Learning to Signal in a Dynamic World*

J. McKenzie Alexander
Department of Philosophy, Logic and Scientific Method
London School of Economics and Political Science

February 21, 2012

Abstract

Sender-receiver games, first introduced by David Lewis in *Convention*, have received increased attention in recent years as a formal model for the emergence of communication. Skyrms (2010) showed that simple models of reinforcement learning often succeed in forming efficient, albeit not necessarily minimal, signalling systems for a large family of games. Later, Alexander et al. (2011) showed that reinforcement learning, combined with forgetting, frequently produced both efficient and minimal signalling systems. In this paper I define a *dynamic* sender-receiver game in which the state-action pairs are not held constant over time, and show that neither of these two models of learning learn to signal in this environment. However, a model of reinforcement learning with discounting of the past does learn to signal; it also gives rise to the phenomenon of linguistic drift.

1. Introduction.

The shadow of David Lewis's book *Convention* looms large over both game theory and philosophy. In addition to offering the first explicit definition of common knowledge, Lewis sought to refute Quine's claim that meaning could not arise solely via convention. In doing so, he provided the first formulation of what has come to be known as the family of sender-receiver games.

The simplest form of the sender-receiver game, shown in figure 1, consists of two players: the Sender and the Receiver, with a minor role played by Nature in determining the outcome of a chance event. Nature selects a state of the world, which the Sender observes without error. The Sender then transmits an arbitrary

*Do not cite without permission of the author.

signal to the Receiver, who observes the signal without error. The Receiver then selects an action to perform. If the action performed is the "correct" action for the state of the world, both the Sender and the Receiver get a positive payoff; otherwise, each receives nothing. Numerous interactions in the animal kingdom exist which have a structure similar to this game (see ch. 2 of Skyrms, 2010).

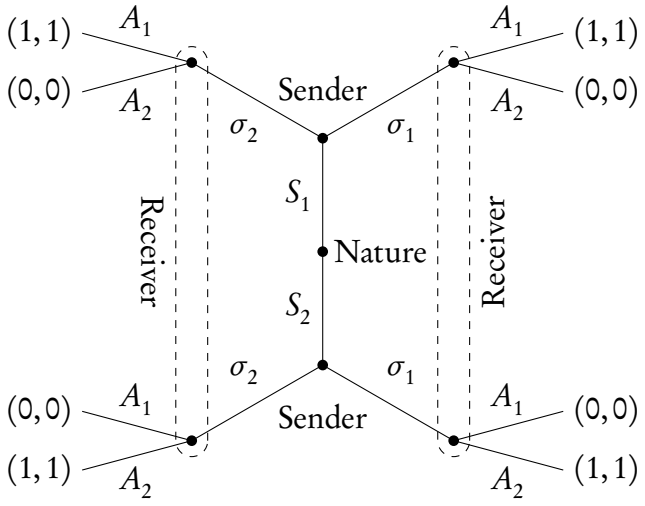


Figure 1: A simple sender-receiver game, in extensive form.

Despite Lewis's brilliance, one aspect which he did not address in *Convention*, much less solve, was just how a group of rational agents facing a collective action problem described by a sender-receiver game might successfully bootstrap their way to a solution. In short, Lewis did not consider the *dynamics* underlying equilibrium selection for sender-receiver games. Thus the question arises: can rational agents to coordinate on an efficient signalling equilibrium in sender-receiver games and, if so, how?

The question has been asked before. Brian Skyrms, in particular, broached it in the final chapter of *Evolution of the Social Contract*, and revisited it in *Signals: Evolution, Learning & Information*. The greatest difference between the two treatments involves a move away from use of the replicator dynamics as the underlying dynamical model to one based on stochastic reinforcement learning. One motivation for doing so derives from a critical shortcoming in attempting to use the replicator dynamics[1] to explain the emergence of signalling systems in sender-receiver games. Namely, the moment we consider signalling games of any complexity greater than the most elementary case, whether or not signalling

[1]By which I mean both the standard replicator dynamics (Taylor and Jonker, 1978) as well as the replicator-mutator dynamics (Hadeler, 1981; Hofbauer, 1985).

systems evolve turns out to depend, often critically, upon the initial state of the population.

More precisely: let a (N,K,M) sender-receiver game be one with N states of the world, K possible signals, and M possible actions. If $N \leq K$, then there are enough signals present for the Sender and Receiver to communicate efficiently, provided they agree on the interpretation. Let us call such an interpretation a *signalling system*.

The simplest interesting sender-receiver game involves two states, two signals, and two actions. Here, if states of the world are equally likely, then under the replicator dynamics a signalling system evolves almost always.[2] If the two states of the world are not equally likely, then whether a signalling system evolves under the replicator dynamics depends upon just how unequal the state probabilities are (Skyrms, 2010, pp. 64–65). Things appear to improve slightly if we consider the replicator-mutator dynamics, for Hofbauer and Huttegger (2008) show that if the state probabilities are not too unequal, then the replicator-mutator dynamics will once again almost always converge to a signalling system. Yet if the number of states, signals, and actions are all greater than two, the story changes again! In these cases, it is possible for the replicator dynamics to converge to a suboptimal outcome where the same signal is used clumsily for more than one state.[3] Whether a signalling system evolves thus depends upon accidental features such as where the system starts.

Ideally, what one would like is a model which shows the following, for all reasonable sender-receiver games. First, how individuals can coordinate upon an *efficient* signalling system, avoiding the partial pooling traps which plague the replicator dynamics. Second, how individuals can coordinate upon a *minimal* signalling system, one using the least number of signals needed. Third, we would like the model to show that signalling systems can be created out of nothing, for it seems *ad hoc* to suppose a pre-existing set of possible signals, to say nothing about imposing a limit on the total number of possible signals.

One advantage of reinforcement learning, which Skyrms shows, is that both the first and third points can be achieved:

> "Using reinforcement learning with invention, starting with no signals, 1,000 trials *all* ended up with efficient signaling. Signalers went beyond inventing the [requisite number of] signals. Lots of synonyms were created. By inventing more signals, the avoided the traps

[2] See Skyrms (1996, pg. 93). Although it is not true, in general, that "the emergence of meaning is a moral certainty," it is true in this special case.

[3] These are the so-called "partial pooling" equilibria. In a $(3,3,3)$ sender-receiver game, one example would be where signals 1 and 2 are both used for state 1, and signal 3 is used for both states 2 and 3.

> of partial pooling equilibria." (Skyrms, 2010, pg. 131)

In the above, Skyrms refers to a two-state, two-action sender-receiver game, although the point holds more generally.

In a later paper, Alexander, Skyrms, and Zabell (2011) augmented Skyrms' original model of signal invention with a model of signal destruction, or "forgetting". One particular model[4] of signal invention and forgetting, according to simulation results, yields both efficient *and* minimal signalling systems quite often.[5] Hence it appears that all three above desiderata are readily satisfiable.

Or so it appears. One should note that most of the work done on sender-receiver games, from Lewis to Skyrms, assumes a *static* environment for the signalling game.[6] The correct pairing between states and acts does not change with time. Yet the real world is not static: seasons change, predators learn, and with that the correct act, given the state of the world, may change to something different than what it was before.[7] If we consider sender-receiver games within a *dynamic* world, where there is not always a constant, correct response to the state of the world, how well does Skyrms's reinforcement model of signal invention cope?

The answer, I shall argue, is that the model does not cope very well at all; nor, for that matter, does the augmented model of inventing and forgetting signals due to Alexander et al. (2011). What does prove to be particularly effective at adjusting to a dynamic environment is a well-known aspect of human psychology: discounting the past. In what follows, I shall show that discounting the past provides a good solution to the problem of learning to signal in a dynamic world. In addition, it will, somewhat surprisingly, turn out to provide a model of *linguistic drift* between signalling systems — largely without sacrificing signal efficiency.

---

[4]The model of signal invention and forgetting is a varient Roth-Erev reinforcement learning. It is equivalent to a Hoppe-Pólya urn where balls are discarded by uniform selection by *type*. Further discussion appears in section 2.

[5]I should note that, if one changes the learning rule, it is possible to do even better. Zollman has shown that if one combines *best-response* with random experimentation upon failure, then it is always possible to learn to signal. However, if errors are permitted, then *best-response* needs to changed to *best-response with inertia*, where the introduction of inertia compensates for the occasional error. In this case, we always arrive at an efficient signalling system. The downside of this learning rule is that it requires knowledge of the set of possible payoffs.

[6]One notable exception is Barrett (2009), who considers the effect of modifying the reinforcement function used in the learning process.

