Coordination Games with Sequential Stochastic Learning and Language Emergence

James W. Shearer

JamesWShearer@gmail.com

Abstract

Lewis signaling game (LSG) and similar coordination games have been used to model the emergence and evolution of language. However both Nash equilibria and learning or evolutionary dynamics often result in suboptimal signaling systems. We present a sequential reinforcement learning (SRL) model based on a novel sequential binary decision process. SRL has low cognitive demands and parameter count and exhibits lateral inhibition without additional assumptions. We prove all scenarios converge to an optimal signaling system in all N state, N signal LSGs with arbitrary state probabilities and further explore its properties with numerical simulations. Next, we model a signaling game with agents who both speak and hear while using one state of learning (instead of two, as is common). Agents have a probability distribution for meanings in a given context. Speaking agents use the distribution to choose a meaning and use SRL model to choose a signal. Hearing agents use Bayes to combine their state of learning with their meaning distribution to guess a meaning. An agent's state of learning is reinforced from habit of speaking and guessing a meaning. Numerical simulations indicate both agents converge to the same optimal system without external reinforcement as happens in language acquisition.

Keywords: Lewis Signaling Game; Reinforcement Learning; Optimal Signaling Systems; Bayes Theorem; Language Emergence

Introduction

Although game theory has its origins in parlor games, once formalized (Von Neumann & Morgenstern, 2007), it was applied to problems ranging from nuclear deterrence to economic behavior to animal conflict (Osborne, 2004). Players in a game choose among strategies with the goal of maximizing their payoff which depends upon the strategies played by others. Players can choose a *pure* strategy, such as playing 'Rock' in the game Rock-Paper-Scissors,[1] or playing a *mixed* strategy consisting of randomly choosing among pure strategies 'Rock,' 'Paper,' and 'Scissor' with a fixed probability distribution. With these strategies available, Nash proved that every game has a Nash equilibrium, where no player can improve their expected payoff by changing their strategy. Thus a Nash equilibrium exhibits a type of stability that could be observed in game theoretic models of human behavior. The mixed strategy with equal probabilities in Rock-Paper-Scissors is a Nash equilibrium and people playing the game generally quickly learn to play something close to that strategy.

[1]Rock-Paper-Scissors is not a coordination game, it is a game of conflict with misaligned interests.

The Lewis signaling game (LSG) (Lewis David, 1969) is an early and influential coordination game having two players, the *sender* of signals, who observes a random state of nature, and a *receiver* of signals who must choose an action. For each state of nature there is exactly one action where both players get an equal positive payoff. All other actions result in zero payoffs. Players maximize their payoffs by finding an optimal solution or *signaling system*. In this case, the sender's strategy is a bijection from states of nature to signals and the receiver's strategy is the bijection from signals to actions that given the sender's strategy, always results in a positive payoff. Signaling systems are exactly the set of strict Nash equilibria (if one player deviates, payoffs will be strictly lower). 'Signaling games' in the literature can refer to a broad category including coordination games, where players' interests are aligned but also games of conflict, games with costly signaling, and more (S. Huttegger, Skyrms, Tarres, & Wagner, 2014), (De Silva & Sigmund, 2024).

Lewis assumed 2 equiprobable states and 2 signals resulting in 2 signaling systems among 8 possible pure strategy pairs which include several suboptimal signaling systems (suboptimal equilibria). A *pooled* equilibrium is a non-strict Nash equilibrium where the sender maps both states to one signal and the receiver maps both signals to a single action. With more than 2 states, *partial pooling* equilibria can occur when several states (and several signals) map to one signal (and one action) and other states (signals) map bijectively to other signals (actions). Except in Lewis's original game with exactly 2 equiprobable states (Argiento, Pemantle, Skyrms, & Volkov, 2009), methods for finding signaling systems often converge to these suboptimal solutions.

The general N state, N signal LSG ($N\times N$ LSG) has $N!$ signaling systems and N^{2N} pure strategy and infinitely many mixed strategy pairs. How can players find and agree upon one signaling system among the many possibilities? A program of repeated play lets players' strategies evolve over time according to a simple principle: strategies with higher payoffs are more likely to be played in the future. The many methods that embody this principle in some form are given in evolutionary game theory (EGT) (Newton, 2018), (Sandholm, 2020) and are based on ideas from biological evolution (Smith, 1974), (Smith, 1984) and stochastic learning theory (Bush & Mosteller, 1953), (Bush & Mosteller, 1955), (Camerer & Hua Ho, 1999). Links between discrete time and

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

continuous time dynamical systems allow for analysis of convergence and stability of equilibria (Hofbauer & Sigmund, 2003), (Benaïm, 1999), (Pemantle, 2007).

A simple stochastic learning rule is illustrated by an urn filled with colored balls, one color for each strategy. Let c_i^n be the number (count) of balls of color i after n games have been played. The urn is initially filled with c_i^0 colored balls. At game $n \geq 1$, a strategy is chosen by randomly drawing a ball from the urn and replacing the ball. If the drawn ball's strategy is successful, then another ball of that color is added to the urn. Adding a single ball corresponds to a payoff of 1. If the strategy is not successful, then the number of balls of each color in the urn is left unchanged. Clearly, a successful strategy ends up with a larger portion of balls in the urn and is more likely to be played in the future, matching the principle mentioned earlier. The probability of drawing color i ball at game $n+1$ is given below where the vector $\mathbf{c}^n = (c_1^n, \ldots, c_N^n)$ is the *state of learning* after n games.

$$\mathbb{P}(\text{ ball } i \,|\, \mathbf{c}^n) \;=\; \frac{c_i^n}{c_1^n + \ldots + c_N^n} \;=\; \frac{c_i^n}{\sum_{i'} c_{i'}^n}$$

This reinforcement learning (RL) rule from psychology (as opposed to machine learning (Sutton & Barto, 2018)) satisfies Thorndyke's Law of Effect (behavior with positive outcomes is reinforced and likely to be repeated) and the Law of Practice (learning is initially fast, then slows) and is generally consistent with animal and human behavior (Harley, 1981), (Roth & Erev, 1995). It is the basis for Model A (Barrett, Skyrms) and Model B (SRL) described below.

