simulations, which yielded results that not only had not been anticipated, but in fact have not yet been explained mathematically even after the fact.

Let us illustrate the point with the two games highlighted above, the convention and cooperation games. Both are instances of social agreement games, and hence subject to the upper and lower bounds presented in the previous section, and yet the practical experience with the two has been radically different. In the case of the convention game, the HCR rule not only led to the emergence of convention, but it did so at a rate that approaches the theoretical lower bound. In contrast, in the cooperation game the HCR rule proved to be very inefficient, rendering it useless for most practical applications. We will elaborate on these points in the following sections.

The sensitivity of the system is perhaps not very surprising, given the experience in other disciplines with complex dynamic systems (disciplines such as physics, economics, and biology; see later section). If any lesson from those disciplines carries over, it is that analytic results must at the very least be complemented by experimental ones. Certainly our experience has been that we have gained more insight into the evolution of convention and cooperation through simulation than through analysis. The remainder of this article reports on some of our more informative experimental findings.

# 3 Convention Evolution

In the previous section we showed that HCR leads to the emergence of successful joint actions, provided a lower bound on the efficiency of that process, and ended by showing that these bounds still hide many interesting properties that are specific to particular stochastic games. In this section we concentrate on the efficiency of convention evolution, and in particular on the effects of different modifications of HCR and parameters of the system on that efficiency. We concentrate on this update rule, although we investigated other update rules that led to interesting phenomena as well (see [10]), since it led to the most interesting phenomena and to efficient convention evolution (as will be discussed in 3.4). Unless stated otherwise, the experimental results appearing in this section and in the following section refer to experiments with 100 agents starting with random initial actions. Each experiment consists of many trials, each of which consists of a run of the stochastic game for a given number of iterations.

## 3.1 The effect of update frequency

The first parameter and modification we consider concern update frequency. In the previous section we assumed that each agent updates[5] its behavior at each iteration. What happens if agents update their behavior less frequently? (This condition might be imposed by internal limitations of the system, or

[5] By 'update' we mean the application of the update function; the result need not be a change in action.

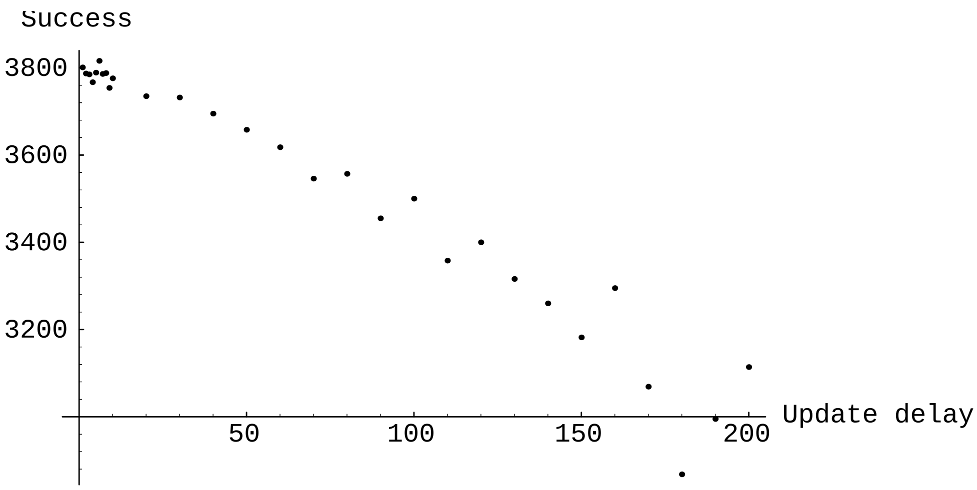


Figure 1: The effects of update frequency

alternatively might be selected voluntarily to impose greater stability on the system.) We found that when the frequency of update decreases, then the efficiency of convention evolution decreases. Our results can be illustrated by Figure 1. In this figure, the $x$ coordinates describe the distance between iterations in which update is performed, while the $y$ coordinates describe the number of trials from among 4000 trials of 1600 iterations each in which more than 95% of the agents reached a convention.

### 3.2 The effect of memory restarts

We investigated the effects of memory size on the efficiency of convention evolution. We consider two forms of limited memory; one is treated in this subsection, and the other one will be treated in Section 3.4. One type of limited memory is a memory that is restarted from time to time. When the memory is restarted, the agents' current actions (the ones they will now start with) are not forgotten, but previous history is. This might be in particular interesting in systems which stop operating for a short while from time to time. For example, a society might be interested in a particular coordination only in some periods of the year, where agents are assumed to forget what they have exactly seen in the previous periods although they still remember their current (latest) action. We investigated the efficiency of convention