[7]One might object that this problem only arises because one has failed to adopt a sufficiently fine-grained partition for distinguishing between states of the world. The difficulty is that the set of possible partitions ranges from the maximally inclusive ("everything") to the maximally individuated (where each referent of "this" falls in its own cell). How should one choose a partition? Barrett (2007) analyses this question and shows that the choice of partition is frequently underdetermined.

2. Dynamic signalling games with reinforcement learning

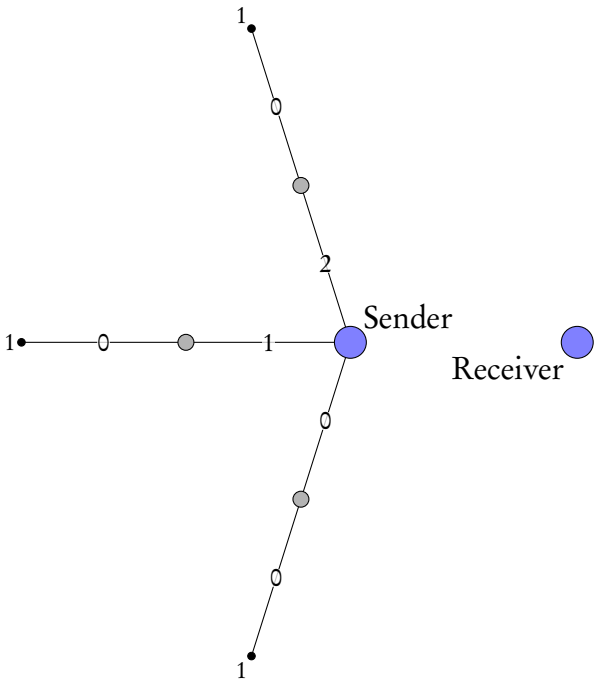
Urn models have long been used to study reinforcement learning. In their recent paper, Alexander et al. (2011) use Hoppe-Pólya urns to model a process of reinforcement learning with signal invention and forgetting. Suppose that the Sender and Receiver play a repeated N-state, N-action signalling game. The Sender begins with N urns, one for each state of the world, with each urn containing a single black ball (known as the *mutator*). The Receiver, on the other hand, begins with nothing.

When the Sender sees the state of the world, he reaches into the appropriate urn and draws a ball. If the mutator is drawn (which always happens in the first round of play), the Sender chooses a new colour not represented in the urn and sends that as a signal to the Receiver. Upon receipt of this new signal, the Receiver creates a Pólya urn with N differently coloured balls — one ball for each possible action. The Receiver draws a ball[8] at random and then performs that action. If the action was the correct response for the state of the world, the Receiver *reinforces* by adding another ball of the same colour to the urn, and *labels* that urn with the colour of the signal it responded to. (The idea being that the Receiver always uses the same urn when responding to a signal.) The Sender, likewise, retains the new coloured ball and reinforces by adding an additional ball of that colour to the urn from which it was drawn. Since a new signal has been created, the Sender adds one ball of the new colour to each of the other urns, since it is possible to send the new signal in those states of the world, too. If the action was incorrect, the Receiver discards the newly-created urn, and the Sender discards the new coloured ball.

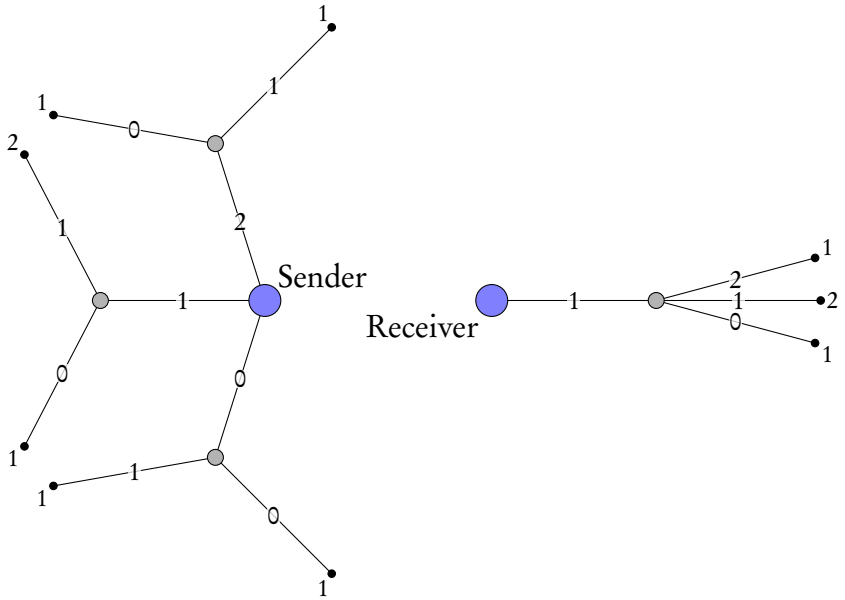
Figure 2 illustrates this process for a 3-state, 3-action signalling game. In 2(a), we see the beginning configuration where the Sender has three urns, one for each state of the world, each urn containing the black ball. Edges connect the Sender node to urn nodes, where the edge is labelled according to the state of the world. (Note that we index states of the world starting at 0.) Edges leading from an urn node to terminal nodes represent the colours present in the urn; the colour is indicated on the edge, with '0' denoting the mutator. The number of balls in the urn of a colour is listed at the terminal node. Figure 2(b) illustrates the outcome after a successful signalling attempt: in state of the world 1, the Sender tried a new signal, to which the Receiver responded successfully. The usefulness of this representation can be seen in figure 3, which illustrates the complete urn configuration for both the Sender and Receiver after 1,000 iterations.[9]

[8] All sampling is done with replacement.

[9] A minor technical point: earlier, I said that after a failed signalling attempt with a new signal, the Sender discards that colour and the Receiver discards the newly created urn. One might wonder, then, why all of the colours in figure 3 increment perfectly from 1 to 11? Shouldn't there



(a) The beginning configuration: the Sender has three urns, each containing a single black ball (indexed by '0').



(b) The configuration resulting from a successful signalling attempt: in state 1, the Sender sent signal 1, to which the Receiver responded correctly. The Receiver keeps her new urn, reinforces, and the Sender reinforces and adds the new signal to the other two urns.

Figure 2: The first correct signalling attempt.

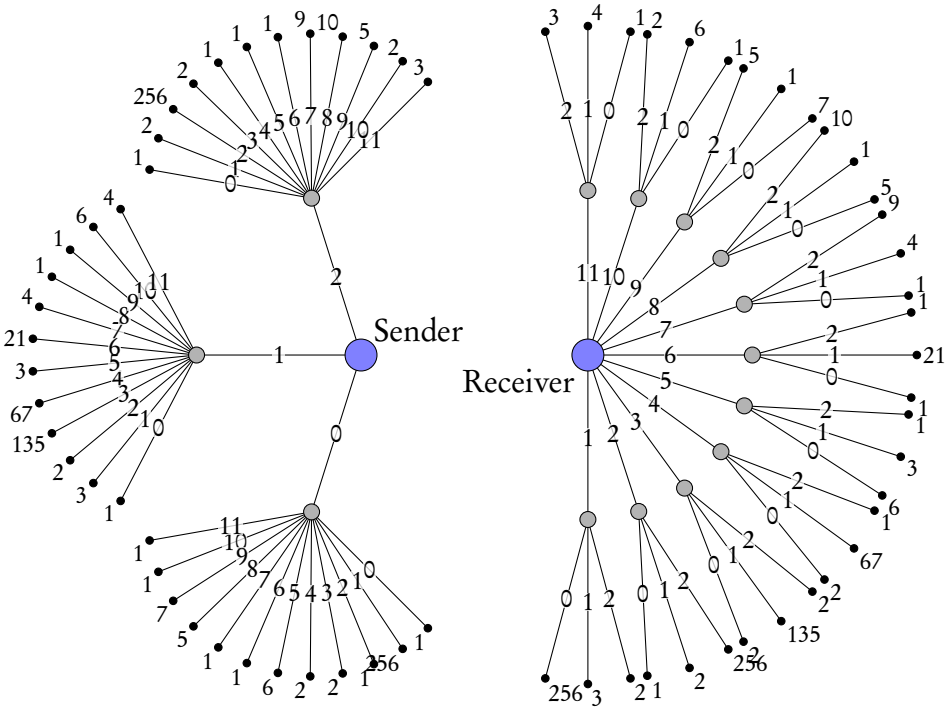


Figure 3: Urn configuration after 1,000 iterations (signal invention only).

