Draft Jan. 2011

Learning to Signal
with Two Kinds of Trial and Error

Brian Skyrms

1. Low Rationality Game Theory

High rationality game theory is built on idealizations that may be hard to justify such as – at the minimum – common knowledge of strategic interaction and common knowledge of rationality of the interacting agents. Low rationality game theory investigates interactions between more limited agents. At the most modest level, agents may not even be aware that they are in strategic interaction, and may just muddle along with trial and error learning. If low rationality dynamics leads to the same result as high rationality equilibrium analysis, it lends support to high rationality game theory. If they disagree, this raises questions.(Roth and Erev 1995, Erev and Roth 1998, Young 2004)

It is of special interest to investigate learning to signal with in a low rationality setting for two reasons. The first is that some fairly robust signaling system must be already in place to support (or even approximate) the assumptions of high rationality. So signaling must presumably have its origin in a low rationality setting. The second is that animals incapable of high rationality can learn to signal.

But there is more than one kind of trial and error learning. Here we focus on two kinds of trial and error, and compare their success in learning to signal. The first is the simplest form of reinforcement learning used by Roth and Erev (1995), Erev and Roth(1998) - which they trace to Herrnstein (1970). The second is Probe and Adjust dynamics. A somewhat more complicated form of probe and adjust was introduced by Kimbrough and Murphy (2009) in an analysis of tacit collusion in oligopoly pricing. The simple form used here was introduced by Skyrms (2010) and analyzed by Huttegger and Skyrms (forthcoming) to investigate learning an optimal signaling network.

2. Two kinds of trial and error learning

Reinforcement

An individual chooses repeatedly between actions $A_1 \ldots A_n$. At each point in time the probability of choosing A_i is proportional to the accumulated past payoffs for choosing A_i. To get the process started, we assume initial "virtual payoffs" equal to 1 for each act. Thus the learner starts by choosing at random, and then the evolution of choice probabilities is driven by payoffs.

The stochastic process may be realized by an urn scheme. There are balls of colors $C_1 \ldots C_n$ in and urn, initially one ball of each color. A ball is drawn at random from the urn – say of color C_i and replaced, and the corresponding action, Ai, is taken. A payoff (possibly zero, but non-negative) is realized and the number of balls equal to that payoff of color C_i is added to the urn. This is repeated.

Probe and Adjust

An individual *probes* by just trying an act at random, and then *adjusts* by (i) adopting the new act if it has a higher payoff than the old one (ii) going back to the old act if the new one got a lower payoff or (iii) flipping a coin to decide if they are equal. In a repeated choice situation an individual *probes and adjusts* with small probability, e, and just chooses the same as last time with probability (1-e). We can start the process by just choosing at random.

Comparison

Although both these processes are kinds of trial and error learning with at least some psychological plausibility, they are quite different in character and perform differently in different learning situations. Their analysis calls for different mathematical methods. The transitions in Roth-Erev reinforcement depend on the number of balls in each urn, and thus require some memory of the whole history of the process. It slows down as the reinforcements pile up, and approximates better and better a mean field deterministic dynamics. The transitions in probe-and-adjust depend only on a comparison of probe and pre-probe payoffs, and so only require a limited memory. Reinforcement learning is analyzed using stochastic approximation theory. (Pemantle 2007) Probe and adjust uses Markov chains. For an initial comparison, we apply them to a two-armed bandit learning problem.

You have two slot machines, R; L, each of which pays off with a different unknown probability. (Trials are independent and identically distributed, and are independent between machines.) Can you learn to play the optimal machine? Roth-Erev reinforcement learning converges to playing the highest paying machine with probability one. (Beggs 2005). Here is a sketch of the stochastic approximation approach.

Let L pay 1 with probability p and 0 with probability (1-p). Let R pay 1 with probability q and 0 with probability 1-q. Our learner has an urn with one R ball and one L ball, and proceeds with reinforcement learning. Let N = number of R balls + number of L balls. The probability of choosing R at a given time is then just the number of R balls divided by N.

We start be calculating the expected change in the probability that the learner chooses R, p(R). First calculate the expected value of p(R) after one trial. One of four things can happen: (1) R is chosen and reinforced, (2) R is chosen and not reinforced, (3)

L is chosen and reinforced, (4) L is chosen and not reinforced. Accordingly, we calculate the expectation of the next value of p(R) as:

	Chosen?		*Reinforced?*		*New Value of p(R)*	
1.	[p(R)	*	q	*	(N p(R) +1)/(N+1)]	+
2.	[p(R)	*	(1-q)	*	p(R)]	+
3.	[(1-p(R))	*	p	*	(N p(R))/(N+1)]	+
4.	[(1-p(R)	*	(1-p)	*	p(R)]	

Subtracting the current value, p(R) gives us the expected increment. We get:

(1/N+1) p(R) (1- p(R) (q-p) *(expected increment)*

The value 1/(N+1) is the *step size*. This tells us that the process slows down at a rate such that the stochastic process approximates the mean field dynamics with higher and higher probability as N builds up. The rest of the equation gives us the mean field dynamics:

d p(R)/dt = pr(R) (1-pr(R) (q-p) *(mean field dynamics)*

Here learning must converge to one of the rest points of the mean field dynamics. If the two machines pay off equally, q=p, then every point is a rest point of the mean field dynamics. In this case the urn is a Polya urn and the learner can converge to anything.