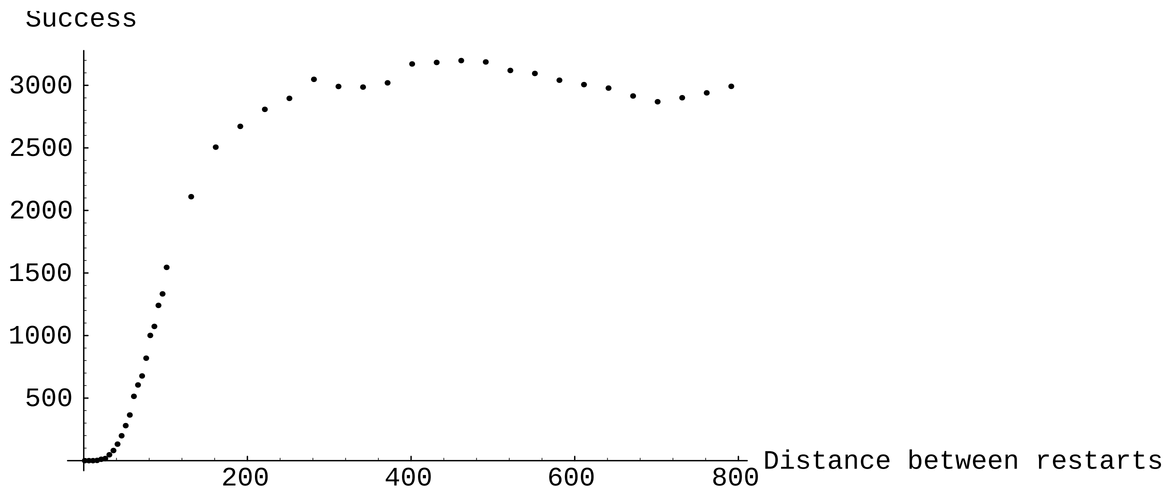


Figure 2: The effects of memory restarts

evolution as a function of the frequency of memory restarts. We found that when the distance between iterations where the memory is restarted decreases, then the efficiency of convention evolution decreases. This can be illustrated by Figure 2. The $x$ coordinates of this graph correspond to the distance between iterations where the memory is restarted. The $y$ coordinates describe the number of trials from among 4000 trials of 800 iterations each, in which more than 85% of the agents reached a convention.

The reader may be tempted to treat this as an 'obvious' result; however, full memory is not always an advantage. Sections 3.3 and 3.4 will provide some examples; here is another example. We ran an experiment in which agents restarted their memory *always and only* after changing their action. In that case the evolution of convention was even more efficient than in the case of full memory; in 3298 from among 4000 trials of 800 iterations each, more than 85% of the agents reached a convention (while with complete information this was true of only 3010 of the trials.)

## 3.3 Co-varying memory size and update frequency

We have so far varied update frequency and memory independently; we now show that these two parameters interact. Consider the results from section

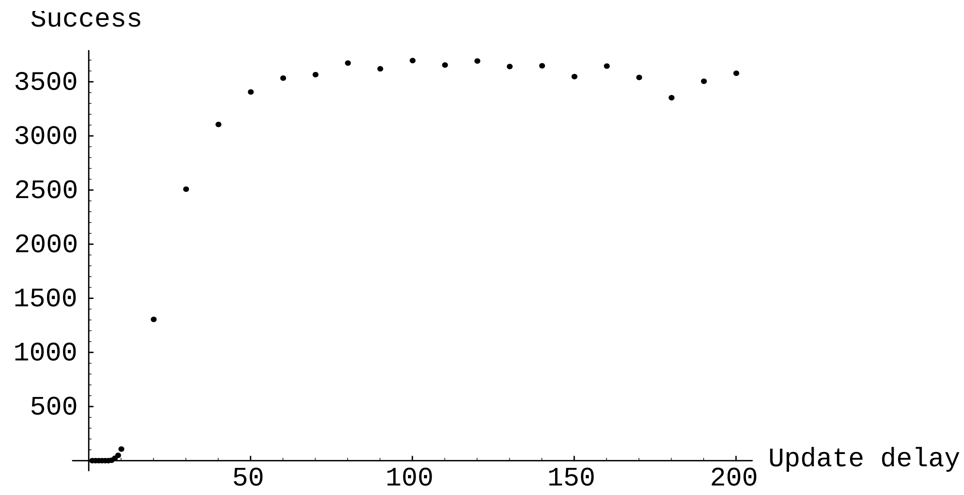


Figure 3: The case in which update frequency = memory restart frequency

3.1, where we showed that the rate of convention evolution is a monotonic increasing function of update frequency. We now show that decreasing memory blocks the degradation of convergence with the decrease in update frequency. Specifically, in this experiment we adopted the memory-restart model, and varied together the memory-restart frequency and the update frequency; that is, at the end of each window each agent updated its choice according HCR for that window. The general result we obtained is that when update becomes infrequent (there is a long delay between action updates), then it is better to restart the memory from time to time than to rely on the whole memory. Our results are illustrated in Figure 3. The $x$ coordinate of this figure corresponds to the update frequency, which is equal to the number of iterations between consecutive memory restarts. That is, in this case, we had a single interval which served both as the update frequency and the memory restart frequency. The $y$ coordinates correspond to the number of trials from among 4000 trials of 1600 iterations each in which 95% of the agents reached a convention. It is illuminating to compare Figure 3 to Figure 1 (where full memory is assumed); when the update frequency drops below about 100 iterations, it becomes better to use the statistics of only the last window than to rely on the entire history.

## 3.4 Limited memory windows

A more continuous form of limited memory is one in which each agent at each time keeps a limited window into its past experience, and bases the HCR rule on only that window. We have considered two forms of windows, one in which it remembers the last $n$ iterations in which it participated in a meeting, and another in which the agent remembers the last $n$ iterations, regardless of whether it participated in a meeting in those.

Our results of these two experiments are illustrated in Figures 4 and 5, respectively. In both of these figures the $x$ coordinates describe the size of the memory window, and the $y$ coordinates correspond to the number of trials from among 4000 trials of 800 iterations each, in which more than 85% of the agents reached a convention. Note that, somewhat surprisingly, in both cases it pays to forget, though some minimal memory is essential (in the first case this minimum is in fact equal to 2 iterations, and therefore this can be seen more easily in the second case).

A right choice of the memory window while applying HCR will give us in fact an update rule which is a close to optimal update rule. The case where the memory size is between $2n$ to $3n$ (where $n$ is the number of agents) gives us the above-mentioned close to optimal behavior, which is in fact a speed of convergence of $O(n \cdot log(n))$. More specifically, given that there are $n$ agents who adopt HCR with a memory window $3n$, we observed that all of the agents reach a convention after less than $3n \cdot log(n)$ iterations (when we vary the number of agents.) The optimality stems from the above fact and from Theorem 2. The important point is that HCR with an appropriate limited memory window can be supplied to the agents as an update rule that will enable an efficient convention evolution in a system where there are no update delays.