In figure 3, we see many more signals being used than are necessary. Suppose that, from time to time, the Sender selects an urn at random, then selects a colour found in that urn at random, and then discards a ball of that colour from that urn. This method proves highly effective at reducing the number of synonyms used, as figure 4 illustrates. (For more details, see Alexander et al. 2011.)

Let us now turn to the question of how well this coordination technique copes with dynamic environments. Suppose we have a sender-receiver game consisting of N states $\{s_1, \ldots, s_N\}$ and N actions $\{a_1, \ldots, a_N\}$. For convenience, initially assume that the correct response to state s_i is act a_i. Two ways in which the game can be modified to include a dynamic environment are as follows: either a new state/action pair is introduced (representing an expansion of the signalling problem), or the correct response to an existing state of the world is altered

be gaps in this sequence? Think of it in the following way: the Sender has a shelf containing infinitely many coloured balls, each uniquely coloured. When the Sender 'discards' a coloured ball after a failed signalling attempt, he merely returns it to the shelf. The next time the Sender attempts to invent a new signal, he will use the same colour as before. But since the Receiver discarded her urn for that colour after the previous failed attempt, this is equivalent to the Sender attempting to signal with a novel colour. It proves useful to index signals in this way because, when signals can be forgotten over time, we know that any gaps occuring are due to a previously successful signal being (eventually) discarded.

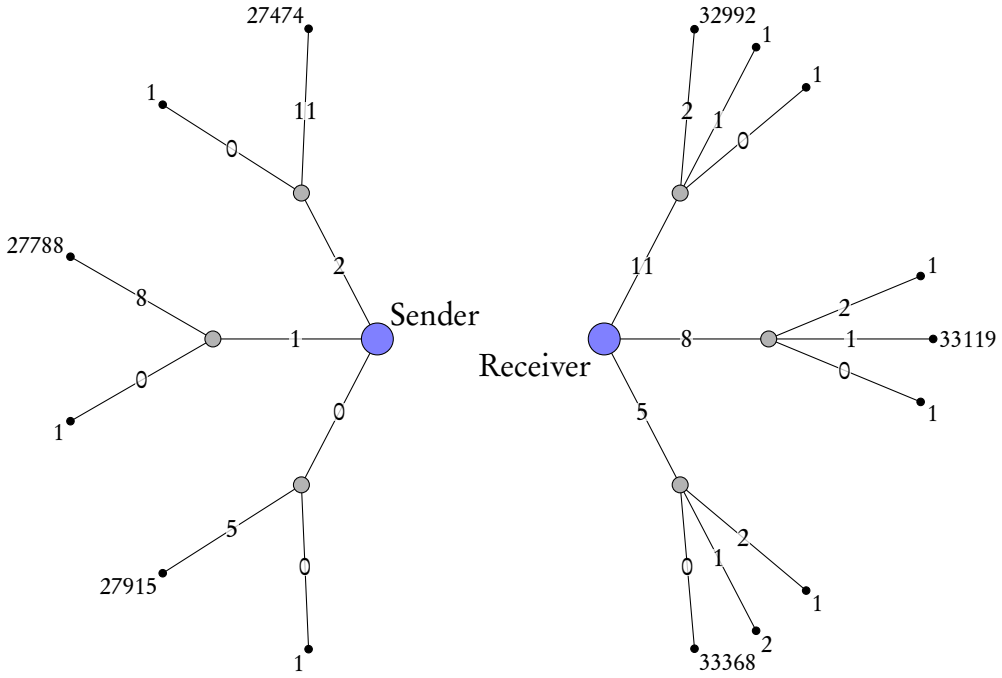


Figure 4: Urn configuration after 100,000 iterations, with both signal invention and forgetting (forgetting rate of 0.333).

(representing a lateral shift in the environment).

When a new state s_{N+1}, with corresponding act a_{N+1} is introduced, the Sender finds himself in possession of a new urn, used when the Sender observes state s_{N+1}. That urn contains, in addition to the mutator, one ball for each of the other signals in use. Likewise, a new ball representing the new available action a_{N+1} is added to each of the Receiver's response urns.

How does the addition of a new state/action pair affect the state probabilities? Recall that Nature chooses a state at random according to some given probability distribution $\Pr(\cdot)$ over N states. In what follows, I assume the new probability distribution $\Pr'(\cdot)$ over $N+1$ states to be defined as follows:

$$\Pr'(s_i) = \begin{cases} \frac{N}{N+1}\Pr(s_i) & \text{if } i \leq N, \\ \frac{1}{N+1} & \text{otherwise.} \end{cases}$$

This definition ensures that the new probability distribution $\Pr'(\cdot)$ is equiprobable if the original distribution $\Pr'(\cdot)$ was.

Alternatively, the correct response to a given state of the world might change. When this occurs, Nature selects, at random, two state-action pairs (s_i, a_i) and (s_j, a_j), permuting the correct response. After the swap, a_j is the correct action

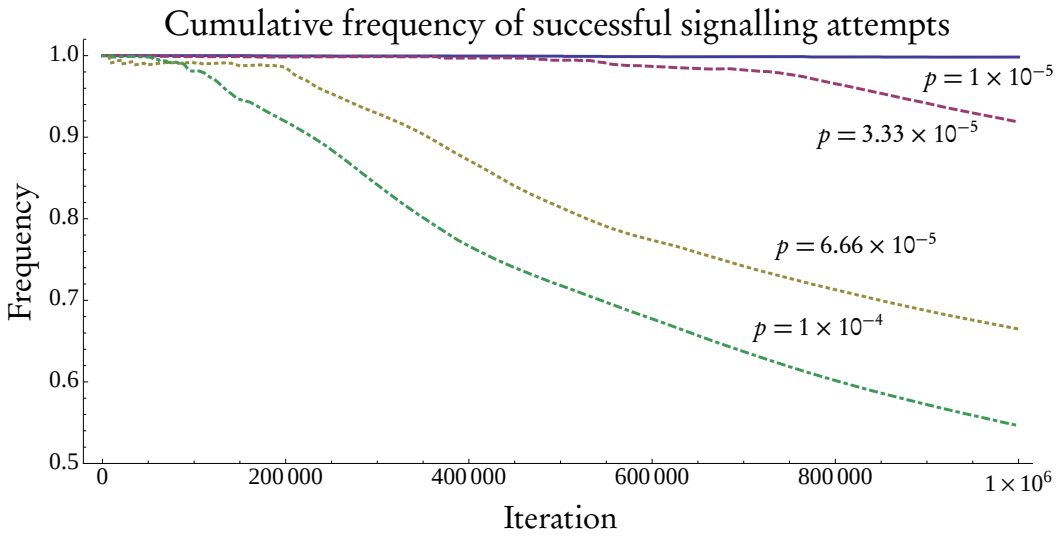


Figure 5: The inability of reinforcement learning to cope with the addition of new state-action pairs (p denotes the probability of adding new states).

to take in state s_i and a_i the correct response to take in state s_j. Unlike the case of adding new state-action pairs, this modification leaves the state probabilities unchanged.

These dynamic modifications transform the problem of learning to signal. Adding new state-action pairs ensures that the problem is of ever-growing complexity. Although we have seen it is possible to generate efficient, minimal signalling systems from nothing, one may wonder if that learning mechanism works in an open-ended signalling problem. Furthermore, the forgetting rule which prunes signals done to a minimal (or nearly minimal set) only needs to eliminate *rarely used* signals. Adjusting to a dynamic environment where the correct state-action pairs change requires the Sender and Receiver to unlearn, in a timely fashion, past associations between state, signal and action.

2.1. Introducing new states

Figure 5 illustrates four typical outcomes for a simulation of Alexander et al.'s model of inventing new signals when faced with the dynamic problem of coping with new state-action pairs. As the probability p of adding new pairs increases, the ability of the Sender and Receiver to coordinate on an efficient signalling system decreases, in some cases greatly, over time. (The downward trend present when $p = 1 \times 10^{-5}$ cannot be seen at this scale.)