If the R pays off more often than L, then there are only two rest points, a stable attracting equilibrium at p(R)=1, and an unstable equilibrium at Pr(R)=0.

L →→→→→→→→→→→→→→→→R

Reinforcement learning must converge to one of these points. The instability of the latter equilibrium suggests that learning will never converge to it, and thus will always converge to always playing R. That is correct, but it requires a special argument to show it. (Hopkins and Posch 2005).

In probe-and-adjust learning, nothing happens except when there is a probe. We can analyze it by looking only at pre-probe and post-probe states. This embedded sequence is a Markov chain. We can analyze as follows: Suppose bandit L pays off with probability p and bandit R pays off with probability q. The state of playing bandit R transitions to that of playing bandit L with probability:

p(1-q) [R doesn't pay, probe, L does pay. Switch to L]

+ ½ pq [both pay off, flip a coin to decide whether to switch]
+ ½ (1-p)(1-q) [neither pays off, flip a coin]

Likewise, L transitions to R with probability:

q(1-p) + ½ pq + ½ (1-p)(1-q)

The matrix of transition probabilities is:

	L	R
L	½ (1+p-q)	½ (1-p+q)
R	½ (1+p-q)	½ (1-p+q)

This is an ergodic Markov chain. No matter where you start, you get to an invariant probability distribution immediately:

Pr(L) = ½ (1+p-q), Pr(R) = ½ (1- p + q).

If, for instance, R pays off 90% of the time and L 50%, probe and adjust plays R 70% of the time. Probe-and-adjust favors the higher paying bandit, but does not learn optimal play.

In the special case where p + q = 1, the invariant distribution is:

Pr(L) = p, Pr(R) = q.

In the long run, the machine is played with the probability of its payoff. This is some version of "probability matching".

3. Signaling Games

Sender-Receiver signaling games were first introduced by David Lewis in *Convention*. There are two players, a sender and a receiver. The sender observes a situation, which nature chooses at random. She chooses a signal, conditional on the situation observed. The receiver observes the signal, and chooses an act, conditional on the signal observed. Payoffs for sender and receiver are determined by the combination of signal and situation. Signals are cost-free and sender and receiver have pure common interest. In the simplest case, there is an act that is "right" for each situation in that if that act is done in that situation both sender and receiver get a payoff of 1, and otherwise they get a payoff of 0. For simplicity, we can index that situations and acts such that the joint payoff on 1 occurs just in case act A_i is done in situation S_i. (Situations are often called "states", but we reserve this term for the state of a Markov chain.)

Lewis calls an equilibrium in which players always receive a payoff of 1 a *signaling system equilibrium.* There are other equilibria. If the sender sends signals with probabilities independent of the situation and the receiver chooses acts independent of the signals, we have a *total pooling* equilibrium. All the situations are pooled, in that the signal sent carries no information about the situation. It is an equilibrium in that neither sender nor receiver can improve her payoff by changing her behavior. If the are more than two situations, there may be *partial pooling equilibria* in which some but not all states are pooled. For example, suppose there are 3 situations, 3 signals and 3 acts, and that the sender sends:

signal 1 in situations 1 and 2,
signals 2 and 3 in some proportion in situation 3

and the receiver does:

acts 1 and 2 in some proportion for signal 1
act 3 for signals 2 and 3

This is a partial pooling equilibrium in which situations 1 and 2 are pooled.

Many generalizations and variations are of interest, but here we concentrate on the basic signaling game.

4. Learning to Signal with Reinforcement Learning: The simplest case.

In applying reinforcement learning to signaling games we do not reinforce whole strategies. After all, part of the point of the low rationality approach is that the agents involved may not be thinking strategically at all. Rather we think of just reinforcing responses to stimuli. A sender observes a situation as a stimulus and responds by sending a signal. For each situation, we equip the sender with an urn, the colored balls corresponding to the signal to send in that situation. And the receiver observes the signal as a stimulus. So for each signal, we equip the receiver with an urn, the colored balls corresponding to the act to take upon receiving that signal.

Consider the simplest Lewis signaling game in which nature flips a fair coin to choose one of two situations, the sender observes the situation and chooses between two signals and the receiver observes the signals and chooses between two acts. Sender and receiver use reinforcement of stimuli. There are now 4 interacting reinforcement processes – two sender's urns and two receiver's urns.

An analysis of these interacting reinforcement processes shows that sender and receiver always learn to signal with probability one. (Argiento, Pemantle, Skyrms and Volkov, 2009) Here is a sketch.

There are 8 quantities to keep track of: the numbers of the two types of ball in each of the four urns. But because the receiver is reinforced just when the sender is, and to the same extent, the numbers of balls in the sender's urns contain all the relevant information about the state of the process. Normalizing by dividing by the total number of balls in both sender's urns, we get four quantities that live in a tetrahedron. The mean

field dynamics is calculated. A quantity is found that the dynamics always increases, ruling out cycles. The reinforcement learning process must then converge to a rest point of this dynamics. The rest points are shown in figure 1..