## 3.5 More complicated decisions

The convention game captures a situation where a selection among a pair of successful joint actions has to be made. This can also be considered as a selection of an option from among two possible options, without an a-priori agreement about which option should be chosen. What happens if

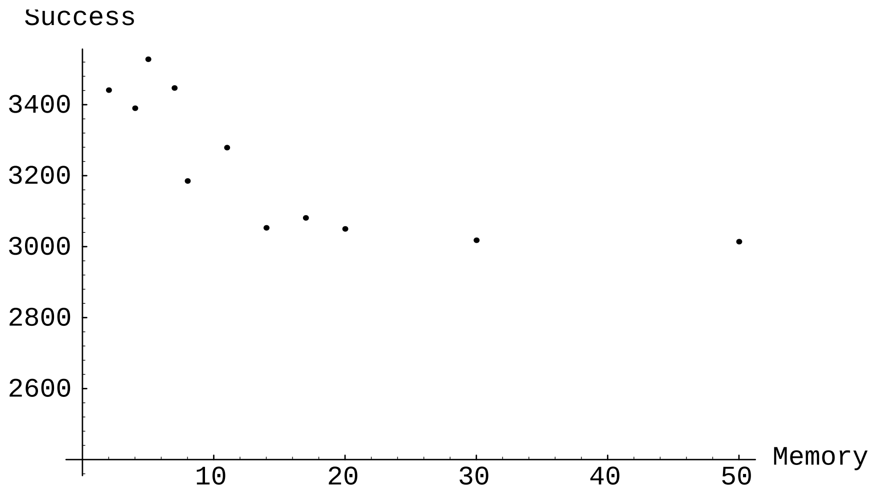


Figure 4: Limited Memory (latest observations)

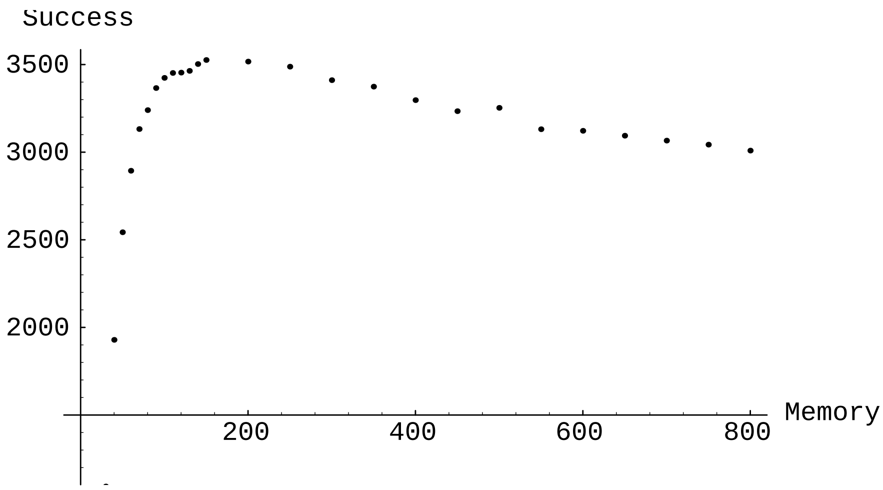


Figure 5: Limited Memory (latest iterations)

the agents have to agree on an option from among more than two available options, that is, on something more complicated than a bit? How does the number of options (potential conventions) affect the efficiency of convention evolution?

In order to answer the above question, we use the following observation: whenever an agent performs a particular action and gets a particular feedback in a convention game, it can interpret it as an observation of the action used by the agent it encountered. For example, if the agent performs action $a$ and gets a feedback of 1, we can say that the agent observed that another agent used the action $a$ as well. Having the above interpretation for the feedback, we can define:

**Definition 9:** The External majority (EM) update rule is an update rule which says: Adopt action $i$ if so far it was observed in other agents more often than the other action and remain with your current action in the case of equality.

We can show:

**Lemma 1:** EM coincides with HCR in the convention evolution setting.

Given the above Lemma, we can assume that the agents use EM and not HCR in the convention game. The advantage behind the use of EM is that there exists a natural extension of it to the case of more than two possible conventions: An agent will adopt the action it observed most often until the given point. Because of Lemma 1, this update rule extends HCR in a precise and natural way. Hence, we assume that the agents adopt this extension of EM.

Our general results are as follows. What we find is that adding more potential conventions decreases the efficiency of convention evolution in a less than logarithmic fashion. In addition we find that the absolute amount of success in convention evolution decreases in less than logarithmic fashion: In order that the number of successes of convention evolution will decrease by factor of 2, we need to increase the number of potential conventions by a factor of more than 4, and in order to decrease it by a factor of 3 we need to increase the number of potential conventions by a factor of more than 8.[6]

[6]We have verified these basic results also in the case of limited memory.

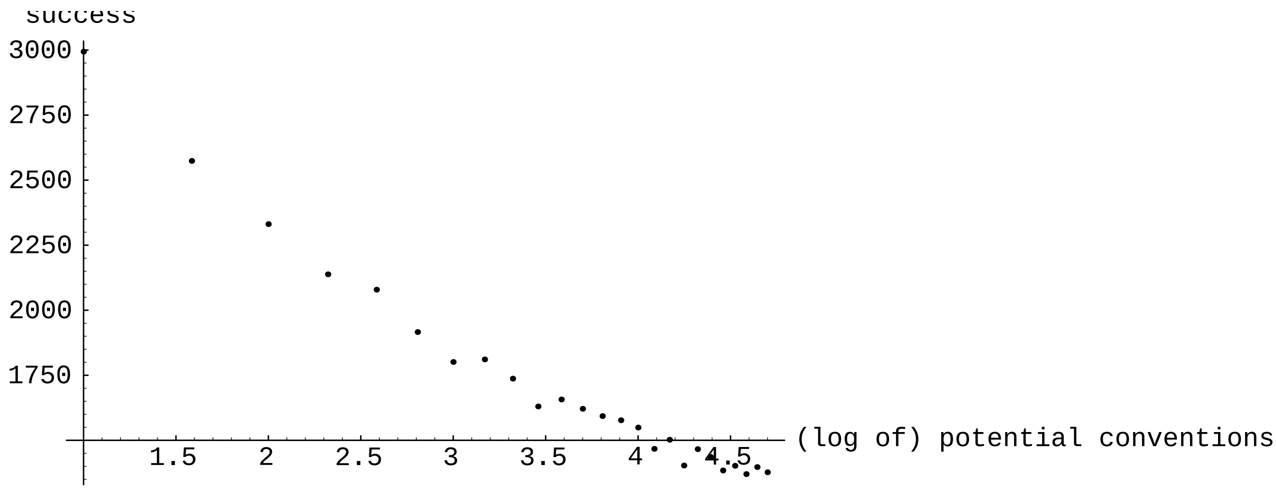


Figure 6: The effects of the number of potential conventions

Intuitively speaking, our results point to the following encouraging fact: the efficiency of convention evolution is not affected too badly by an increase in the number of potential conventions.