It is easy to prove that, in the limit, for any fixed forgetting rate $r > 0$, and any

initial probability distribution, the cumulative frequency of successful signalling attempts for the Sender and Receiver converges to zero. To see this, consider a simpler learning problem than the one modelled: that when a new state-action pair appears, the new urn the Sender gets only contains the mutator ball, rather than one ball for each signal currently in use. Suppose that the Nth state-action pair has just been added, so that the probability that Nature presents this new state to the Sender is $\frac{1}{N}$. When the Sender observes that the state of the world is s_N, he will draw the mutator ball and send a new signal to the Receiver. The Receiver, upon receipt, chooses an action at random, and the probability she selects action a_N, the correct response, is also $\frac{1}{N}$. Since p, the rate at which new state-action pairs are added, is fixed, as N gets sufficiently large eventually $\frac{1}{N^2} \ll p$, which means that new state-action pairs will almost certainly be added *before* the Sender and Receiver manage to make any progress whatsoever on figuring out the correct response to s_N. Furthermore, suppose that the forgetting rate is fixed at r. The probability that the Sender will attempt to discard a ball from urn N is thus $\frac{r}{N}$. Since the model of forgetting used by Alexander et al. (2011) first chooses a *colour* (other than black) before discarding a ball of that colour, if N is sufficiently large, then $\frac{1}{N^2} \ll \frac{r}{N}$, which means that the approximate rate of positive reinforcement for successful signalling in state N is much lower than the rate of negative reinforcement. In short, not only are new states being added before the Sender and a Receiver have a chance to solve the previously existing signalling problem, the size of the signalling problem becomes so large that, eventually, the Sender and Receiver forget their recently-discovered correct responses faster than they play them (so as to obtain further reinforcement). Hence, in the limit, the cumulative frequency of successful signalling converges to zero.

It is no surprise that increasing the size of the learning problem by an arbitrary extent eventually swamps the ability of the learning rule to cope with it. Even *best-response for all we know with inertia*, the learning rule which "[learns] to signal with probability one in all Lewis signaling games" (Skyrms, 2010, pg. 105), cannot cope with the problem of a dynamic world in which states are added over time. The reason is slightly different: *best-response for all we know with inertia* does not include negative reinforcement, so once the Sender and Receiver manage to signal successfully once, that convention will continue. (Even in the face of occasional error, given the inertia.) Yet, as above, the time required for the Sender and Receiver to coordinate on a signalling scheme for state s_N grows as $O\left(\frac{1}{N^2}\right)$. Eventually new states will be added faster than the Sender and Receiver can augment their signalling scheme, and so the cumulative frequency of successfully signalling attempts converges to zero. However, because *best-response for all we know with inertia* will eventually establish a successful signalling scheme for all

of the new states (it just takes longer and longer to do so), a curious corollary is that, *in the limit*, the Sender and Receiver will learn to successfully signal for all of the infinitely many states — even as the cumulative frequency of successful signalling attempts converges to zero![10]

2.2. Swapping state/action pairs

Now consider the alternate problem in which the correct response to a given state of the world periodically changes, as when the seasons change or a predator acquires a new mode of attack. Suppose that we have a 3-state, 3-action signalling game with a forgetting rate of 0.333 (chosen because that rate frequently yields efficient and minimal signalling systems), and a state swap probability of $p = 1.0 \times 10^{-5}$. How well do reinforcement learners cope with this dynamic environment?

As figure 6 illustrates, reinforcement learners do not cope with state swaps very well at all. (The alternating shaded regions identify periods where the state-action pairs are held constant: a swapping of state-action pairs occurs at the transitions.) Although reinforcement learners do very well at solving the initial signalling problem, the fact that a Hoppe-Pólya urn does not place an upper limit on the number of balls of a given colour means that the Sender and Receiver can "lock in" to a given signalling system.[11] This lock-in proves difficult to unlearn, as the moving frequency in figure 6 shows. The moving frequency plots the average rate of successful signalling attempts over the last 100 iterations. Initially, the Sender and Receiver establish a signalling system for the original problem. After the first swap occurs, the lock-in persists for over 75,000 iterations, with successful signalling attempts occurring about a third of the time. Yet once the Sender and Receiver coordinate on a signalling system for the new problem, eventually another swap will occur, requiring that the process of unlearning begin again.

Hence we see that, although the model of reinforcement learning of Alexander et al. (2011) frequently achieves efficient, minimal signalling systems, the method of reinforcement learning used has too slow of a response curve to cope effectively with a changing environment. Although figure 6 only shows the outcome for a 3-state, 3-action signalling problem, it is clear that the problem exists for other N-state, N-action signalling problems as well. Increasing the number of states does reduce the extent to which lock-in occurs to a successful signal in a given urn, but there is also a corresponding decrease in the frequency with which signals are removed from urns. (In the case of equiprobable states, these exactly cancel each other out.)

[10] This is, after all, just the Tristram Shandy paradox.

[11] See Barrett (2007); Barrett and Zollman (2009), though, who look at models of reinforcement learning with a cap.

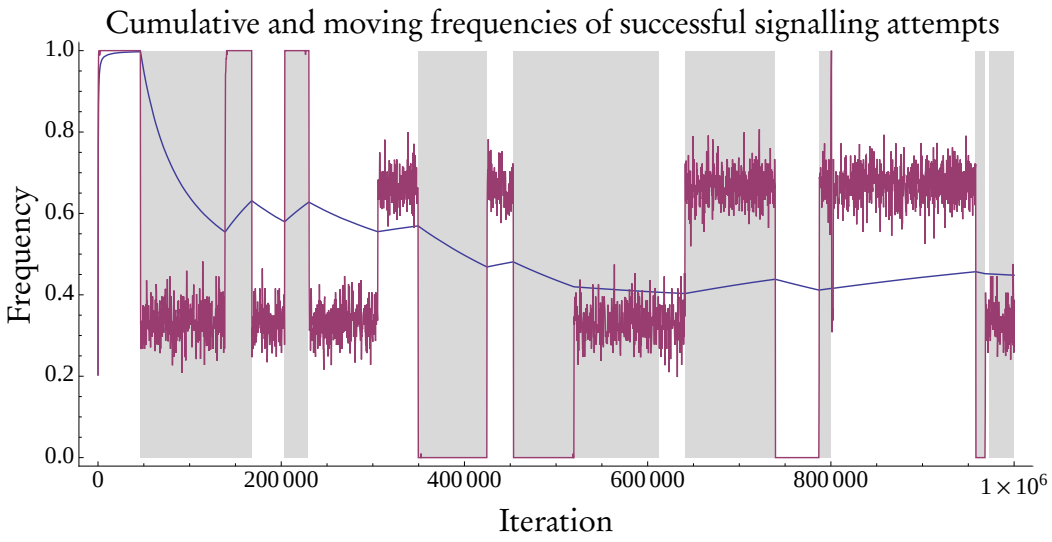


Figure 6: The inability of reinforcement learners to adjust to a dynamic environment (3-states, 3-actions).

In contrast, *best-response for all we know with inertia* easily adjusts. Suppose that a swap occurs in a signalling game having N states and N actions. The next time the players encounter either swapped state, they will recognise that their previously coordinate upon action is no longer the correct response. They then will search through the space of possible actions at random until they happen upon the correct response. (A precise calculation of the expected waiting time until this happens can be found in appendix A.) Yet one worry is that even this not particularly strategic form of best-response requires some knowledge of the payoff matrix. Is there no learning rule in between the reinforcement learning discussed by Skyrms (2010) and best-response which does the trick?

3. Discounting the past

In economics and finance, it is frequently assumed that rational agents trade off present amounts against uncertain future gains by discounting the future. Much debate exists over the exact form such temporal discounting takes[12] but *that* temporal discounting occurs is beyond question. Similarly, others (see Charles Wolf, 1970; Caplin and Leahy, 2004) have suggested that rational agents discount the *past* as well.

Setting aside the question of whether it is *rational* to discount the past, it is

[12]For example, whether discount rates are constant over time, the extent to which discount rates are frame dependent, whether people use exponential or hyperbolic discounting, and so on.

certainly a part of human psychology. Memories are imperfect and fade over time. When I decide whether to carry an umbrella when I leave the house, I am more likely to rely on what I learned yesterday, regarding the weather, than what I learned last week or last month.

Consider, then, the following modification of the Hoppe-Pólya urn model of inventing new signals: suppose that the Sender begins with N urns, one for each state of the world, as before. However, instead of the urns containing discrete coloured balls, let us use coloured liquids, instead. If the urn has a slow leak, and all of the liquids drip out at the same rate, except for the black liquid, this corresponds to a model of signal invention and reinforcement learning with discounting of the past. The rate of the leak being the degree older information is discounted in favour of more recent information.

More precisely, let us represent an urn u as a tuple of ordered pairs. An urn with n signals, plus the mutator, would be as follows:

$$u = \langle (m, w), (s_1, w_i), \ldots (s_n, w_n) \rangle.$$

The ordered pair (m, w) represents the mutator and its weight, and the ordered pairs (s_i, w_i) represent signals and their associated weights.[13] As before, signal s_i is sent with probability $\frac{w_i}{w+\sum_k w_k}$. The probability that the mutator is selected, and an attempt is made at inventing a new signal, is $\frac{w}{w+\sum_k w_k}$.