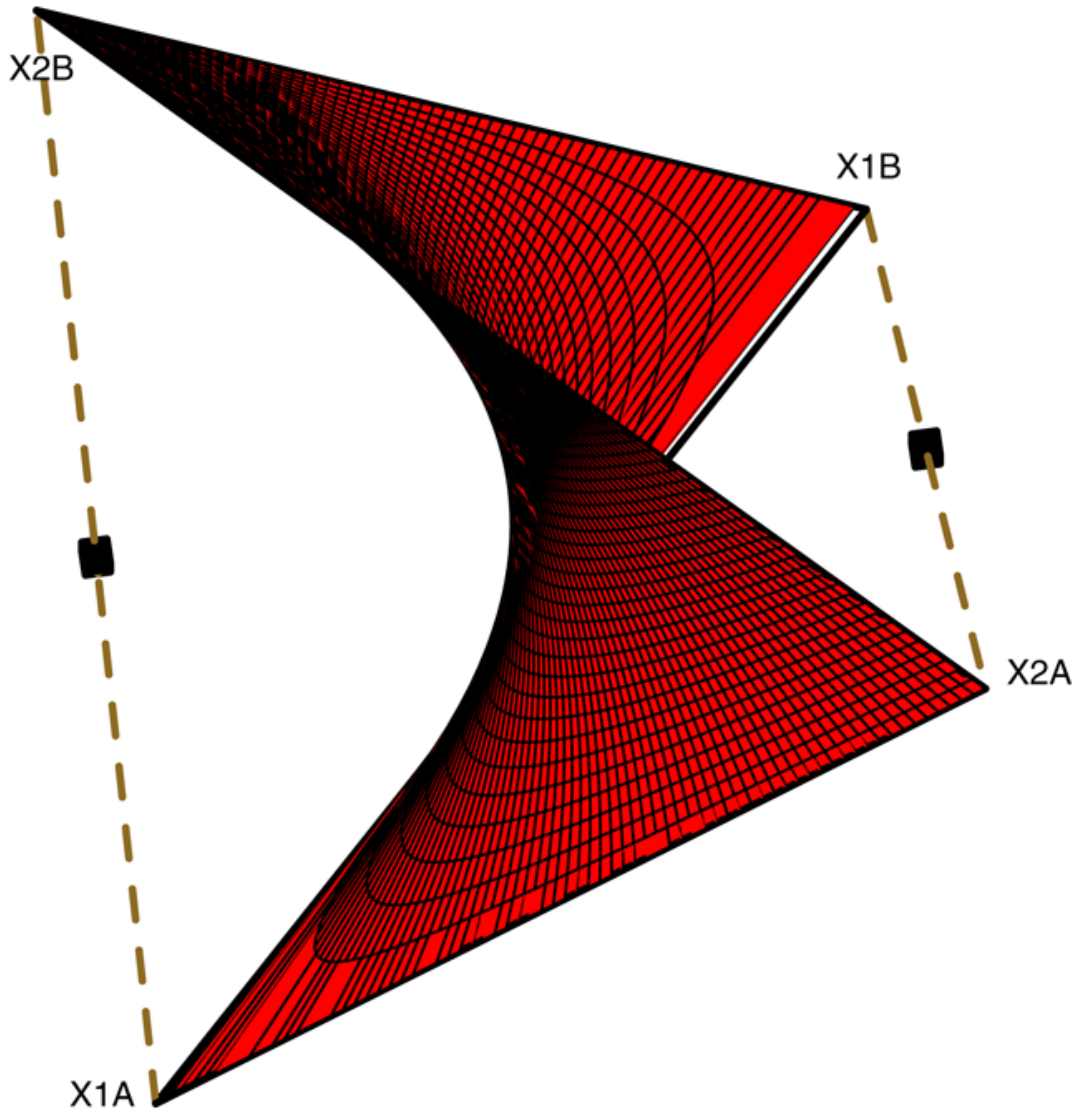

Figure 1: Rest points of the mean-field dynamics for reinforcement learning.

The two signaling systems are shown as square dots. At a signaling system players are always reinforced. One signaling system ends up with the urn for situation 1 full of signal A balls and the urn for situation 2 full of signal B balls; the other has urn 1 full of signal B balls and urn 2 full of signal A balls. The curved surface consists of pooling equilibria, where the probability of signals is independent of the situation. On one side of the

surface, the mean field dynamics leads to one signaling system, on the other to the second signaling system. It is then shown that the probability that reinforcement learning converges to a point on the surface is zero. (The argument requires a separate treatment of the parts of the surface in the interior if the tetrahedron and the parts on the boundary.) The rest point that are left are the signaling systems. With probability one, reinforcement learners learn to signal.

5. Learning to Signal with Probe and Adjust: The simplest case.

Consider probe and adjust learning applied to the same simple signaling game. As before we assume that each sender treats each situation as a separate choice stimulus. Each receiver treats each signal received as a separate choice stimulus. In order to apply probe and adjust, individuals must not remember not just what happened last time, but what happened last time for each stimulus. So the sender remembers what he did the last time he was confronted with situation 1 and the payoff he received and likewise for situation 2. The receiver remembers what he did last time and payoff received when observing signal one and likewise for signal 2. Memory requirements are still quite modest.

Most of the time agents just repeat what they did last time for the same stimulus. Every once and a while an agent probes and adjusts. We assume here that only one agent probes and adjusts at a time. The other agent keeps doing the same thing during the probe-and-adjust process. Now we can again find an embedded Markov chain which consists of pre and post probe-and-adjust transitions. Here the *state of the system* of both players consists of a pair: a map from situations to signals for the sender and a map from signals to acts for the receiver. This is just the *memory* of what they did last given the respective stimuli. (Note that this system state enables us to calculate the payoffs that they got last time they did something.)