Convergence to suboptimal systems

It is known that coordination games with methods for repeated play all follow a similar general principle, often leading to similar qualitative behavior, (Börgers & Sarin, 1997), (Hopkins, 2002), (Hopkins & Posch, 2005). Hence, it is not surprising that convergence to suboptimal solutions is observed in many coordination games including LSG. An early example of this phenomena is shown in figure 1 of (Nowak & Krakauer, 1999). This was both preceded and followed by numerous other studies which sometimes found methods (such as forgetting, mutations, lateral inhibition through punishment, hybrid methods) to reduce and sometimes, with special parameter values, mostly eliminate the chance of convergence to suboptimal systems (J. A. Barrett, 2006), (Hofbauer & Huttegger, 2008), (J. Barrett & Zollman, 2009), (S. M. Huttegger, Skyrms, Smead, & Zollman, 2010), (Skyrms, 2010), (S. Huttegger et al., 2014), (De Silva & Sigmund, 2024). Other studies have analyzed why this happens (Pawlowitsch, 2008), (Spike, Stadler, Kirby, & Smith, 2017). One method, found independently by (Oliphant & Batali, 1997) and (Mühlenbernd, 2013) appears to always converge to optimal signaling systems by using agents who optimize, reason with Bayes and always update to deterministic strategies. This 'best response' method in game theory assumes agents contemplate their next move by estimating the other agents probabilities of play and then choose one best response that maximizes expected return.

For the SRL model, we assume humans make quick choices when learning a new task.[2] Instead of contemplating a menu of options (strategies) they first consider a single strategy and make a binary yes or no decision. If yes, they use it. If no, they consider another strategy and make another yes or no decision.[3] This continues until a 'yes' occurs and they use that strategy or they run out of possible strategies and stop and do nothing.[4] Learning occurs when a strategy is successful. Consequently that strategy is reinforced, becoming more likely to be used (have a 'yes' decision) in the future. As learning occurs, some of these yes / no decisions become almost instantaneous and so this sequential decision process becomes immediate as happens after learning.

Reinforcement learning models for LSGs

The basic RL rule is the foundation for both our SRL rule (Model B) and the RL learning rule (Model A) given in (J. A. Barrett, 2006), (Skyrms, 2006), (Argiento et al., 2009) and (Skyrms, 2010). Both models have low cognitive requirements and the same number of parameters.

Model A (Barrett, Skyrms) and Model B (SRL)

Model A In Model A the sender has one urn for each state of nature i and each urn is filled with balls representing signals j. If the sender observes state i^*, they pull a ball j from urn i^*, send signal j corresponding to the ball and then return the ball to the urn. Similarly, the receiver has one urn for each signal j and each urn is filled with balls representing actions i. The receiver observes signal j and pulls a ball of action type i from urn j, performs action i and returns the ball to the urn. If the action matches the state ($i = i^*$) then the communication is successful and the signal and action choices are reinforced by adding a j ball to the sender's i^* urn and adding an $i = i^*$ ball to the receiver's urn j.

Model A has relatively low cognitive demands where only memory and relative size of cumulative payoffs are used. However, this model suffers from not using the entire state of knowledge when making a decision – once the state of nature is observed, the sender only consults one urn and states of learning in other urns is ignored. For example, if state of nature i^* is observed by the sender, and an urn $i \neq i^*$ has very high probability of drawing signal j, then that knowledge should make it less likely to use signal j with state i^*. There is nothing to prevent both urns from using signal j with high probability leading to suboptimal partial pooling.

To compensate for this issue, some models use a *lateral inhibition* mechanism. This includes punishment (removing

[2] In the words of (Skyrms, 2010), page 90, "[RL agents] do not have to know that they *are* in a game."

[3] Cf. 'statisficing' and (Gigerenzer & Selten, 2002).

[4] Strictly speaking, 'doing nothing' isn't playing an LSG and is clearly suboptimal. However, agents are not assumed to be optimizers and instead are responding to reinforcements, producing strategies that in the limit learn to play an LSG with perfect success.

balls) for communication failure or removing competing balls when reinforcing, (Catteeuw & Manderick, 2014), (J. A. Barrett, Cochran, Huttegger, & Fujiwara, 2017), (J. A. Barrett & Gabriel, 2024). Successfully implementing punishment can depend on parameters in the LSG model such as number of states and signals N or state probabilities. A more subtle form of lateral inhibition is enforcing pure strategies such as win-stay/lose-randomize or variants of a 'best response' strategy.

Partial pooling equilibria have lower expected payoffs than signaling systems (which are Pareto optimal). In Model A, a low probability state of nature is much more likely to be part of a partially pooled equilibrium, since the difference in expected payoff between a signaling system and the partially pooled system can be small due to the rare chance of needing a separate signal for the low probability state. Adding noise to Model A such as forgetting in learning models and mutations in evolutionary models, improves the probability of converging to a signaling system, (J. A. Barrett, 2006), (J. Barrett & Zollman, 2009), (Pawlowitsch, 2007).