If the weights are restricted to the natural numbers, we have an ordinary Hoppe-Pólya urn with discrete balls. If the weights are permitted to range over the nonnegative reals, we have a Hoppe-Pólya urn which admits discounting of the past. If $\beta \in (0, 1]$ denotes the discount factor, a discounted urn u' is obtained from the urn u in the expected way:

$$u' = \langle (m, w), (s_1, \beta w_i), \ldots (s_n, \beta w_n) \rangle.$$

Notice that only the weights attached to signals have the discount factor applied to them; the weight attached to the mutator is not discounted.

Discounting the past prevents lock-in by effectively placing an upper limit on the amount of reinforcement. Even if the weight on a signal was reinforced at the end of every round by adding 1, the cumulative weight would never exceed $\frac{1}{1-\beta}$. For psychologically plausible discount factors of 0.95 or 0.99, the cumulative weights are capped at either 20 or 100, respectively.

Although discounting caps the total weight a signal may have, it does not prune disused signals. A signal that was only used once will, with a discount factor of 0.99, have a weight of approximately 2.25×10^{-44} after 10,000 iterations.

[13] Strictly speaking, the notation for signals and their weights should include an additional index for the urn they are in. I have suppressed that, here, to reduce notational clutter.

Since signals are purely conventional and invented, there is presumably a cost to maintaining signals as viable possibilities. Given that, let us introduce a *cutoff threshold* τ such that, if the weight of a signal in an urn dips below τ, then that signal will be removed. From the Receiver's point of view, the contents of her urns represent possible *actions*, rather then signals. Because the set of possible actions are determined by what is physically possible, and aren't conventional and invented, when the Receiver discounts the weights in her urn, the cutoff threshold isn't applied.

To summarise, the model of inventing and discarding signals with discounting the past is similar to that of Alexander et al. (2011) with the following variations: at the end of each round of play, the weights attached to the contents of both the Sender's and Receiver's urns are discounted by a common factor of $\beta \in (0, 1)$. A cutoff threshold of $0 < \tau$ is applied to the weights in the Sender's urn, removed signals whose weight dips below τ. How does this model cope with a dynamic environment?

3.1. Learning to signal in a dynamic world

Concerning the case of adding new state-action pairs, in the limit, the Hoppe-Pólya urn with discounting suffers from the same problem as the original urn model: eventually, the number of new states will swamp the ability of the Sender and Receiver to coordinate on a correct response and retain it. This is not a surprise: *discount-the-past* deinforces weights more aggressively than the model of forgetting used by Alexander et al. (2011) by acting on all weights in all urns at the end of each iteration.

That said, judicious selection of initial weight for the mutator, discount factor, and cutoff threshold has *discount-the-past* coping rather well the problem of adding new states, up to a point. Suppose that the initial mutator weight is 0.001, the discount factor is 0.99, and the cutoff threshold is 0.0001.[14] Comparison of figure 7 with figure 5 shows *discount-the-past* outperforming the original Hoppe-Pólya urn model previously discussed.

It should be noted that the outperformance is largely an artefact of the initial weight attached to the mutator. In the original Hoppe-Pólya urn model, the mutator had a weight of one, and any new signals introduced had a weight of two, with reinforcement by one afterwards for each correct signalling attempt. As (Skyrms, 2010, pg. 97) noted, for the case of Roth-Erev[15] reinforcement learning,

[14] There is little reason in having the cutoff threshold less than an order of magnitude below the mutator weight. As the mutator weight isn't discounted, it is more likely that the Sender will attempt to generate a new signal than revive a disused-but-correctly-interpreted old signal.

[15] It should be noted that, at this point, Skyrms was only considering the ease with which a signalling system could be reached via a process of reinforcement learning. The set of possible

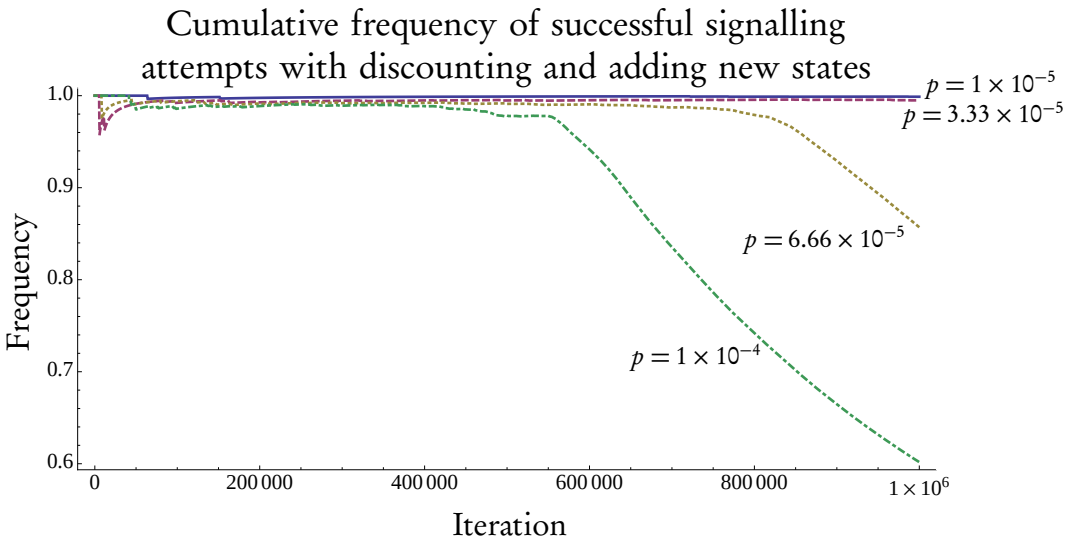


Figure 7: Discounting the past copes with the addition of new state-action pairs, up to a point. In the plot, p denotes the probability of adding new states. Here„ $m = 0.001$, $\tau = 0.0001$ and $\beta = 0.99$.

the initial weights make an enormous difference in whether a partial pooling equilibrium emerges. Given that, it comes as no surprise that initial weights make a difference in this case as well.

If we lower the initial weight of the mutator to 0.01, the original urn model behaves more like *best-response*, with a small probability of error. What was initially a 2-to-1 chance of re-using a correctly interpreted signal on the second try now becomes a 200-to-1 chance. This speeds up the initial learning phase, giving the Sender and Receiver more time to build up a buffer before deinforcement overwhelms their ability to establish a signalling system as the problem space grows.[16]

The aggressive deinforcement provided by discounting the past turns out to be extremely efficient in adjusting to dynamic environments where state-action pairs are swapped. Figure 9 illustrates two typical simulation results — one for a 5-state, 5-action signalling game and the other for a 10-state, 10-action signalling game. The two plots show both the cumulative frequency of successful signalling attempts and a moving average of length 1,000. The sharp dips in the moving

signals was fixed beforehand, with the question being whether the Sender and Receiver could spontaneously arrive at a signalling system. The problem of *inventing* signals, which lead to the Hoppe-Pólya urn model discussed here, was not broached until later.

[16]Nevertheless, both learning rules do not do as well as *best-response for all we know with inertia*. The reason, of course, is that a best-response rule, once it establishes a correct response to a signal sent in a given state of the world, will continue to employ that same response in the future.