There are 4 possible sender configurations and 4 receiver configurations, so there are 16 possible states of the system, as shown in figure 2.

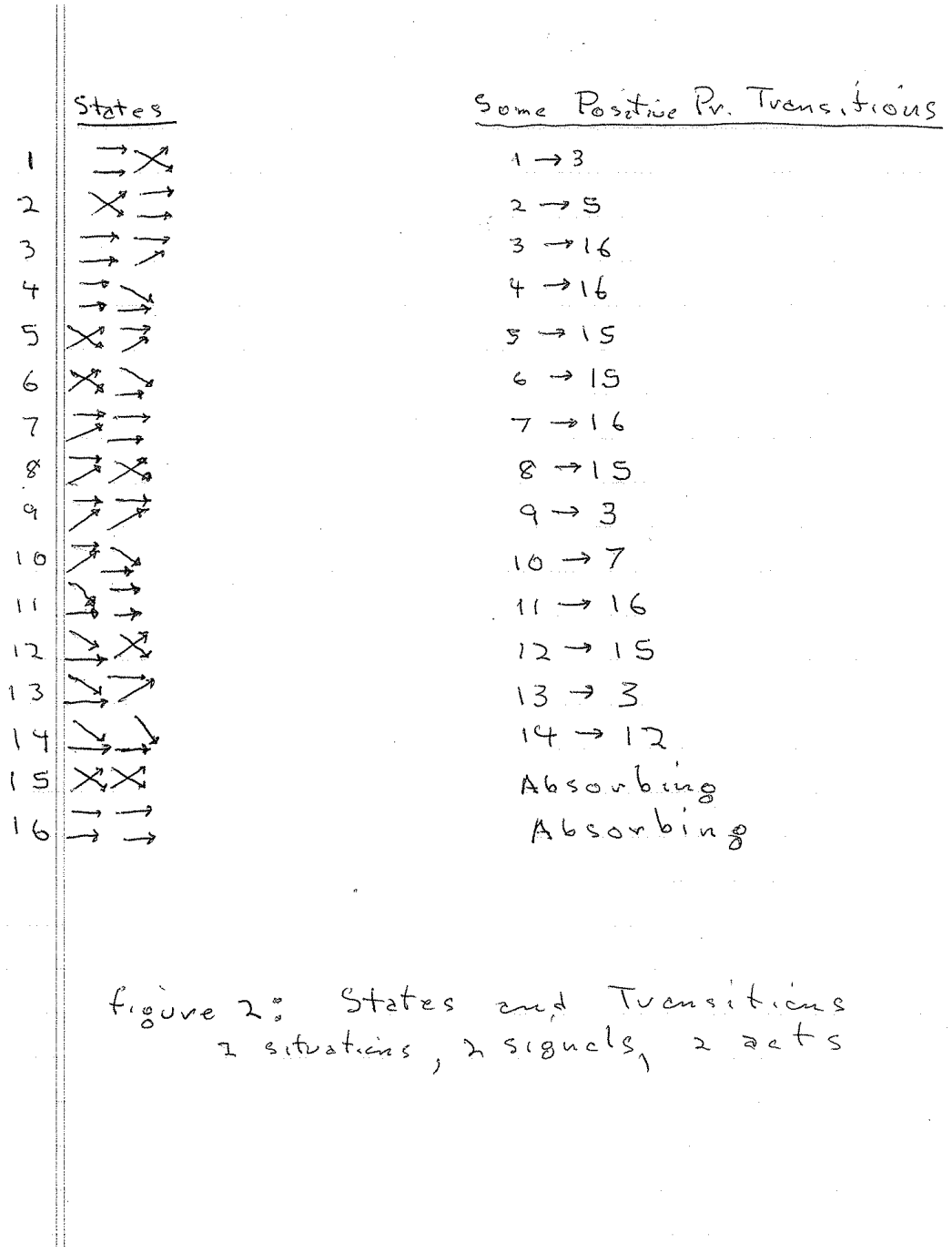

figure 2: States and Transitions
2 situations, 2 signals, 2 acts

v

(figure 2 here)

The transition probabilities are calculated as follows. Nature chooses the sender or receiver to probe by flipping a fair coin. Nature chooses a situation by flipping a fair coin. If sender is chosen sender probes a new signal for the situation chosen. If receiver is chosen, sender sends the old signal for the situation and receiver probes a new act for the signal.Then if the probe gave a higher payoff the new configuration is adopted; if it gave

the same payoff the new configuration is adopted with probability ½; if it gave a lesser payoff the system remains in the original state.

This Markov chain is no longer ergodic, like that for the bandit problem. States 15 and 16, which correspond to the two signaling systems, are *absorbing states*. Once entered the Markov chain will not leave them. Any probe will get a smaller payoff, and the agents will adjust back to the original state. Furthermore, as we will see, they are the only absorbing states. If the agents enter one of these states, we will say that they have *learned to signal*. This means that they will continue in the signaling system except for occasional fruitless probes that will lead them to return to it. Figure 2 shows some (not all) positive probability transitions between states.