Some specific results are illustrated in Figure 6. The $x$ coordinate describes on a logarithmic scale the number of potential conventions, while the $y$ coordinate describes the number of successful trials (more than 85% reached a convention) from among 4000 trials of 800 iterations each.

# 4 Cooperation Evolution

As we see, the HCR update rule gives us close to optimal behavior as far as the efficiency of convention evolution is concerned. One might have hoped that it will give us also efficient cooperation evolution. However, although we proved that the HCR update rule yields the emergence of cooperative behavior, we are able to show that it is hopeless in the cooperation setting. We observed that phenomena and the other phenomena mentioned in this section for many instances of the cooperation game, but in order to demonstrate our results we will use a particular assignment of the game parameters: $x = 2, v = 5, w = 6$ and $n = 100$. Our experiments show that (e.g., in the case of 100 agents) even after hundreds of thousands of iterations we get that many of the agents are non-cooperative. Moreover, even when we start with a small number of non-cooperative agents the society evolves in a way that many additional agents become non-cooperative, and even after hundreds of thousands of iterations we never reach a situation in which most of the agents are cooperative!

In the previous section we experimented with the effects of various parameters on the efficiency of convention evolution. We showed that an appropriate modification of several parameters (such as the memory size) improves the efficiency of convention evolution. We have therefore investigated how these parameters affect the efficiency of cooperation evolution. As it turns out, most of the related modifications (changing memory size, changing the update frequency, etc.) produced no real change; we remain with the theoretical convergence to cooperation, and total inefficiency of the process in practice. However, some of the modifications did lead to insights, and in this section we describe one modification that revealed particularly illuminating phenomena. This modification is concerned with the ability of agents to communicate and exchange their past histories with one another. In the original framework we assumed that when agents meet, they play an instance of the prisoners' dilemma game. Here we allowed *some* pairs of agents to simply exchange their past histories when they are selected to meet, rather than play the game. The fact that some pairs of agents play the game and some don't introduces a form of non-homogeneity to the setting; it turns out that this non-homogeneity gives rise to a surprising phenomenon. We describe this phenomenon in two non-homogeneous settings: one in which each agent has its own 'neighborhood of friends,' and another in which the entire populations is divided into two sub-societies.

### 4.1 The effects of communication and initial conditions in a neighborhood-based setting

The model embodied in the following definition is that each agent has a certain set of friends. (This set happens to be defined as follows. The agents are arranged on a ring, and two agents are friends if the distance between them along the ring is less than some threshold. Other definitions of neighborhood are possible, of course.) The stochastic game is played; when two non-friends meet they play an instance of the prisoners dilemma (with full memory of the past), but when two friends meet they simply exchange histories, and set each of their histories to be the combined history of both.

**Definition 10:** An n-2-g stochastic social game with *communication radius* $k$ is an n-2-g stochastic social game where when agents $i$ and $j$ are selected, and $j \in [i-k, i+k](mod\, n)$, then agents $i$ and $j$ exchange their past histories instead of playing $g$. If agents $i$ and $j$ are selected but $j \notin [i-k, i+k](mod\, n)$, then $i$ and $j$ play $g$. Given that the agents obey the HCR update rule we assume that whenever a pair of agents communicate instead of participating in $g$, they update their benefits from action $x$ to be the sum of their respective individual benefits from $x$ until the given point.

In this section we will take $g$ to be the prisoners' dilemma.

The above communication structure is commonsensical and can be related to our everyday experience. Therefore, we urge the reader to speculate about the following: Assuming that most of the agents are initially cooperative, how does the parameter $k$ affect the evolution of cooperation? Does it speed up the process and encourage the non-cooperators to change color, or does it slow the process even more? Before we answer this question, let us first establish an upper bound. The following theorem shows that, if there are at least two non-friends among the agents, then eventual convergence to cooperation is guaranteed also in the case of communicating agents.:

**Theorem 3:** Given an n-2-g stochastic social game with communication radius $k < n-1$, where $g$ is the cooperation game, and where agents start with any action and use the HCR update rule, the following property holds: for every $\epsilon > 0$ there exists a bounded number $M$ such that if the system runs for $M$ iterations then the probability that the system will arrive at a situation in which a successful joint action is chosen is greater than $1-\epsilon$. Once we arrived at that situation, it will never be left.

Although the result is interesting by itself, we have found that it hides surprising phenomena. Specifically, we have shown the following through simulations:[7]

[7]In the experimental results appearing in this section we assume an upper bound on the differences an agent might have between its benefits given different strategies. When the difference is higher than a given (large) number, we assume that it is cut off to that number. This is done in order to prevent computational overflow.

- The evolution of cooperation is nonmonotonic in the communication radius.
- The evolution of cooperation is nonmonotonic in the initial number of non-cooperative agents.