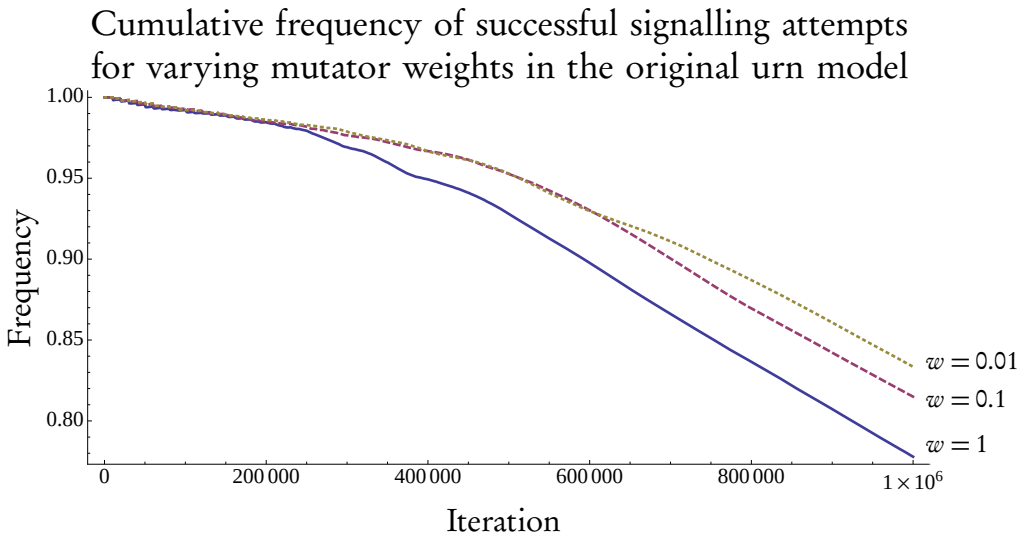


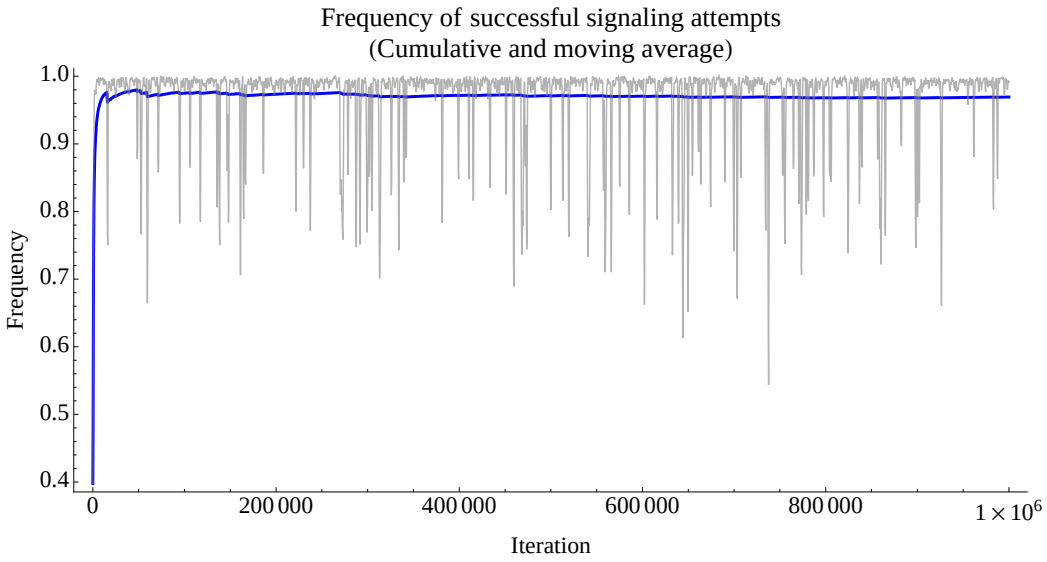
Figure 8: The effect of varying initial mutator weights on the Hoppe-Pólya urn model, with the addition of new states. Beginning from a 1-state, 1-action signalling problem, a new state was added everything 10,000 iterations. Here, $\tau = 0.001$ and $\beta = 0.99$, with the initial mutator weight w as shown.

average indicate points where the correct state-action response for two randomly selected states were swapped. Even so, the cumulative effect was minimal, with the Sender and Receiver maintaining a signalling system with a success rate greater than 95% in both cases.

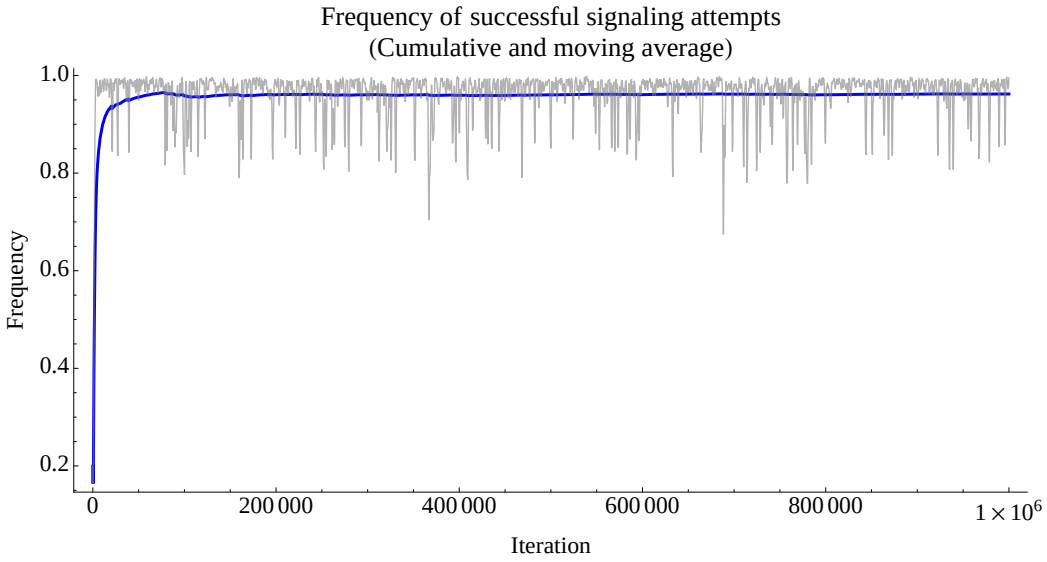
The ability of the Sender and Receiver to coordinate on a signalling system requires an appropriate matching of the discount factor and cutoff threshold to the size of the signalling problem. One can show that, for any fixed discount factor β and cutoff threshold τ, there are N-state, N-action signalling problems which are too large for the players to solve in general. Hence, it follows trivially that they will be unable to cope with the additional dynamic challenge of learning when the correct act for a given state has changed. To see this, first assume (without loss of generality) that all states are equiprobable. Then the expected waiting time between the ith and $i+1$th observation of state k by the Sender is N iterations. Suppose that $N > \frac{\log(\tau/2)}{\log(\beta)}$, and suppose that the Sender had invented a new signal after the ith observation of state k, and the Receiver had responded to it correctly. Then the typical weight attached to that signal, the next time state k is encountered, would be

$$2\beta^N < 2\beta^{\frac{\log(\tau/2)}{\log(\beta)}} = \tau.$$

Since $2\beta^N$ falls below the cutoff threshold, that signal would have been eliminated



(a) Simulation results for a 5-state, 5-action signalling problem.



(b) Simulation results for a 10-state, 10-action signalling problem.

Figure 9: Adjusting to a dynamic environment involving a swap of state-action pairs by discounting the past. Here, the simulations used a Hoppe-Pólya urn model with a discount factor of 0.99, initial mutator weight of 0.1, and a cutoff threshold of 0.01. The probability of swapping state-action pairs was 0.0001.

from the urn before the $i+1$th encounter of state k. In other words, even if the Sender and Receiver had managed to establish a signalling convention through trial-and-error, they will be unable to retain that convention over time due to infrequent reinforcement.[17]

In summary, Skyrms's model of Hoppe-Pólya reinforcement learning, augmented with the ability to forget signals, cannot cope with a dynamic environment in which the correct response to a given state of the world is exchanged with another. However, Hoppe-Pólya reinforcement learing with *discounting* proves remarkable effective at adjusting to a dynamic environment. And *both* models of reinforcement learning become overwhelmed for any fixed set of learning parameters (i.e., forgetting rate, size of the discount factor, and so on) as the signalling problem size grows.

Since *best-response for all we know* also adjusts to the two types of dynamic environments considered here, why bother at all with reinforcement learning, discounted or not? There are three reasons. First, reinforcement learning does not require that players have some knowledge of the possible payoffs. Second, as the analysis in appendix A shows, *best-response for all we know* does not handle the dynamic problem of swapping states significantly better than that of reinforcement learning with discounting, once the problem becomes sufficiently large. Finally, reinforcement learning with discounting turns out to automatically generate an additional phenomenon associated with natural languages: linguistic drift.

3.2. Generating linguistic drift

In the models of reinforcement learning, both with or without signal invention, studied by Skyrms (2010), signalling systems are unlikely to change, once established. In order for an urn model of reinforcement learning to switch from one signalling system to another, an extremely unlikely series of events would have to occur. The same holds true for the model of signal invention with forgetting of Alexander et al. (2011). Yet natural languages are continuously evolving entities. And although the English spoken by Chaucer is radically different from the English spoken today (so much so as to be largely unintelligible to contemporary speakers of English), the transition between the two forms occurred with successful communication at each point along the way.