States 15 and 16 are signaling systems. States 3, 4, 5, 6, 7, 8, 11 and 12 can move to a signaling system with one probe. Each of the other states can move to these with one probe. From any state there is a (short) positive probability path to a signaling system. It follows that *with probability one, probe and adjust learns to signal.* We will investigate how this logic generalizes.

6. *Reinforcement:* N Equiprobable Situations, N Signals, N Acts

Suppose we keep everything the same except increasing the numbers of situations, signals and acts. Situations are still assumed to be equiprobable and the number of signals and acts matches the number of situations. We know that there is now a new class of equilibria, the partial pooling equilibria. The question is what significance, if any, they have for reinforcement learners. A full analysis of this situation is not quite yet available, but it is known (Hu 2010, Hu et. al. in preparation) that:

(1) Convergence to total pooling has probability zero.
(2) Convergence to partial pooling has positive probability.

Learning to signal perfectly is no longer guaranteed. Learning to signal imperfectly is.

Extensive numerical simulations show that the extent of convergence to partial pooling is not a negligible outcome. Starting from an initial condition of one ball of each color in each urn, Barrett (2007) finds numerical convergence to partial pooling after 10^6 iterations of learning in 10^3 trials as follows:

N	Partial Pooling
3	9.6 %
4	21.9 %
8	59.4 %

7. *Probe and Adjust*: N Situations, N Signals, N Acts.

As in the simplest case, when we consider transitions between pre-probe-adjust states and post-probe-adjust states, we have an embedded Markov chain. The state of the system for this chain consists of a record of what was done last in each situation by the sender and last for each signal by the receiver. It is thus a pair of sender and receiver functions, <f, g> These are functions since each only remembers what she last did in the each decision situation. These need not be one-to-one, as the sender may have sent the same signals in multiple situations and the receiver may have done the same act on receiving multiple signals.

We show that the analysis of the simplest game generalizes. Specifically:

(1) Signaling systems are absorbing states of the Markov chain.
(2) Signaling Systems are the unique absorbing states.
(3) From any state, there is a positive probability path to a signaling system.

(1) Starting from a signaling system every *probe* changes a payoff from 1 to 0. Then then *adjust* returns to the signaling system. Signaling systems are absorbing states of the embedded Markov chain.

(2) If a state is not a signaling system, some probe either gives the same payoff or a lesser one. Thus some probe leads away with positive probability.

(3) Here is an algorithm that generates a positive probability path from any state of the system to an absorbing state. (This algorithm generates the transitions shown in figure 2 for the simplest signaling game.)

Start with S_1. The composition of sender and receiver functions $g(f(S_1)$ map it to an act. If it is A_1, move on. If it is not A_1, nature chooses the situation and the receiver to probe. Receiver probes A_1, and adjusts to choose A^1 for that signal since the probe moved payoff from 0 to 1.

Continue as follows:

Consider S_n. If sender maps it to a signal that does not yet appear on the path, proceed at above. The composition of sender and receiver functions $g(f(S_n)$ maps

it to an act. If the act is A_n, move on. If it is not A_n, nature chooses the situation and the receiver to probe. Receiver probes A_n, and adjusts to choose A_n for that signal since the probe moved payoff from 0 to 1.

If sender maps it to a signal already visited on this path [$f(S_1)$ --- $f(Sn_{-1})$] then nature chooses the situation, sender probes an unused signal [not now in the range of f]. There must be one since in this case more than one signal is mapped to the same situation. Previous payoff must have been a 0, since the old signal led to A_j ($j<n$). so adjust sticks with the probe with positive probability.

If $g(f(S_n) = A_n$ move on. Otherwise Nature chooses receiver to probe, receiver probes A_n, and adjusts by keeping $g(f(S_n)=A_n$, since the probe changes zero to 1

Next S_i.

It follows that the Markov chain always reaches an absorbing state.

One might ask how individuals begin to learn. In defining the states of the Markov chain, we assumed there was always a history. Shouldn't we redo the Markov analysis allowing partial functions for sender and receiver histories, and allowing random choices when a new stimulus with no prior history is reached? The answer is that this doesn't change anything. There is still a positive probability path from any state to a signaling system.

Agents always learn to signal perfectly. Here *probe and adjust* learning always achieves optimal signaling in a setting where reinforcement learning sometimes does not. One might ask how this result generalizes. First we look at cases where the number of signals is different from the number of states and acts, keeping states equiprobable. Then we relax the condition of equiprobable states.

9. *Extra Signals*: N states, M signals, N acts. (M>N)

If we have excess signals, reinforcement learning can still fall into partial pooling equilibria. (Hu 2010) For probe and adjust, the picture has changed. The embedded Markov chain generated by probe and adjust no longer has any absorbing states. Efficient states - those where g(f(s)) is the identity - of the system consist of an N by N by N signaling system together with M-N signals that are not sent. These signals are mapped onto some acts by the receiver's memory. Suppose some unused signal is mapped onto A_i. If the sender probes by sending the unused signal instead of the signal he usually sends in S_i, then the probe does not change his payoff. Then with some probability he adjusts by retaining the new signal in place of the old. (nb. a receiver cannot probe changing an unused signal, since a receiver is not probing a new function, g,

but only what to do if confronted with a signal. If a signal is unused, a receiver is not confronted with it.)