Model B In Model B, we assume both the sender and receiver have a habit of considering various behaviors in a fixed order. The sender has a list of signals labeled by the index $j = 1,\ldots,N$, perhaps ordered from simpler or easier to produce to more complicated or by the order in which they were created or learned. The receiver has a list of actions labeled by the index $i = 1,\ldots,N$, also perhaps ordered from simple to complicated or by the order in which they were learned or discovered. For convenience, we relabel the states of nature so that state i and action i result in a positive payoff.

The sender has one urn for each signal j and each urn is filled with balls of type $i = 1,\ldots,N$ representing states of nature. The sender first observes the state of nature i^*. The sender pulls balls from each sender urn in order $j = 1,2,\ldots,$ and stops either when a drawn ball i from urn j^* matches the observed state, $i = i^*$, or when draws from all urns don't match i^*. In the former case, the sender uses signal j^*, and in the latter case, the game stops without signaling success and without a payoff. In either case, all balls are returned to their respective urns. The sender pulling a ball from urn j represents the sender contemplating how often in the past the sender has used signal j with state i^* compared to other states $i \neq i^*$.

The receiver has one urn for each action i and each urn is filled with balls of type $j = 1,\ldots,N$ representing signals. If the receiver observes signal j^*, then the receiver pulls balls from each receiver urn in order $i = 1,2,\ldots,$ and stops either when a drawn ball j from urn i matches the observed signal, $j = j^*$ or when draws from all urns don't match j^*. In the former case, the receiver uses action i, and in the latter case, the game stops without signaling success and without a payoff. In either case, all balls are returned to their respective urns. The receiver pulling a ball from urn i represents the receiver contemplating how often in the past the receiver has used action i with signal j^* compared to other signals $j \neq j^*$.

If $i \neq i^*$, then communication is unsuccessful and the game ends with no changes to urns. If action i matches the state of nature $i = i^*$, then we have signaling success. Both sender and receiver get a reinforcement (payoff) where sender's urn j^* gets an i^* ball added and receiver's urn i^* gets a j^* ball added. This results in higher probability for sender to use signal j^* with state of nature i^* and higher probability for the receiver to use action i^* when observing signal j^*.

One Signal: Model A pools, Model B doesn't

Suppose there is one signal. States of nature are provided by the environment, but signals must be invented by players and in this case there is only one, called 'signal 1.'

In Model A, every sender urn contains only signal 1 type balls. Hence, the sender maps every state to signal 1, and all convergent systems are pooled. If there is one action then the receiver always uses action 1. If there are multiple actions, then a Nash equilibrium has the receiver mapping the signal to the action corresponding to highest probability state(s). In this very simple case, a signal can never correspond to a single state of nature. And though this is a plausible consequence of having only one signal, this learned behavior may make the creation and use of a new signal more difficult.

In Model B the sender has only urn 1 (for signal 1) filled with balls i representing states of nature and the receiver has one urn for each action i, but each urn only has balls of type signal 1. The sender upon observing state i^* will send signal 1 if sender draws ball i^* and otherwise will do nothing. The receiver upon observing signal 1, always uses action 1 since that urn only contains signal 1 balls. Thus only when state 1 occurs and sender draws ball 1 is there successful communication and balls are added to the appropriate urns for reinforcement.

Over time the sender's urn 1 fills with balls of type $i = 1$ and the sender learns to use signal 1 with state 1 and do nothing with other states. The receiver always takes action 1 when observing signal 1 resulting in successful communication. Thus Model B converges to an *efficient partial signaling system* where the sender uses signal 1 for state 1 and the receiver always uses action 1 and for other states of nature, no signal is sent and no action is needlessly performed. In all other states of nature, the sender is silent, which makes it easy to learn a new signal when one becomes available.

Lateral inhibition We introduce notation for Model B. Let c_{ij} be the count of type i balls in the sender's jth urn and let d_{ij} be the count of type j balls in the receiver's ith urn. We define matrices $\mathbf{c}^n = [c_{ij}^n]$ and $\mathbf{d}^n = [d_{ij}^n]$ to be the *states of learning* after n games for sender and receiver. Let x_{ij} be the probability of randomly drawing a type i ball from sender's jth urn and let y_{ij} be the probability of randomly drawing a type j ball from receiver's ith urn:

$$x_{ij} = \mathbb{P}(\text{ state } i \mid \text{ signal urn } j) = \frac{c_{ij}}{\sum_{i'} c_{i'j}},$$

$$y_{ij} = \mathbb{P}(\text{ signal } j \mid \text{ action urn } i) = \frac{d_{ij}}{\sum_{j'} d_{ij'}}.$$

Suppose sender observes state i. The probability of sending signal 1 is given by x_{i1}. The probability of sending signal 2 requires first a decision to not use signal 1, which has probability $1-x_{i1}$. Thus the probability of sending signal 2 is given by $x_{i2}(1-x_{i1})$ and the probability of sending signal 3 is $x_{i3}(1-x_{i2})(1-x_{i1})$, etc. Similar formulas hold for actions with y_{ij} terms. If there is a high probability of using signal 1 with state i, then subsequent signals have a low probability of being used and being reinforced. Thus if the probability of using signal 1 with state i converges to 1 then the probability of using signal 1 with other states $i' > i$ converges to 0 (due to not being reinforced). This lateral inhibition indicates why Model B converges to an optimal signaling system.

Theoretical and Numerical Results for Model B

In the appendix we prove Model B converges almost surely (i.e., with probability 1) to an optimal signaling system for any $N \times N$ LSG with arbitrary positive state probabilities and any positive initial states of learning. The proof uses techniques similar to those in (Beggs, 2005) and starts by deriving inequalities before using Doob's martingale convergence theorem. Other proof techniques for RL involve compact operators (Norman, 1968), (Norman, 1972), and stochastic approximation (Benaïm, 1999), (Hopkins & Posch, 2005), (Pemantle, 2007). We use these methods to prove global convergence results for some RL models having direct application in linguistics; see (Wechsler, Shearer, & Erk, 2025), (Boguraev, Katrin, Mahowald, Shearer, & Wechsler, 2025) for models and some proofs.