In other words, if we hold the number of initial cooperators fixed, then having either no communication at all or very rich communication is superior to having only some communication Even more interestingly, for any fixed communication radius above a certain threshold, having either most of the agents start out as cooperators or most of them start out as non-cooperators is vastly superior to having some of them start out as cooperators and some as non-cooperators; in fact, while in the former cases we were often able to obtain almost total cooperation in the society, in the latter case the system never converged, even after millions of iterations. The fact that it is advantageous to start out with all non-cooperators is surprising, and at first blush perhaps even paradoxical; after all, on the way from becoming all non-cooperators to all cooperators, the nonconcurrent nature of our system forces it to pass continuously through all states with more even mixes of cooperators and non-cooperators. How can it then be that if we start out in those intermediate states the system never converges in practice? The answer lies in the fact that, in the presence of communication, the distribution of cooperators and non-cooperators does not uniquely determine the state of the system. In particular, when we start out with an even mix, the agents have no statistics about the state of the system; however when we start out with all non-cooperators, by the time the system arrives at a more even mix the agents have already accumulated substantial statistics about its current state.

The results are best illustrated in a three-dimensional chart. We observed the phenomena of nonmonotonicity in many simulations, in which the number of iterations varied from several thousands to one million. We demonstrate a particular set of results in Figure 7.[8] The $X$ coordinate describes the initial ratio of non-cooperative agents, where $X = x$ means that

[8] For particular initial configurations, it is hard to observe from that graph the fact that the resulting number of cooperative agents is non-monotonic in the communication radius. This is due to the fact that in these cases the resulting situation always include many non-cooperative agents.

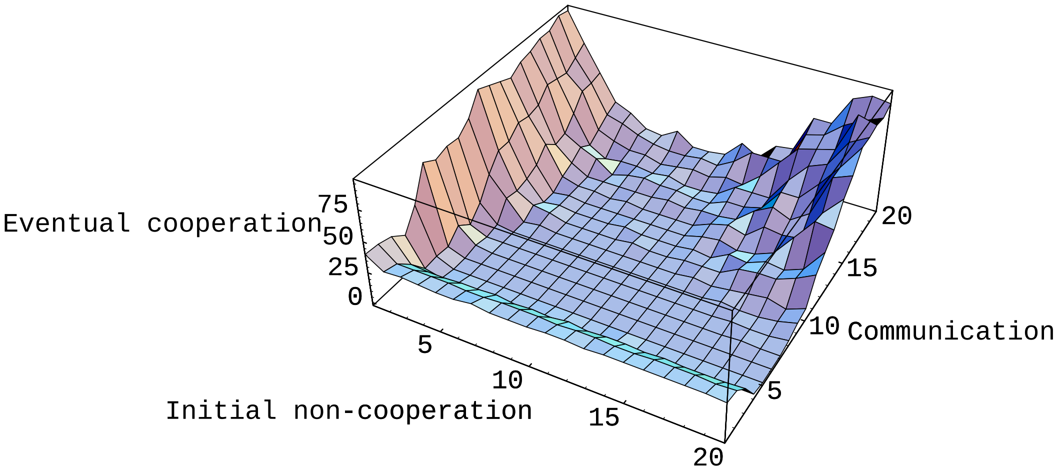


Figure 7: Cooperation as a function of initial cooperation and communication radius

the probability for each agent to be initially non-cooperative is $\frac{(2x+1)}{40}$. The $Y$ coordinate describes the communication radius, where $Y = y$ means that the communication radius is $2y$. The $Z$ coordinate describes the average number of cooperative agents in 50 trials after 5000 iterations in each of the trials.

### 4.2 The effects of communication and initial conditions in a disjoint sub-societies case

The previous subsection described a natural case of adaptive behavior when communication among agents is feasible, and pointed to an interesting property of it. However, adaptive behavior of a similar type can be discussed also for other types of communication structures. A natural and complementary version of communication structure is concerned with non-overlapping sub-societies.

**Definition 11:** A bipartite n-2-g stochastic social game, is an n-2-g stochastic social game in which, when agents $i$ and $j$ are selected, they exchange their past histories if $j <= \frac{n}{2}$ and $i <= \frac{n}{2}$, or $j > \frac{n}{2}$ and $i > \frac{n}{2}$, and play $g$ otherwise. Given that the agents obey the HCR update rule we assume that whenever a pair of agents communicate instead of participating in $g$, they update their benefits from action $x$ to be the sum of their respective individual benefits from $x$.

As before, the property of eventual convergence is retained:

**Theorem 4:** Given a bipartite n-2-g stochastic social game, where $g$ is the cooperation game, and where agents start with any action and use the HCR update rule, the following property holds: for every $\epsilon > 0$ there exists a bounded number $M$ s.t. if the system runs for $M$ iterations then the probability that the system will arrive at a situation where only successful joint actions would be chosen is greater than $1 - \epsilon$. Once we arrived at that situation, it will never be left.

And again, as in the previous subsection, the above theorem hides an interesting phenomenon:

- The evolution of cooperation is nonmonotonic in the initial conditions.

More precisely, when almost all the agents start out as cooperators, the system ends up with many cooperators. The number of final cooperators degrades as the number of initial non-cooperators increase, but only up to a certain point. Beyond that point, increasing the number of non-cooperators actually increases the number of final cooperators.

Here too we observed the phenomenon in many simulations, in which the number of iterations varied from several thousands to one million. We demonstrate a particular set of results in Figure 8. The $X$ coordinate describes the ratio of initially non-cooperative agents, where $X = x$ means that the probability of each agent to be initially non-cooperative is $x$. The $Y$ coordinate describes the average number of cooperative agents in 20 trials and after 200000 iterations in each of the trails.