For example, in figure 3, probe and adjust can lead from state A to B and state B to A but not outside the set. Check all other possible probes. They don't lead anywhere. This is an absorbing set (an ergodic set). Within the set, signals 2 and 3 can be thought of as sequential synonyms. One is used for situation 2 for a while and then the other is. After the system has been absorbed into this set, there is a long term invariant distribution within the set. Signals 2 and 3 are each used half the time for situation 2. There are lots of such ergodic efficient sets.

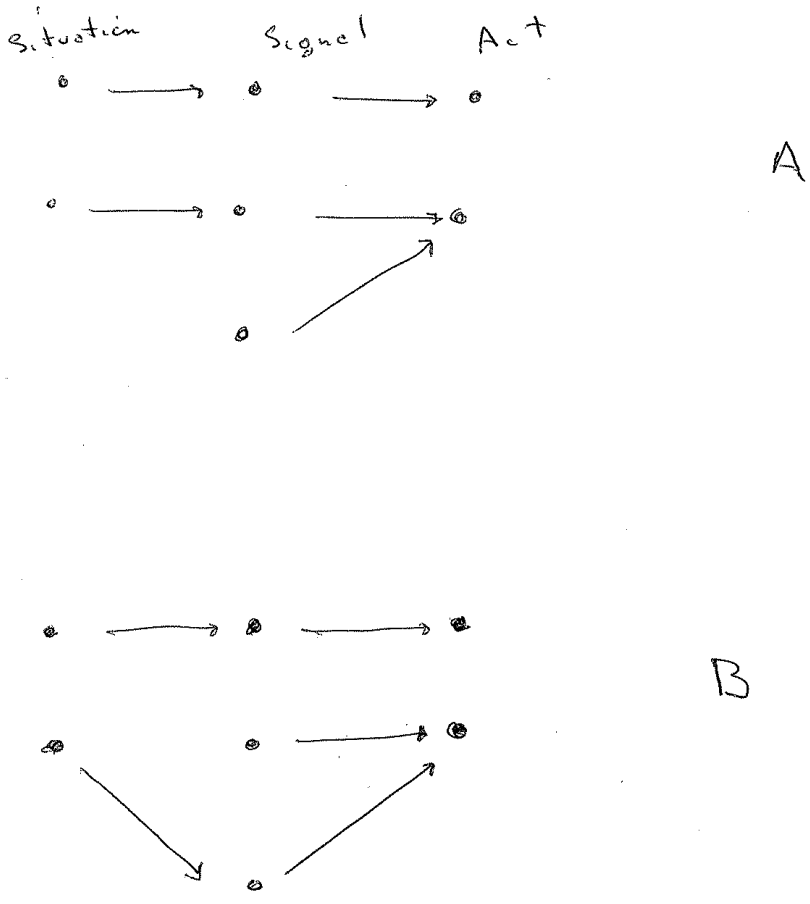

Fig 3

(figure 3 here)

It is still true that from any state there is a positive probability path to an efficient ergodic set. In fact, the same algorithm as before works for the same reasons. Applying the algorithm until one runs out of states give a path to a member of an efficient ergodic set. Probe and adjust still learns to signal.

***Too few signals*: N states, M signals, N acts. (M<N)**

In this case, partial pooling is the best that can be done. There are not enough signals for a signaling system to map each state onto the appropriate act. The best average payoff achievable is (M/N), gotten when for each signal, s, g(s) in $f^{-1}(s)$. The structure of these efficient signaling equilibria is investigated in Donaldson et. at. (2007). Consider the case of M=2, N=3. In the efficient equilibria the sender pools 2 situations. There are 3 choices of situations to pool. For each choice there is a choice which signal to assign to the pooled states. So there are 6 sender's strategies that are components of efficient equilibria. On the receiver's side, for such a sender's strategy there are two situations pooled, so it makes no payoff difference if the receiver does the right act for one or for the other. Thus 2 receiver's strategies pair with each of the 6 sender's strategies to make 12 optimal equilibria. This is shown if figure 4.