Numerical methods reveal rates of convergence and perhaps uncover new, interesting phenomena. To that end, we use Monte Carlo simulation with both typical and unusual and randomly generated initial conditions and state probabilities with $N = 2, 3, 4, 5, 7$. In all cases there is a clear progression showing all scenarios are converging to an optimal signaling system. We found that different scenarios often converge to different signaling systems. There is a tendency for the first signal to be used with the highest probability state of nature and this is more pronounced when action decisions are in order of decreasing state probabilities. If the order of signals is from simple to more complicated (longer), then this corresponds to a tendency towards efficient signaling in the information theoretic sense.

Since there are $N!$ possible optimal signaling systems, and many scenarios converge to different systems, it is not possible to describe outcomes of a simulation in terms of signaling systems. Instead we report the number of successful communications resulting in payoffs to the sender and receiver. As a scenario converges to an optimal signaling system, it's probability of successful communication converges to 100%. Non-optimal signaling always leads to a positive probability of failure.

For example, with N= 5, all initial conditions set equal to 1, and a randomly chosen set of state probabilities (0.012888, 0.31682, 0.26332, 0.29888, 0.10809) we have, after 10^7 utterances, 99.1% of 10,000 scenarios have a greater than 99.9% chance of successful communication. This high success rate is shared by even the low probability states, which have a mean success rate 99.8% (a low probability state will converge more slowly due to fewer chances to learn). The percentage of scenarios with signal 1 being chosen for state i, after 10^7 utterances is 47.5% for the highest probability state 0.31682 and remains above 20% for the other two highest probabilities.

Model C and D: two agents alternate between speaking and hearing

Several generalizations of LSG assume agents are both senders (speakers) and receivers (hearers). All agents have two separate states of learning, one for production (speaking or sending) and one for interpreting (receiving or hearing), (Nowak & Krakauer, 1999), (S. M. Huttegger, 2007). Since a single meaning, expressed and interpreted, is a common link in speaking and hearing, we propose Model C, where each agent has their own single state of learning that looks like the sender's state of learning in Model B. However, in Model C, it is used for both production (speaking) and interpretation (hearing). Model C borrows from both the LSG and language acquisition frameworks. In the latter, the speaker and hearer observe the same scene and are aware of the context. Language is acquired by both habit from usage and inferring meaning when hearing, with minimal direct positive or negative reinforcement. So we start by assuming the reinforcement in Model C happens when speaking and inferring meaning and not when speaker and hearer are shown to agree on meaning via *deus ex machina* as in LSG.

In Model C, there are two agents who alternate between speaking and hearing. They both observe the same scene and know the context. Agent 1 has state of learning **c** and agent 2 has state of learning **d** (not to be confused with the receiver's state of learning **d** in Model B). These states of learning can be interpreted as the urn model used by the sender in Model B. Instead of the speaker observing a state of nature, the speaker chooses a meaning using their probability distribution, p_i = probability of meaning (or state) i. The other agent, who is the hearer at the moment, has a similar probability distribution q. The speaker then uses their state of learning **c** exactly as in Model B to choose a signal j (or to remain silent, and then the game is over). If the speaker chose signal j for meaning i, then regardless of what the hearer does, this use is a habit that is reinforced by adding an i ball to the jth urn (increasing c_{ij} by 1).

The hearer now makes a guess as to the meaning of the signal j. The hearer has their probability distribution q regarding what is a meaning they would probably choose. If they had no previous knowledge they could make a guess using q. But they have a current state of learning **d**. Bayes Theorem allows them to combine q with the state of learning **d** to get a posterior distribution and randomly guess a meaning i'. Because of this inferred experience, the hearer adds an i' color ball to their jth urn (increasing $d_{i'j}$ by 1).

One can argue that Bayesian analysis is exceeding the standard notions of 'bounded rationality.' On the other hand, this mathematical frame may well capture an intuitive thinking mechanism in humans. Bayesian methods are used in language acquisition and other language models, (Xu & Tenenbaum, 2007), (Kemp, Perfors, & Tenenbaum, 2007) in this way. A second issue is that both agents use the same, fixed ordering of signal urns. We relax this in Model D by assuming each agent has their own probability distribution for choosing a different order for signal urns each time they speak. A downside of this is the now more complicated Bayesian analysis for the hearer. But again, we assume this is an approximation for how the hearer incorporates past experience with their priors. Lastly, these models never allow for occasional reinforcement with clear understanding between agents. This is for future work.

We use Monte Carlo simulation with 10,000 scenarios and up to 10^6 utterances with 3 meanings (states). Typically we see convergence to a signaling system in Model C with close, but different p and q. For Model D, we also can get convergence to a signaling system with close and different p and q. There are lots of parameters (initial conditions for $\mathbf{c}$, $\mathbf{d}$, four probability distributions) which can affect the outcome. The main point is that with NO reinforcement due to correct interpretation by the hearer, we still can often get convergence to an optimal signaling system by bounded rational agents with close meaning distributions and using sequential reinforcement learning with Bayesian reasoning when interpreting.