# 5 Related work on complex dynamic systems

Several lines of research are related to our work. These include work in population genetics, statistical mechanics, computational ecologies, quantitative

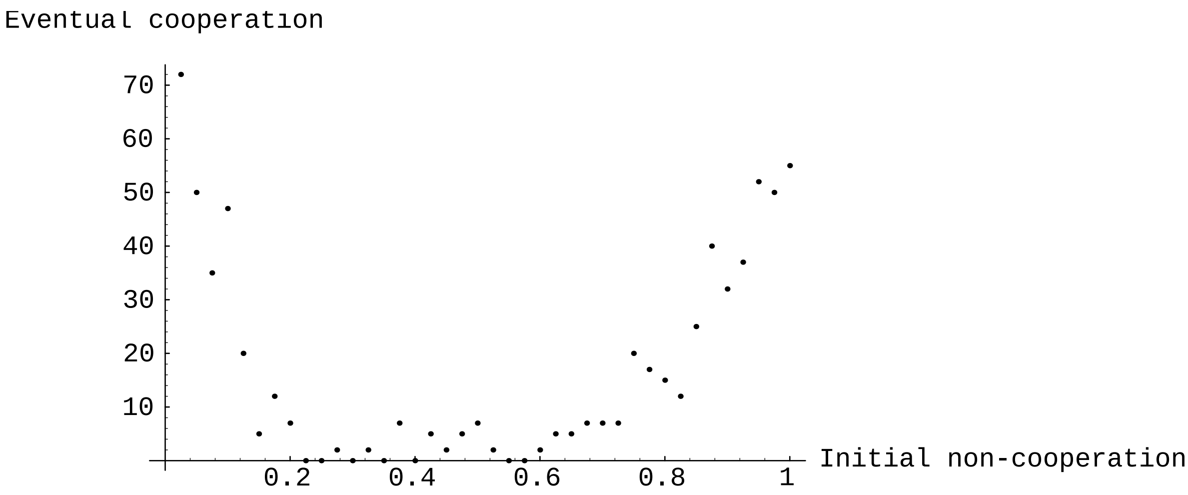


Figure 8: Cooperation as a function of initial cooperation at the disjoint sub-societies case

sociology, and mathematical economics. The discussion in this paper would not be complete without at least a brief description of the work carried out in these fields. We must acknowledge, however, that the combined body of material is so rich that neither we, nor anyone else with whom we have discussed these matters (and we have discussed them extensively), fully understand the connections between the different lines of work. Each of these involve a setting with multiple elements (whether they are called particles, individuals, cells, or agents), which repeatedly undergo relatively simple local changes. The questions usually asked center around interesting global system properties that emerge over time out of these local changes, such as convergence and phase transitions. It is often tempting to try to carry over lessons from one setting to another. Indeed, some of these areas were inspired by one another; for example, work in quantitative sociology was inspired by work in statistical mechanics, and work in economics was inspired by work in population genetics. However, these inspirations have tended to be in spirit rather than in detail; the actual dynamic systems in the various areas are for the most part quite different, and also very sensitive in the sense that even small changes in them result in quite different system dynamics. This has also been our experience with our own framework; initially we had hoped to borrow results from other areas, but our framework then turned out to be sufficiently different from any of the others so as to make such borrowing impossible, or at least very difficult. We are still very much interested in understanding technical connections with these other areas and our own work, anticipating cross influences, but at this point all we will do is briefly describe work in these other areas.

Statistical mechanics models are a powerful tool for explaining a variety of phenomena in physics. An important family of models in that area goes under the general name of the *Ising model* [6]. In a typical Ising model we have a set of spins, each of which can be in -1/1 state, and which are organized into some fixed spatial arrangement (such as a one-dimensional sequence or a two-dimensional grid). At each point in time the system is in some configuration (that is, the spins each have a particular value), and this system has a certain measure energy, or entropy. The energy has a component representing local interactions among the spins, and a component (that is sometimes omitted) representing the effect of some global magnetic field. The interaction among spins is usually limited to neighboring spins; a typical formula for the energy of the system will include the sum of all multiples $x \cdot y$, where $x$ and $y$ are the values of neighboring spins. In terms of this energy a probability distribution is defined over the space of all configurations, which determines the likelihood of the system actually being in any particular configuration. This probability distribution has strong independence properties; the probability of a particular spin having a particular value is sensitive only to the values of the neighboring spins, and is independent of the values of spins that are spatially removed from it.

The Ising model has proved useful for the investigations of various physical properties such as spontaneous magnetization (that is, a majority of spins ending up with the same value), but it is also abstract enough to have motivated other applications. For example, in some work within *quantitative sociology* the spins were interpreted as 'opinions,' the energy between individual spins as 'tension among individuals with differing opinions,' and the the orientation of the magnetic field as 'the opinion of the government.'

As described, the Ising model does not provide a dynamical system, in the sense that it does not provide (e.g. differential) equations that describe the evolution of the system over time; what it does instead is define stable (i.e., low energy) states of the system. This is also true for the more elaborate framework of spin-glass. However, both have been augmented to include a dynamic component, and these dynamic models have also have found applications in other fields. One is quantitative sociology, where the models have been used to predict opinion shifts over time within large populations [16]. Statistical mechanics also provided the inspiration to *Computational Ecology* [4]; this work is based on the idea that the existence of many agents in an advanced computerized framework creates a computational ecology in which agents cooperate as well as compete with one another. A computational framework, similar in its spirit to quantitative sociology, is developed and analyzed using the tools of statistical mechanics. The notions used there are 'strategies' of individual agents, the utility of having identical strategies ('cooperation') as well as its disadvantage, due to resource conflicts ('competition'). A precise continuous framework is built, which allows several predictions on the behavior of those "computational ecologies" (such as chaotic behavior in some situations).

These multi-agent frameworks borrow a powerful tool from statistical mechanics, but as a result they have a heavily 'non-local' or 'non-mechanical' flavor; the dynamics speak about how certain global statistics change over time (such as the average number of cooperating agents), rather than about how an individual agent changes its local state on the basis of its current local state and/or history. This of course is quite unlike our own framework, in which the dynamics of the atomic elements are the basis for change, and any statistical properties are derived from these.

Work in *population genetics* [1] is closer to ours in this sense. Here we have a set of individuals, each belonging to one of several types. The system evolves in 'generations'; in each generation the individual evolves in a way that is defined by its type and the environment (which includes the other agents), and at the end of the generation the 'fitness' of each agent is computed, applying some given fitness function. The probability that an individual will survive into the next generation is proportional to its fitness. Additionally, usually between generations a process of 'recombination' takes place, in which some pairs of individuals in the populations (the 'parents') may combine to produce a new individual ('the offspring'), whose type will in general be defined by, but different from, the types of the parents.