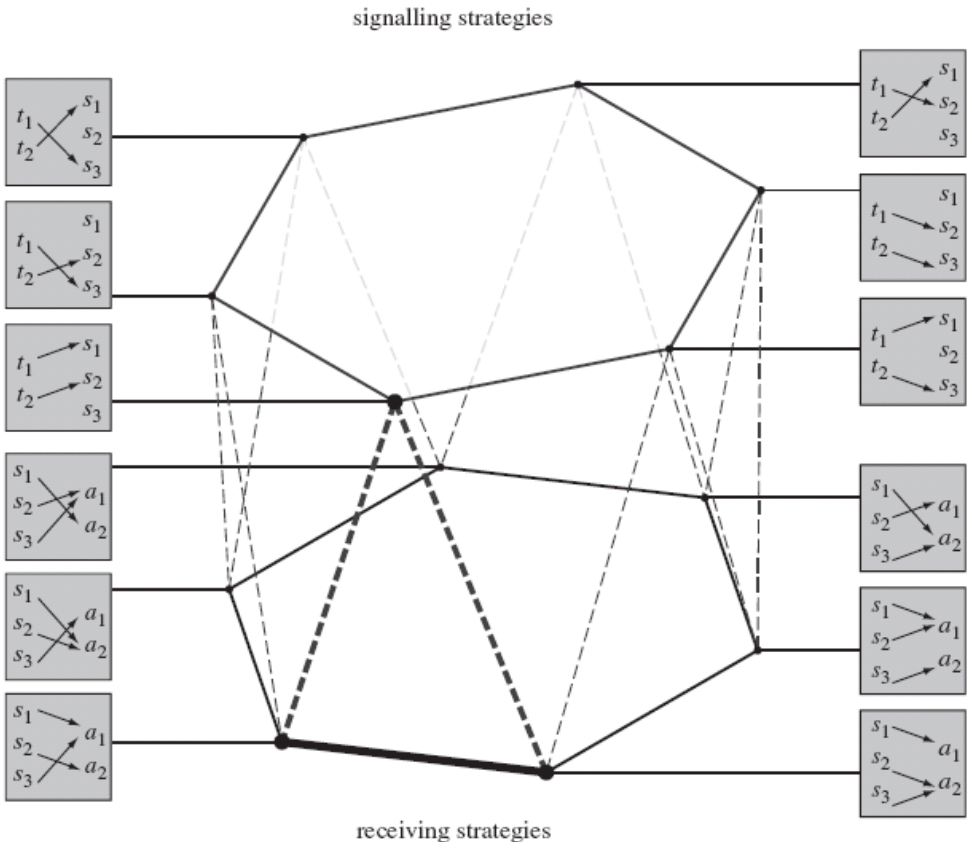

Figure 4: From Donaldson, et. al.

Probe and adjust can move from one of these equilibria to another. Adopting the notation of the figure (s for situation, t for signal, a for act), suppose that we start at:

	Sender	**Receiver**
E1:	s1 => t1	t1 => a1
	s2 => t2	t2 => a2
	s3 => t2	

The sender pools situations 2 and 3. Suppose nature chooses s3, sender sends t2 and receiver probes a3. Receiver gets a payoff of 1 which, at worst, ties previous payoff for doing a2 upon seeing t2. So with positive probability receiver switches, taking us to:

	Sender	**Receiver**
E2:	s1 => t1	t1 => a1
	s2 => t2	t2 => a3
	s3 => t2	

This step is reversible; a probe can take us back to E1.

Now the receiver never does a2, so it doesn't matter if the sender pools s2 with s1 or with s3. Suppose nature chooses s2, and sender probes t1. This leads to 0 payoff which matches sender's previous payoff in s2, so with positive probability sender switches, leading to:

	Sender	**Receiver**
E3:	s1 => t1	t1 => a1
	s2 => t1	t2 => a3
	s3 => t2	

(This step is also reversible.) But now sender is pooling s1 and s2, so by the same logic as before a receiver's probe can lead to:

	Sender	**Receiver**
E4:	s1 => t1	t1 => a2
	s2 => t1	t2 => a3
	s3 => t2	

And now receiver never does a1, in turn setting up a sender's probe that can lead to:

	Sender	**Receiver**
E5:	s1 => t2	t1 => a2
	s2 => t1	t2 => a3
	s3 => t2	

The process continues through a cycle of the 12 efficient equilibria. All these efficient states form one ergodic set. As probe and adjust moves through this set, signals shift their meaning.

To show that there is a positive probability path from any state to a state in the efficient ergodic set, use the same algorithm as before. When you run out of signals you are done.

Thus, in this case also, probe and adjust leads to an efficient signaling system.

Situations with Unequal Probabilities

So far, states have been supposed to have equal probabilities. There is to date no rigorous analysis of reinforcement learning in situations with unequal probabilities. But preliminary analyses and computer simulations point to the conclusion that in this case reinforcement learning can lean to total pooling equilibria. Consider the case of two situations, signals, and acts in which situation 1 is highly probable. Then a total pooling equilibrium in which the receiver just does act 1 and ignores the signals, and the sender ignores the state and always sends the same signal, is not so implausible. Players usually get paid off without bothering much with signaling. Simulations show reinforcement learning sometimes converging to a signaling system, sometimes to total pooling.

How does *Probe and Adjust* do with unequal probabilities? Notice that none of our foregoing analysis made use of the assumption of equal situation probabilities. We only need that situation probabilities are all positive in order to construct the positive probability paths leading to absorbing sets or states. Probe and adjust dynamics learns to signal perfectly in N situation, M signal, N act signaling games when there are enough signals (M >=N).

In the case where there are too few signals, M<N efficiency imposes an extra requirement. Since there are not enough signals, states have to be pooled. In an efficient configuration, the highest probability states must be serviced in a way that maximizes expected payoff.

For example, suppose that M=2, N=3, and states 1,2,3 have probabilities .2,.3,.5 respectively. One efficient configuration will have the sender map states 1 and 2 onto one signal that the receiver maps to state 2, and have the sender map state 3 onto the other signal, which the receiver maps to act 3. Then the average payoff is .8. But Probe and Adjust may lead from this efficient state to and inefficient one, as shown in figure 5.