One interesting consequence of reinforcement learning with discounting in

[17] We may assume that all states are equiprobable without loss of generality because this assigns the greatest possible probability to all states. Any distribution which is not equiprobable necessarily assigns more than $\frac{1}{N}$ weight to at least one state and less than $\frac{1}{N}$ weight to at least other state. States assigned less than $\frac{1}{N}$ weight will have successful signals discounted out of the Sender's urn more rapidly, and so the result holds for those states, too.

sender-receiver games is that linguistic drift occurs. Figure 10 illustrates one such outcome for the Hoppe-Pólya urn model with a discount factor of 0.97. Initially, the Sender and Receiver establish a signalling scheme using signals 1, 3, and 4 to denote the three states of the world. Further into the game, signal 4 is dropped, with signal 32 used in its place. And later still, signal 3 is exchanged for signal 61. As the cumulative frequency plot shows, this occurs without the players communicative ability being impaired at all.

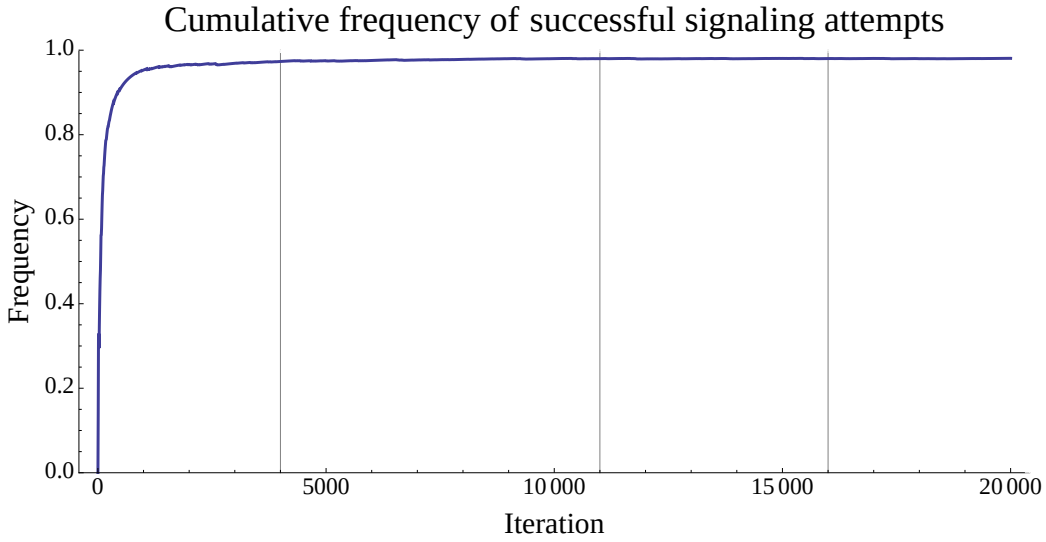
The mechanism underlying drift is straightforward. With a discount factor of $\beta = 0.97$, the maximum weight attainable by a signal is $\frac{100}{3} \approx 33.3$. (In practice, since any state occurred only a third of the time, the effective weights were around a third of the maximum.) Hence there was roughly a 1 in a 100 chance that the mutator would be selected when a given state of the world occurred. This ensures a steady stream of synonymous signals being generated. Since the weight attached to a newly-created synonym is considerably lower than the previous signal, the odds are that the synonym will be played less frequently and, eventually, eliminated once its weight dips below the cutoff threshold. But this does not always occur: sometimes the synonym not only persists but winds up being used more frequently than the former signal. When this happens, the weight attached to the *former* signal can decrease over time until it is eliminated.

4. Conclusion

The majority of work done on Lewis sender-receiver games has assumed a fixed relation between the state of the world and the correct action to be performed in that state. Dynamic sender-receiver games, where the state-action relation can vary over time, present a much harder problem for boundedly rational agents to solve. Although positive reinforcement of correct responses is required, as this enables a solution to the signalling problem to be learned, this must be balanced against the need to prevent lock-in, as this prevents players from being able to adjust rapidly to changes in the environment.

We have seen that none of the models of reinforcement learning discussed by Skyrms (2010) or Alexander et al. (2011) are capable of solving the two types of dynamic sender-receiver games considered here. However, a model of reinforcement learning in which the past is discounted proves surprisingly effective at being able to handle dynamic sender-receiver games. In some cases it is roughly on par with that of *best-response for all we know*.

One interesting possibility for future research concerns revisiting the model of Barrett (2007), in which sender-receiver games are used to show how incommensurable sets of kind terms may be generated from a common starting point. Given that discounting the past gives rise to linguistic drift, one wonders if it



	Signals					Actions		
	mutator	1	3	4		Act 0	Act 1	Act 2
State 0	0.1	0	11.1	0	Signal 1	ϵ	11.5	ϵ
State 1	0.1	11.5	0	0	Signal 3	11.1	ϵ	ϵ
State 2	0.1	0	0	9.52	Signal 4	ϵ	ϵ	9.52

(a) 4,000 iterations

	Signals					Actions		
	mutator	1	3	32		Act 0	Act 1	Act 2
State 0	0.1	0	9.76	0	Signal 1	ϵ	10.8	ϵ
State 1	0.1	10.8	0	0	Signal 3	9.76	ϵ	ϵ
State 2	0.1	0	0	11.3	Signal 32	ϵ	ϵ	11.3

(b) 11,000 iterations

	Signals					Actions		
	mutator	3	32	61		Act 0	Act 1	Act 2
State 0	0.1	10.2	0	0	Signal 3	10.2	ϵ	ϵ
State 1	0.1	0	0	13.9	Signal 32	ϵ	ϵ	8.14
State 2	0.1	0	8.14	0	Signal 61	ϵ	13.9	ϵ

(c) 16,000 iterations

Figure 10: An illustration of linguistic drift between signalling systems while maintaining nearly perfect communication. The game was a 3-state, 3-action sender-receiver game, all states equiprobable, with a discount factor $\beta = 0.97$. (ϵ represents a negligible weight.)

would be possible to show drift occurring in *kind terms*. Barrett (2009) has shown that changing the reinforcement function can cause the players to arrive at a signalling system formally incommensurable with one used earlier. However, as Barrett notes, this happens when "the agents are punished so severely by the new reinforcement functions for failing to capture finer-grained distinctions that their propensities are pushed low enough that they must retool and start over again in evolving the new language." If players effectively start from scratch, it is not surprising that the new signalling system reached may be incompatible with the one used previously. But suppose that players could move from one set of kind terms to a new set of kind terms, incommensurable with the first, communicating each step of the way. Would that not constitute a counterexample to Kuhn's claim that communication across incommensurable paradigms is impossible? Determining whether such linguistic drift is possible remains an open question.

A. Markov chain analysis

We can calculate the mean time for *best-response for all we know, with inertia* to adjust to a swap of the state-action pairs by modelling it as a Markov process. Assume that we have an N-state, N-action signalling game with state probabilities specified by $\langle p_1, \ldots, p_N \rangle$. Suppose also that the Sender and Receiver had coordinated upon a signalling system when the correct response to two states of the world were exchanged.

Let us adopt the following notation for describing states of the Markov chain:

$[s_k]$ The correct response to state s_k has changed, although this fact is currently unknown by the players.

$s_k?$ The players know that the past response to state s_k is no longer correct, but they have not yet figured out a signalling scheme which produces the correct response.

$\widehat{s_k}$ The Receiver has performed the correct action to a new signal for state s_k, thereby establishing a new component of a signalling system.

Now suppose that the correct response for states i and j have been swapped. This is the starting state of the Markov process, indicated by the leftmost node in figure 11. Because neither the Sender nor the Receiver know that the correct response to states i and j have changed, this node is labelled '$[s_i][s_j]$' as per the above.

Once the players know that the previous responses to states i and j are no longer correct, let us assume that the Sender throws away the signals formerly associated with those states and sends a new signal each time state i or j occurs. Let us also assume that the Receiver, upon receipt of a new signal, selects an action at random. Thus the transition probability from a state labelled with '$s_i?$' to a node labeled with '$\widehat{s_i}$' is $\frac{p_i}{N}$ (and *mutatis mutandis* for state j). It is obvious that the state '$\widehat{s_i}\widehat{s_j}$' at the right of figure 11 is the sole absorbing state.