Conclusion

We present a new reinforcement learning model (Model B) and prove it converges to an optimal signaling system for any $N\times N$ Lewis signaling game with arbitrary positive state probabilities and positive initial conditions. Model B is based on the idea that humans presented with a new task consider strategies one at time and make a sequence of binary decisions regarding whether or not to use a strategy. If signals in Model B are contemplated in order from simple (shorter) to more complicated (longer), then one gets a plausible mechanism for efficient languages in the information theoretic sense. Lastly instead of two separate states of learning for speaking and hearing (production and interpretation), we use a single state of learning for production and Bayesian reasoning with a context dependent probability to interpret. This can help explain why comprehension is easier than production and why some concepts are not learned (future work). These models can achieve optimal signaling with minimal direct reinforcement as in natural language acquisition.

Appendix: LSG theorems and proofs

We say Model B *converges* if all $x_{ij},\ y_{ij}$ converge to either 0 or 1. Theorem 1 shows that if Model B converges, it converges to an optimal signaling system. In Theorem 2, we show all convergent 2 state, 2 signal LSGs converge to a signaling system. The same proof, but with many more terms, works in the general $N\times N$ case.

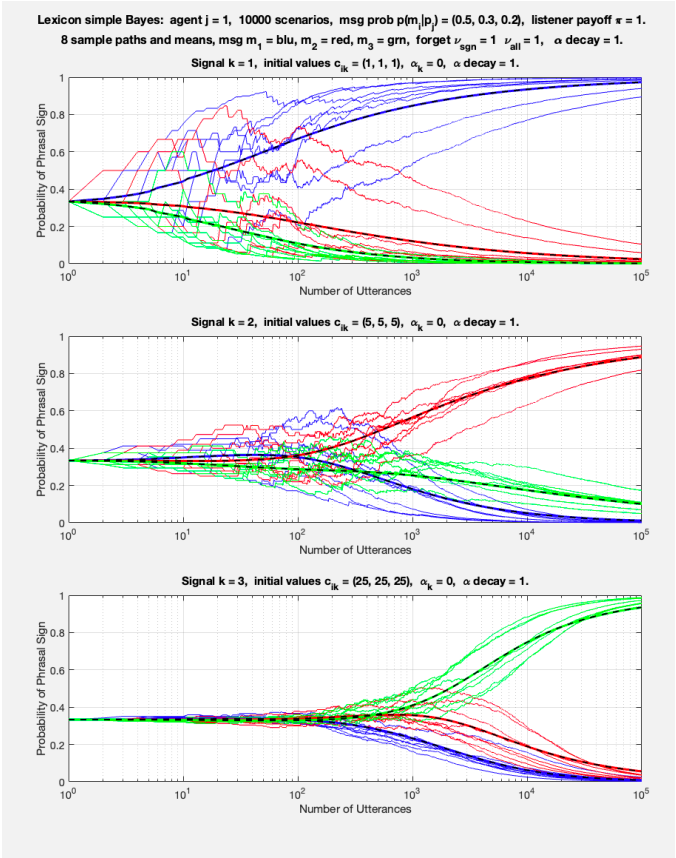

Figure 1: Model C: Agent 1 (plot above) and agent 2 converge to the same optimal system. They have different 'meaning' probabilities, .50, .30, .20 vs .55, .30, .15, but same increasing initial conditions that allow for staggered convergence.

Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space associated with the Markov process $\mathbf{c}$, $\mathbf{d}$ and let $\mathcal{F}_n$ be the filtration associated with $\mathbf{c}^n$, $\mathbf{d}^n$. We suppress dependence on n when clear.

$$U_{ij} = \prod_{j'=1}^{j-1}(1-x_{ij'}), \qquad V_{ij} = \prod_{i'=1}^{i-1}(1-y_{i'j})$$

$$\begin{aligned} p_i &= \text{probability of state of nature } i \\ p_{ij} &= \text{probability of } i,j \text{ payoff in game } n+1 \\ &= \mathbb{P}\left(c_{ij}^{n+1} = c_{ij}^n + 1 \,\middle|\, \mathbf{c}^n\right) = p_i x_{ij} y_{ij} U_{ij} V_{ij} \end{aligned}$$

A few things to note:
(1) If $x_{i^*j} \to 1$, then $x_{ij} \to 0$ for all $i \neq i^*$ since $\sum_{i'} x_{i'j} = 1$.
(2) If $x_{ij} = c_{ij}/\sum_{i'} c_{i'j} \to 1$ then we must have $c_{ij} \to \infty$ since otherwise $x_{ij} \not\to 1$. Since $c_{ij}^n = d_{ij}^n + c_{ij}^0 - d_{ij}^0$ we also have $d_{ij} \to \infty$. Similarly for y_{ij}, c_{ij} and d_{ij}.
(3) If all $x_{ij},\ y_{ij}$ converge to 0 or 1, then all U_{ij} and V_{ij} also converge to either 0 or 1.

We take expected values conditioned on state of learning

$\mathbf{c}^n$, $\mathbf{d}^n$ and suppress dependence on n on the right hand side.

$$\begin{aligned}
&\mathbb{E}\left[\frac{c_{ij}^{n+1}}{c_{i^*j^*}^{n+1}}\,\middle|\,\mathbf{c}^n\right] \\
&= \frac{c_{ij}+1}{c_{i^*j^*}}p_{ij}+\frac{c_{ij}}{c_{i^*j^*}+1}p_{i^*j^*}+\frac{c_{ij}}{c_{i^*j^*}}(1-p_{ij}-p_{i^*j^*}) \\
&= \frac{c_{ij}}{c_{i^*j^*}}\left\{1+\frac{1}{c_{ij}}p_{ij}-\frac{1}{c_{i^*j^*}}\left[1-\frac{1}{c_{i^*j^*}+1}\right]p_{i^*j^*}\right\} \\
&= \frac{c_{ij}}{c_{i^*j^*}}\left\{1+\frac{x_{ij}}{c_{ij}}\frac{p_{ij}}{x_{ij}}-\frac{x_{i^*j^*}}{c_{i^*j^*}}\left[1-\frac{1}{c_{i^*j^*}+1}\right]\frac{p_{i^*j^*}}{x_{i^*j^*}}\right\}
\end{aligned}$$