This setting is thus more local, or mechanical, than the frameworks discussed earlier; the activity of each agent within a generation, its fitness function, and the reproductive process all have a transition-oriented, automatalike flavor. An important global component remains, however, namely the fitness function. The fitness function, which represents unspecified external selection forces, is computed on the basis of global system properties, and is applied to all agents equally. For example, if we consider the earlier 'opinion' example, a typical definition of the fitness of an individual with an opinion in a population would be the proportion of individuals in the population having the same opinion. This global element turns out to have a strong influence on the dynamics of the system. (This is perhaps the deepest difference between the population genetics setting and our own; but of course may other difference exist, including the notion of an accumulated history, limited memory, communication, and our particular stochastic process of encounters.)

The model used in population genetics has strongly influenced work in *mathematical economics*; this is especially true of work published in recent years. In general, mathematical economics has attempted to explain and predict behaviors of 'rational' agents. The notion of rationality is based on associating a well-defined 'utility' with alternatives actions (or 'strategies') available to an agent; a rational agent is that which selects actions which maximize its utility. A central concept in mathematical economics is that of a Nash-equilibrium: a joint behavior, deviation from which by any one agent will lower that agent's utility (and hence every agent, being rational, will not deviate from it, and the state will remain stable). Nash equilibria are clearly an important notion, as are others such as Pareto optimality. We have already discussed in the paper why these notions do not hold any a priori special significance in our setting.

The work in economics that is closest in spirit to ours is that on 'evolutionarily stable strategies', or ess's [5]; this work is also the most strongly influenced by population genetics within economics. A typical setting in this line of research looks as follows. Agents within a given population meet each other randomly, and when they do so they play some particular pre-defined game (such as the prisoners' dilemma). For a while every agent plays the game in a fixed fashion (that is, the agent does not switch strategies). But then, after a certain period (a 'generation'), every agent individually calculates what would have been the best choice in hindsight for that period, and switches to that choice; this is called the 'best response' rule. The question is then asked whether the system will converge to a certain state, or exhibit other interesting behaviors.

There are clear similarities between this framework and ours, but, again, also significant differences. First and foremost, in that work there is an assumption that the periods are long, long enough for almost all agents to glean almost perfectly reliable statistics on the state of the system. This assumption, inherited directly from the global nature of the fitness function in population genetics, is directly at odds with our framework. Other important differences include the absence of any system parameters such as memory length and communication. It is also the case that whereas we have been concerned mostly with the efficiency of social change, work in this area of economics seems to have neglected this entirely, concentrating only on qualitative notions such as convergence and oscillation. Still, this area is probably the closest in spirit to our work, and we are interested in developing stronger links with it.

# 6 Summary

We have defined the notion of stochastic social games, defined two that are closely related to the economics literature, introduced a particular update rule called HCR, and investigated its properties – first giving coarse analytic results, and then reporting on finer grained experimental ones.

Beside the novelty of our work, we believe that it also creates a bridge between work in economics and work in machine earning, two disjoint areas heretofore, and in the process generalizes both. We generalize research in machine learning by introducing co-learning, and we adapt the evolutionary approach discussed in economics by importing computer-sciency elements in the spirit of reinforcement learning. More specifically on the latter, since our perspective is that of system designers, we introduce into the game-theoretic framework additional degrees of freedom: an external criterion for satisfactory social behavior (in contrast with internal rationality criteria such as Nash-equilibrium), and an external update rule which governs the way in which each agent updates its behavior based on purely local information.

As was mentioned in the introduction, this work is part of our work on multi-agent systems. In addition to continuing the investigation described here, we are interested in applying the lessons to system design. We are currently experimenting with co-learning-based dynamic load balancing, and hope to report on it in the future.

# References


[1] L. Altenberg and M. W. Feldman. Selection, Generalized Transmission, and the Evolution of Modifier Genes. I. The reduction principle. *Genetics*, pages 559–572, November 1987.

[2] R. Axelrod. *The Evolution of Cooperation.* New York: Basic Books, 1984.

[3] A. Blumer, A. Ehrenfeucht, D. Haussler, and M.K. Warmuth. Occam's Razor. *Information Processing Letters*, 24:377–380, 1987.

[4] Bernardo A. Huberman and Tad Hogg. The Behavior of Computational Ecologies. In Bernardo A. Huberman, editor, *The Ecology of Computation.* Elsevier Science, 1988.

[5] M. Kandori, G. Mailath, and R. Rob. Learning, Mutation and Long Equilibria in Games. Mimeo. University of Pennsylvania, 1991.

[6] R. Kinderman and S. L. Snell. *Markov Random Fields and their Applications.* American Mathematical Society, 1980.

[7] David Lewis. *Convention, A Philosophical Study.* Harvard University Press, 1969.

[8] Y. Moses and M. Tennenholtz. On Computational Aspects of Artificial Social Systems. In *the Proceedings of DAI-92*, 1992.

[9] G. Owen. *Game Theory (2nd Ed.).* Academic Press, 1982.

[10] Y. Shoham and M. Tennenholtz. Emergent Conventions in Multi-Agent Systems: initial experimental results and observations. Proceedings of KR-92.

[11] Y. Shoham and M. Tennenholtz. On the Synthesis of Useful Social Laws for Artificial Agent Societies. In *Proc. of AAAI-92*, pages 276–281, 1992.

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

[13] R.S. Sutton. Special issue on reinforcement learning. *Machine Learning*, 8(3–4), 1992.

[14] L. G. Valiant. A theory of the learnable. *Comm. ACM*, 27(11):1134–1142, 1984.

[15] C.J.C.H. Watkins. *Learning With Delayed Rewards.* PhD thesis, Cambridge University, 1989.

[16] W. Weidlich and G. Haag. *Concepts and Models of a Quantitative Sociology; The Dynamics of Interacting Populations.* Springer-Verlag, 1983.