Let $\vec{k} = \langle k_1, \ldots, k_9 \rangle$ denote the vector of mean hitting times of the absorbing state. Obviously, $k_9 = 0$. The values of the remaining k_i are related according to

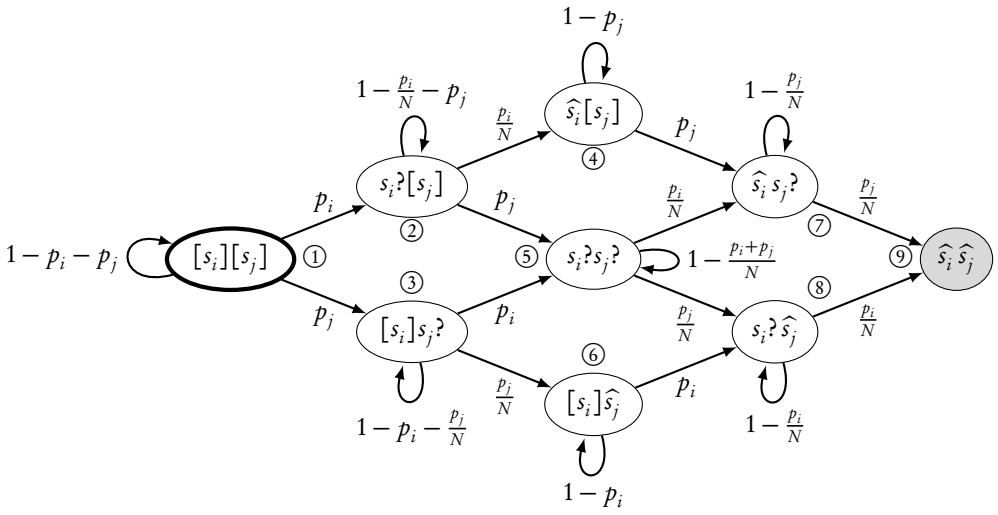


Figure 11: The Markov process describing how the Sender and Receiver adjust to a swapping of state-action pairs using *best-response for all we know, with inertia*. States have been labelled with numbers for purposes of reference.

the following equations:

$$
\begin{aligned}
k_1 &= 1 + k_1\left(1-p_i-p_j\right) + k_2 p_i + k_3 p_j \\
k_2 &= 1 + k_2\left(1-\frac{p_i}{N}-p_j\right) + k_4\frac{p_i}{N} + k_5 p_j \\
k_3 &= 1 + k_3\left(1-p_i-\frac{p_j}{N}\right) + k_5 p_i + k_6\frac{p_j}{N} \\
k_4 &= 1 + k_4\left(1-p_j\right) + k_7 p_j \\
k_5 &= 1 + k_5\left(1-\frac{p_i+p_j}{N}\right) + k_7\frac{p_i}{N} + k_8\frac{p_j}{N} \\
k_6 &= 1 + k_6\left(1-p_i\right) + k_8 p_i \\
k_7 &= 1 + k_7\left(1-\frac{p_j}{N}\right) \\
k_8 &= 1 + k_8\left(1-\frac{p_i}{N}\right)
\end{aligned}
$$

Solving this for k_1, the mean hitting time of the absorbing state when we begin at

state 1, gives:

$$k_1 = \frac{1}{p_i} + \frac{p_i}{p_j\left(p_i+p_j\right)} + \frac{Np_j^2}{p_i\left(p_i+p_j\right)\left(Np_i+p_j\right)} + \frac{(N+1)N^2p_j}{\left(Np_i+p_j\right)\left(p_i+Np_j\right)} + \frac{Np_i^2\left(Np_i+N^2p_j+Np_j+p_j\right)}{p_j\left(p_i+p_j\right)\left(Np_i+p_j\right)\left(p_i+Np_j\right)}.$$

Assuming that all states are equiprobable, so that $p_i = p_j = \frac{1}{N}$, the expected time for the players to arrive at a solution to the two swapped states when $N = 3$ is 16.875. When $N = 20$, the time increases to 620.476, and when $N = 30$, it is 1380.48.

Now consider the related problem of how quickly players who use Hoppe-Pólya urns with discounting can adapt to the problem of swapped states. For simplicity, assume that discounting has already depleted the Sender's urns for states i and j of everything except the mutator. Figure 12 illustrates a simple Markov chain modelling this system.

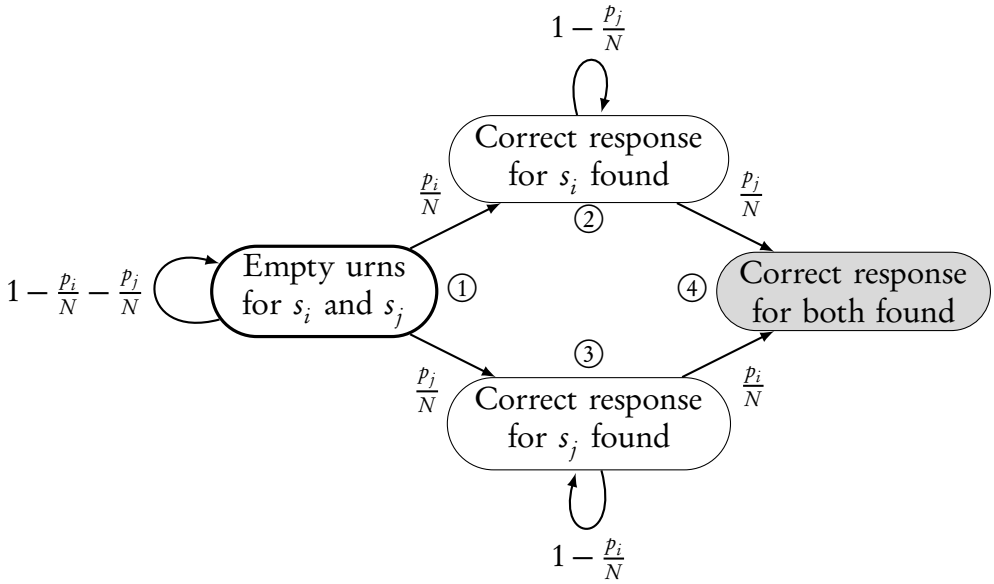


Figure 12: A Markov process approximating how the Sender and Receiver adjust to a swapping of state-action pairs using Hoppe-Pólya urns with discounting.

As per the standard method, the set of recurrence equations which need to be

solved are as follows:

$$k_1 = 1 + k_1\left(1 - \frac{p_i}{N} - \frac{p_j}{N}\right) + k_2\frac{p_i}{N} + k_3\frac{p_j}{N}$$
$$k_2 = 1 + k_2\left(1 - \frac{p_j}{N}\right)$$
$$k_3 = 1 + k_3\left(1 - \frac{p_i}{N}\right).$$

The solution is $k_1 = \frac{Np_i}{p_j(p_i+p_j)} + \frac{N}{p_i}$. If, as before, we assume equiprobable states, then the mean hitting time of the absorbing state when $N = 3$ is 13.5. Larger signalling problems, such as when $N = 20$ have a mean hitting time of 600, which is slightly *less* then that of *best-response for all we know, with inertia*. For $N = 30$, the time is 1350, also slightly less than the best-response time. Of course, not reflected in these times is the wait needed to make the assumption that the Sender's urns for i and j are empty except for the mutator. With a discount factor of 0.95 and a cutoff threshold of 0.1, the additional waiting time is on the order of 60 iterations. Overall, best-response is slightly more efficient than discounting for larger signalling problems, but not significantly so.

References

J. McKenzie Alexander, Brian Skyrms, and Sandy Zabell. Inventing new signals. *Dynamic Games and Applications*, 2011.

Jeffrey A. Barrett. Dynamic partitioning and the conventionality of kinds. *Philosophy of Science*, 74:527–546, 2007.

Jeffrey A. Barrett. Faithful description and the incommensurability of evolved languages. *Philosphical Studies*, 147(1):123–137, 2009.

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

Andrew Caplin and John Leahy. The social discount rate. *Journal of Political Economy*, 112(6):1257–1268, 2004.

Jr. Charles Wolf. The present value of the past. *Journal of Political Economy*, 78 (4):783–792, 1970.

K. P. Hadeler. Stable polymorphisms in a selection model with mutation. *SIAM Journal of Applied Mathematics*, 41:1–7, 1981.

J. Hofbauer. The selection-mutation equation. *Journal of Mathematical Biology*, 23:41–53, 1985.

J. Hofbauer and S. Huttegger. Feasibility of communication in binary signaling games. *Journal of Theoretical Biology*, 254:843–849, 2008.

Brian Skyrms. *Evolution of the Social Contract*. Cambridge University Press, 1996.

Brian Skyrms. *Signals: Evolution, Learning, & Information*. Oxford University Press, 2010.

Peter D. Taylor and Leo B. Jonker. Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, 40:145–156, 1978.