Setting $j=j^*$ in the above, we get

$$= \frac{c_{ij^*}}{c_{i^*j^*}}\left\{1+\frac{1}{\sum_{i'}c_{i'j^*}}\left(\frac{p_{ij^*}}{x_{ij^*}}-\left[1-\frac{1}{c_{i^*j^*}+1}\right]\frac{p_{i^*j^*}}{x_{i^*j^*}}\right)\right\} \quad (1)$$

Setting $i=i^*$, replacing c with d and x with y, we get

$$= \frac{d_{i^*j}}{d_{i^*j^*}}\left\{1+\frac{1}{\sum_{j'}d_{i^*j'}}\left(\frac{p_{i^*j}}{y_{i^*j}}-\left[1-\frac{1}{d_{i^*j^*}+1}\right]\frac{p_{i^*j^*}}{y_{i^*j^*}}\right)\right\} \quad (2)$$

Theorem 1 *Suppose on set $A\subset\Omega$ with $\mathbb{P}(A)>0$, all scenarios in Model B, N×N LSG, converge to the same language.*

a) If $x_{ij}\to 1$, $U_{ij}\to 1$ then $y_{ij}\to 1$, $V_{ij}\to 1$.

b) If $y_{ij}\to 1$, $V_{ij}\to 1$ then $x_{ij}\to 1$, $U_{ij}\to 1$.

Proof: We prove a) by induction on $j^*=1,\ldots,N$, and b) is proved similarly. We assume a) is true for all i and any $j<j^*$, then show it is true for j^*. Let $\mathbb{E}_A[\cdot]=\mathbb{E}[\mathbf{1}_A\cdot]$ denote expectation on set A and $\mathbf{1}_A$ is the indicator function of A.

Step 1 Assume $x_{i^*j^*}\to 1$ and $U_{i^*j^*}\to 1$.

Step 2 We show $y_{ij^*}\to 0$ for all $i<i^*$ (hence $V_{i^*j^*}\to 1$).

There is nothing to prove if $i^*=1$, so we assume $i^*>1$ and use induction on $i=1,\ldots,i^*-1$ with induction hypothesis at ith step being $y_{i'j^*}\to 0$ for all $i'<i$.

If $U_{ij^*}\to 0$ then $x_{ij}\to 1$ for some $j<j^*$ with $U_{ij}\to 1$ (or use $U_{i1}=1$ if $j=1$). Assuming a) holds for $j<j^*$ we have $y_{ij}\to 1$, hence $y_{ij^*}\to 0$ since $\sum_{j'}y_{ij'}=1$.

If $U_{ij^*}\not\to 0$ then $U_{ij^*}\to 1$. If $i=1$ then $V_{ij^*}=1$. If $i>1$, then by the Step 2 induction hypothesis, $y_{i'j^*}\to 0$ for all $i'<i$, hence $V_{ij^*}\to 1$. We want $y_{ij^*}\to 0$ and instead assume $y_{ij^*}\to 1$, hence $V_{i^*j^*}\to 0$, and derive a contradiction. Using equation (1) from above with $y_{ij^*}U_{ij^*}V_{ij^*}\to 1$ and $V_{i^*j^*}\to 0$ we have

$$\begin{aligned}
&\mathbb{E}_A\left[\frac{c_{ij^*}^{n+1}}{c_{i^*j^*}^{n+1}}\,\middle|\,\mathbf{c}^n\right] \\
&\geq \mathbf{1}_A\frac{c_{ij^*}}{c_{i^*j^*}}\left\{1+\frac{1}{\sum_{i'}c_{i'j^*}}\left(p_iy_{ij^*}U_{ij^*}V_{ij^*}-V_{i^*j^*}\right)\right\}
\end{aligned}$$

For all large n the term in braces is greater than 1 on set $B\subset A$ with positive measure. Taking expectations we see $\mathbb{E}_B[c_{ij^*}/c_{i^*j^*}]\not\to 0$, contradicting $x_{i^*j^*}\to 1$ on A.

Step 3: We show $y_{i^*j}\to 0$ for all $j\neq j^*$ (hence $y_{i^*j^*}\to 1$).

We have $U_{i^*j^*}\to 1$ implies $x_{i^*j}\to 0$ for $j<j^*$ and $x_{i^*j^*}\to 1$ implies $U_{i^*j}\to 0$ for $j>j^*$. From step 2 we have $V_{i^*j^*}\to 1$. Using equation (2) with $x_{i^*j}U_{i^*j}\to 0$ for all $j\neq j^*$, $x_{i^*j^*}U_{i^*j^*}V_{i^*j^*}\to 1$ and $D_1=1-1/(d_{i^*j^*}+1)$, we get

$$\begin{aligned}
&\mathbb{E}_A\left[\frac{d_{i^*j}^{n+1}}{d_{i^*j^*}^{n+1}}\,\middle|\,\mathbf{d}^n\right] \\
&\leq \mathbf{1}_A\frac{d_{i^*j}}{d_{i^*j^*}}\left\{1+\frac{p_{i^*}}{\sum_{j'}d_{i^*j'}}\left(x_{i^*j}U_{i^*j}-D_1x_{i^*j^*}U_{i^*j^*}V_{i^*j^*}\right)\right\}
\end{aligned}$$