## Appendix: Sketch of Proofs

**Proof of Theorem 1 (sketch):**

Recall that the payoff matrix of a social agreement game has the structure

| | a | b |
|---|---|---|
| a | $x, x$ | $u, v$ |
| b | $v, u$ | $y, y$ |

in which $x, y, u, v \neq 0$, either $x > 0$ or $y > 0$ or both, and either $u < 0$ or $v < 0$ or $x < 0$ or $y < 0$.

We prove the theorem by case analysis. In principle we need to examine 16 cases, each case determined by the polarity (either positive of negative) of $x, y, u, v$. However, since we can assume without loss of generality that $x > 0$, and since the case in which $u, v, x$ and $y$ are all greater than 0 is ruled out, we are left with seven cases. Notice that the cooperation game is a special case of the case in which $y < 0, u < 0, v > 0$, and the convention game is a special case of the case in which $y > 0, u < 0, v < 0$. We will provide the proof for these two cases; proofs for the other cases can be obtained in a similar fashion.

Consider the case where $y > 0, u < 0, v < 0$. First, observe that there always exists a pair of agents with identical actions. Then, notice that the following process can be generated with a probability $p = \frac{1}{f(n)}$ and leads to a successful joint action (all agents will adopt the same action) in $g(n)$ iterations, where both $f(n)$ and $g(n)$ are bounded by an exponent of the form $n^s$ where $s$ is polynomial in $l$ (the memory size) and $n$. The process is defined as follows: a pair of agents $(i, j)$ with the same action is selected and meet each other until all of the rest of the agents forget their past. Afterwards, $i$ meets a member $x \neq j$ , and afterwards meets $j$. The last step continues in a loop where at each time $i$ meets a new $x$ until it meets all the members in the society. It is easy to see that this process will bring to a successful joint action (all agents will adopt the same action). As a result, if the system runs for $M = k \cdot g(n) \cdot f(n)$ iterations then the probability that a successful joint action will not be reached (not all of the agents will adopt the same action) is at most $e^{-k}$. Taking $k > -log(\epsilon)$ yields the desired result.

Consider the case where $y < 0, u < 0, v > 0$. In the sequel we will refer to an agent who adopts the action $a$ as a "cooperative" agent and to an agent who adopts $b$ as a "non-cooperative" agent. The structure of the proof is as the structure of the proof regarding the case where $y > 0, u < 0, v < 0$, but the basic process will now change. This process will now at first guarantee that there will be at least two cooperative agents. In order to guarantee this, the process will include in its beginning a procedure of creating a pair of cooperative agents (if no such pair exists). This procedure selects two non-cooperative agents and two additional agents, and let the latter pair meet until the former pair will forget its past. Afterwards the process selects the former (non-cooperative) agents to participate in a meeting. This will create a pair of cooperative agents. In a second stage this pair of cooperative agents will meet until the other agents will forget their past, and then pairs of non-cooperative agents will meet sequentially. This will create a society where at most one agent is non-cooperative. In order to make this agent cooperative the process will end with the following procedure: the non-cooperative agent will meet a cooperative agent until it will become non-cooperative as well, and then a pair of cooperative agents will be selected and meet each other until the rest of the agents will forget their past. The process will end by an encounter in which the two non-cooperative agents meet each other.

The above process will take place with probability $p = \frac{1}{f(n)}$ and will guarantee that after $g(n)$ iterations all of the agents will become cooperative, where appropriate exponential bounds can be given for $f(n)$ and $g(n)$. Hence, the number $M$ can be calculated as for the convention game, and the desired result can be obtained.

∎

**Proof of Theorem 2 (sketch):**

Let $Y_n(m))$ be a random variable which contains the number of agents that did not participate in any iteration of n-2-g until iteration $m$. It is easy to see that $E[X_n(m)] \geq k \cdot E[Y_n(m)]$ for some constant $k > 0$ and for every $n$ and $m$. In particular, $E[X_n(T(n))] \geq k \cdot E[Y_n(T(n))]$ for every $n$. Hence, it suffices to show that if $E[Y_n(T(n))]$ converges to 0 as a function of $n$, then $T(n)$ is at least of the order of $n \cdot log(n)$. The probability that a particular agent will not be chosen along $T(n) = (n-1) \cdot f(n)$ iterations is bounded by $(1 - \frac{1}{(n-1)})^{2 \cdot (n-1) \cdot f(n)}$ which converges to $e^{-2f(n)}$. If $e^{-2f(n)} > \frac{1}{n}$ then we will get that $E[Y_n(T(n))] > 1$ and hence there is no convergence to 0. But, in order to have $e^{-2f(n)} \leq \frac{1}{n}$ we must have $f(n) \geq 0.5 \cdot log(n)$ (where we consider w.l.o.g the natural log). This gives us the desired lower bound.

∎

**Proof of Lemma 1 (sketch):**

For ease of exposition let us denote the actions as 0 and 1, and let $c_i$ be the accumulated payoff for action $i$ of a given agent $j$. Notice that $c_i$ equals to the number of times that $j$ met an agent which used $i$ minus the number of times $j$ met an agent which used $1-i$, when its ($j$'s) action was $i$. According to HCR, an agent chooses $c_i$ if it is larger than $c_{1-i}$, but $c_0 - c_1 =$ (number of times you met 0 when you had 0 minus number of times you met 1 when you had 0) - (number of times you met 1 when you had 1 minus number of times you met 0 when you had 1), which equals to the number of times you met 0 minus the number of times you met 1. Hence, we get that the comparison between the accumulated payoffs coincide with the comparison between the number of times the different actions were encountered in other agents. This gives us the desired result.

∎

**Proof of Theorem 3 (sketch):**

The proof is similar to the proof of the cooperation game case of Theorem 1, but the basic probabilistic process will be changed a bit: 1. The pair of agents selected in the beginning of the process should be from different sub-societies. 2. The above pair of (initially cooperative or which becomes cooperative using the above-mentioned process) agents will meet until the rest of the agents forget their past; Afterwards, a cooperative behavior can be spread out in the society using the communication mechanism. ∎

**Proof of Theorem 4 (sketch):**

The proof is similar to the proof of Theorem 3. ∎