Since $x_{i^*j^*}\to 1$ we have $d_{i^*j^*}\to\infty$ and $D_1\to 1$. The term in parentheses is less than $-1/2$ for large n on $B\subset A$ with $\mathbb{P}(A)-\mathbb{P}(B)=\varepsilon$, ε arbitrarily small. Using $\sum_{j'}d_{i^*j'}^n\leq n+\sum_{j'}d_{i^*j'}^0$ and taking expectations we get

$$\begin{aligned}
\mathbb{E}_B\left[\frac{d_{i^*j}^{n+1}}{d_{i^*j^*}^{n+1}}\right] &= \mathbb{E}_B\left[\mathbb{E}_B\left[\frac{d_{i^*j}^{n+1}}{d_{i^*j^*}^{n+1}}\,\middle|\,\mathbf{d}^n\right]\right] \\
&\leq \mathbb{E}_B\left[\frac{d_{i^*j}^{n}}{d_{i^*j^*}^{n}}\right]\left\{1-\frac{p_{i^*}}{2\left(n+\sum_{j'}d_{i^*j'}^0\right)}\right\}
\end{aligned}$$

Thus $\mathbb{E}_B[d_{i^*j}^n/d_{i^*j^*}^n]\to 0$, and using Doob's martingale convergence theorem we have $d_{i^*j}^n/d_{i^*j^*}^n\to 0$ on $B\subset A$ almost surely and $\varepsilon>0$ can be arbitrarily small. Hence $y_{i^*j}\to 0$ almost surely for all $j\neq j^*$ on A. ■

Theorem 2 *If the hypotheses in Theorem 1 hold with $N=2$, then Model B converges to an optimal signaling system on A.*

Proof: Suppose not. Given theorem 1, the only possibility is $x_{11},\ y_{11}\to 1$, $x_{12},\ y_{21}\to 1$ and other terms converge to 0. Then $c_{12}/c_{22}\to\infty$. Let $C>0$, $B\subset A$ be generic constants and subsets. $\mathbb{P}(A)-\mathbb{P}(B)$ can be made arbitrarily small.

Step 1: For any $\varepsilon>0$ there is n_0 such that $p_1+\varepsilon\geq c_{11}/n\geq p_1-\varepsilon$, all $n\geq n_0$ on $B\subset A$, $\mathbb{P}(A)-\mathbb{P}(B)<\varepsilon$.

For all large n, $x_{11},y_{11}>1-\varepsilon/4$ on B. If $c_{11}^{n+1}=c_{11}^n+\eta^{n+1}$, use Bernoullis $\xi_1^n\geq\eta^n\geq\xi_2^n$ with ξ_i means p_1, $p_1-\varepsilon/2$.

Step 2: $c_{21}/c_{11}\leq C/n^{\delta_4}$ for large n on B', $\delta_k=1-k\varepsilon/p_1$.

$$\mathbb{E}_B\left[\frac{c_{21}^{n+1}}{c_{11}^{n+1}}\,\middle|\,\mathbf{c}^n\right]\leq\mathbf{1}_B\frac{c_{21}}{c_{11}}\left\{1-\frac{\delta_3}{n}\right\}\leq\mathbf{1}_B\frac{c_{21}}{c_{11}}e^{-\delta_3/n}$$

$$\mathbb{E}_B\left[\frac{1}{(N+1)^{\delta_4}}\frac{c_{21}^{N+1}}{c_{11}^{N+1}}\right]\leq\prod_{k=n}^{N}e^{-\varepsilon/2k}\mathbb{E}_B\left[\frac{c_{21}^n}{c_{11}^n}\right]\to 0 \text{ as } N\to\infty$$

Doob's martingale convergence theorem implies $c_{21}^n/[n^{\delta_4}c_{11}^n]\to 0$ almost surely on B, and $c_{21}^n/c_{11}^n<Cn^{\delta_4}$ on $B'\subset B$, all large n.

Step 3: $x_{11}=1/(1+c_{21}/c_{11})\geq 1-C/n^{-\delta_4}$ on B', $n>n_0'$.

Step 4: $c_{12}/c_{22}\not\to\infty$ on $B'\subset A$, contradiction.

$$\frac{p_{12}}{c_{12}}=p_1\frac{y_{12}}{c_{12}}x_{12}(1-x_{11})\leq\frac{C}{c_{11}/n}\frac{C}{n^{\delta}},\quad\delta=1+\delta_4>1$$

$$\mathbb{E}_{B'}\left[\frac{c_{12}^{n+1}}{c_{22}^{n+1}}\,\middle|\,\mathbf{c}^n\right]\leq\mathbf{1}_{B'}\frac{c_{12}}{c_{22}}\left\{1+\frac{p_{12}}{c_{12}}\right\}\leq\mathbf{1}_{B'}\frac{c_{12}}{c_{22}}\left\{1+\frac{C'}{n^{\delta}}\right\}$$

$$\mathbb{E}_{B'}\left[\frac{c_{12}^N}{c_{22}^N}\right]\leq\prod_{n=n_0}^{\infty}e^{C'/n^{\delta}}\mathbb{E}_{B'}\left[\frac{c_{12}^{n_0}}{c_{22}^{n_0}}\right]=C''<\infty\quad\blacksquare$$

References

Argiento, R., Pemantle, R., Skyrms, B., & Volkov, S. (2009). Learning to signal: Analysis of a micro-level reinforcement model. *Stochastic processes and their applications*, *119*(2), 373–390.

Barrett, J., & Zollman, K. J. (2009). The role of forgetting in the evolution and learning of language. *Journal of Experimental & Theoretical Artificial Intelligence*, *21*(4), 293–309.

Barrett, J. A. (2006). Numerical simulations of the lewis signaling game: Learning strategies, pooling equilibria, and the evolution of grammar.

Barrett, J. A., Cochran, C. T., Huttegger, S., & Fujiwara, N. (2017). Hybrid learning in signalling games. *Journal of Experimental & Theoretical Artificial Intelligence*, *29*(5), 1119–1127.

Barrett, J. A., & Gabriel, N. (2024). Reinforcement with iterative punishment. *Journal of Experimental & Theoretical Artificial Intelligence*, *36*(7), 1361–1383.

Beggs, A. W. (2005). On the convergence of reinforcement learning. *Journal of economic theory*, *122*(1), 1–36.

Benaïm, M. (1999). Dynamics of stochastic approximation algorithms. *Séminaire de probabilités (Strasbourg)*, *33*, 1-68.

Boguraev, S., Katrin, E., Mahowald, K., Shearer, J., & Wechsler, S. (2025). Reinforcement learning produces efficient case-marking systems. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 47).

Börgers, T., & Sarin, R. (1997). Learning through reinforcement and replicator dynamics. *Journal of economic theory*, *77*(1), 1–14.

Bush, R. R., & Mosteller, F. (1953). A stochastic model with applications to learning. *The Annals of Mathematical Statistics*, 559–585.

Bush, R. R., & Mosteller, F. (1955). *Stochastic models for learning.* John Wiley & Sons, Inc.

Camerer, C., & Hua Ho, T. (1999). Experience-weighted attraction learning in normal form games. *Econometrica*, *67*(4), 827–874.

Catteeuw, D., & Manderick, B. (2014). The limits and robustness of reinforcement learning in lewis signalling games. *Connection Science*, *26*(2), 161–177.

De Silva, H., & Sigmund, K. (2024). Dynamics of signaling games. *SIAM Review*, *66*(2), 368–387.

Gigerenzer, G., & Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. MIT press.

Harley, C. B. (1981). Learning the evolutionarily stable strategy. *Journal of theoretical biology*, *89*(4), 611–633.

Hofbauer, J., & Huttegger, S. M. (2008). Feasibility of communication in binary signaling games. *Journal of theoretical biology*, *254*(4), 843–849.

Hofbauer, J., & Sigmund, K. (2003). Evolutionary game dynamics. *Bulletin of the American mathematical society*, *40*(4), 479–519.

Hopkins, E. (2002). Two competing models of how people learn in games. *Econometrica*, *70*(6), 2141–2166.

Hopkins, E., & Posch, M. (2005). Attainability of boundary points under reinforcement learning. *Games and Economic Behavior*, *53*(1), 110–125.

Huttegger, S., Skyrms, B., Tarres, P., & Wagner, E. (2014). Some dynamics of signaling games. *Proceedings of the National Academy of Sciences*, *111*(supplement_3), 10873–10880.

Huttegger, S. M. (2007). Evolution and the explanation of meaning. *Philosophy of science*, *74*(1), 1–27.

Huttegger, S. M., Skyrms, B., Smead, R., & Zollman, K. J. (2010). Evolutionary dynamics of lewis signaling games: signaling systems vs. partial pooling. *Synthese*, *172*, 177–191.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical bayesian models. *Developmental science*, *10*(3), 307–321.

Lewis David, K. (1969). Convention: a philosophical study. *Cambridge MA: Harvard*.

Mühlenbernd, R. (2013). *Signals and the structure of societies*. Unpublished doctoral dissertation, Universität Tübingen.

Newton, J. (2018). Evolutionary game theory: A renaissance. *Games*, *9*(2), 31.

Norman, M. F. (1968). Some convergence theorems for stochastic learning models with distance diminishing operators. *Journal of Mathematical Psychology*, *5*(1), 61–101.

Norman, M. F. (1972). *Markov processes and learning models* (Vol. 84). Academic Press New York.

Nowak, M. A., & Krakauer, D. C. (1999). The evolution of language. *Proceedings of the National Academy of Sciences*, *96*(14), 8028–8033.

Oliphant, M., & Batali, J. (1997). Learning and the emergence of coordinated communication. *Center for research on language newsletter*, *11*(1), 1–46.

Osborne, M. J. (2004). An introduction to game theory. *Oxford University Press google schola*, *2*, 672–713.

Pawlowitsch, C. (2007). Finite populations choose an optimal language. *Journal of Theoretical Biology*, *249*(3), 606–616.

Pawlowitsch, C. (2008). Why evolution does not always lead to an optimal signaling system. *Games and Economic Behavior*, *63*(1), 203–226.

Pemantle, R. (2007). A survey of random processes with reinforcement. *Probability surveys*, *4*, 1–79.

Roth, A. E., & Erev, I. (1995). Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and economic behavior*, *8*(1), 164–212.

Sandholm, W. H. (2020). Evolutionary game theory. *Complex social and behavioral systems: game theory and agent-based models*, 573–608.

Skyrms, B. (2006). Signals presidential address. In *Meeting of the philosophy of science association.*

Skyrms, B. (2010). *Signals: Evolution, learning, and information*. Oxford University Press.

Smith, J. M. (1974). The theory of games and the evolution of animal conflicts. *Journal of theoretical biology*, *47*(1), 209–221.

Smith, J. M. (1984). Game theory and the evolution of behaviour. *Behavioral and Brain Sciences*, *7*(1), 95–101.

Spike, M., Stadler, K., Kirby, S., & Smith, K. (2017). Minimal requirements for the emergence of learned signaling. *Cognitive science*, *41*(3), 623–658.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Von Neumann, J., & Morgenstern, O. (2007). Theory of games and economic behavior: 60th anniversary commemorative edition. In *Theory of games and economic behavior*. Princeton university press.

Wechsler, S., Shearer, J. W., & Erk, K. (2025). The emergence of grammar through reinforcement learning. *arXiv preprint arXiv:2503.01635*.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological review*, *114*(2), 245.