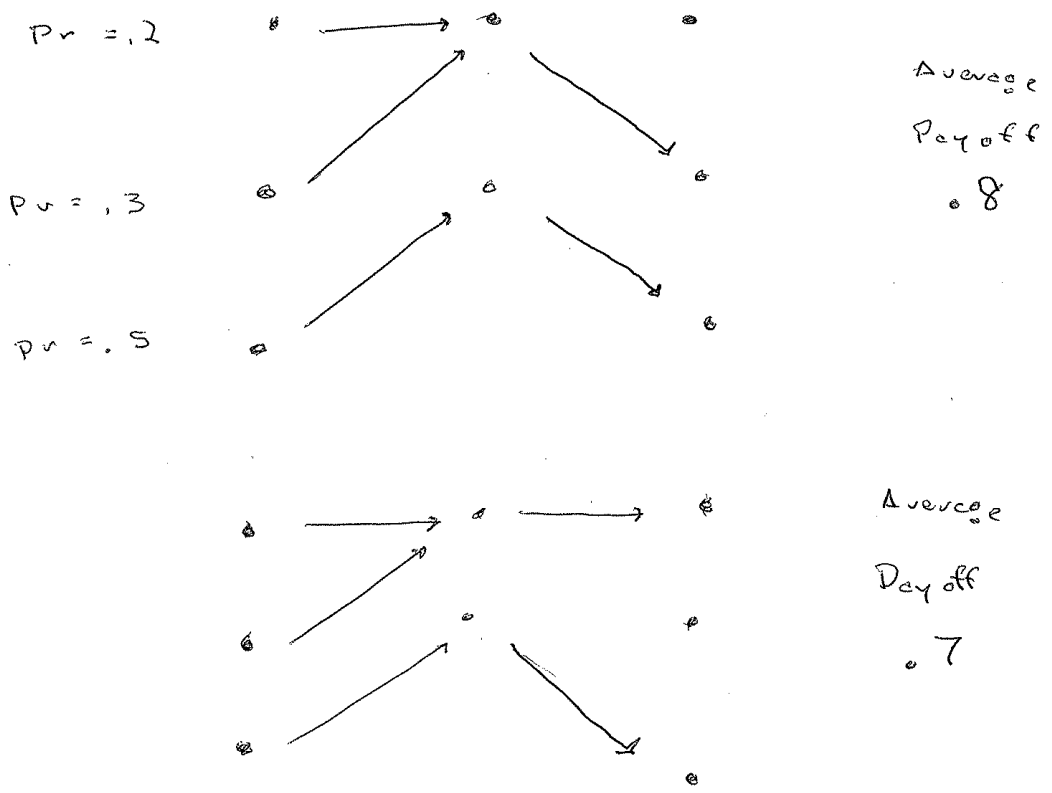

figure 5

(Figure 5 here)

In state 1 sender send signal 1 and receiver probes act 1. This gets a payoff of 1, which matches the receiver's memory of a payoff of 1 for doing act 2 on receiving signal 1. The tie is broken by a coin flip, so with some probability receiver stays with the probe. The receiver has no memory of the *frequencies* of payoffs – only of the magnitude of the last payoff. So the probe and adjust can move around the whole ergodic set of the preceding section.

This weakness of probe and adjust here is just the other side of the coin from its strength. The fact that absorption to an ergodic set is what allows it to so reliably learn to signal with situations of unequal probabilities.

Conclusions

Evidently our two kinds of trial and error learning have different strengths and weaknesses. Probe and adjust jumps deterministically and in signaling games locks onto successful acts. But in bandit problems it jumps too often and never learns to play the optimal bandit. Reinforcement learning gradually modifies act probabilities and is asymptotically efficient in bandit settings, but it can sometimes get stuck in suboptimal equilibria in some signaling games.

Various modifications can make these two kinds on trial and error more like one another. If we start with very small initial propensities in reinforcement learning, it tends to almost jump deterministically on a success. Simulations show that with very small initial propensities, reinforcement learners almost always learn to signal perfectly. Pushing this to the extreme, suppose we start out with the urns empty, with the rule that for empty urns you choose at random, but for non-empty urns you use Roth-Erev reinforcement. An initial reinforcement then causes sender and receiver to lock-in of the actions that produced the reinforcement. With this dynamics such learners *always learn to signal.* (See Barrett and Zollman 1999) But this modification deprives reinforcement of its exploration properties. In bandit problems, for instance, such a learner would always stick with the bandit that first paid off.

On the other hand, the behavior of *probe and adjust* in bandit problems can be improved if we endow the learner with a little more memory. She might remember the average payoffs from the last n trials, then probe n times and compare the two figures. Then, as before, she would switch if the probe is better, go back if the probe is worse, and flip a coin to break ties. For example, for p=.9 and q=.5, we saw that probe and adjust played the optimal bandit just 70% of the time – with *3-probes and adjust* this rises to 86%.

Evidently different kinds of trial and error learning can give quite different results. And their effectiveness depends both on the application and the definition of success. Versions of trial and error learning that work more generally are possible at the cost of some complication (Marden et. al. 2009, Young 2009).

If we ask whether we can learn to signal by trial and error, the answer is positive Both kinds of trial and error sometimes learn to signal in all signaling games and always learn to signal in the simplest signaling game. Probe and adjust always learns to signal in a wider variety of settings.

Related Work

The need for low rationality game theory is emphasized in Roth and Erev (1995), Erev and Roth(1998), Fudenberg and Levine (1998),and Young (2004) (2009). Equilibrium structure of signaling games with many or few signals is analyzed in Donaldson et. al. (2007). Two population replicator dynamics of a 2 situation, 2 signal, 2 act signaling game where the situations have unequal probabilities is analyzed, with and without mutation, in Hofbauer and Huttegger (2008). Blume et. al. (1998) and Blume et al (2001) present experimental evidence for the emergence of signaling by learning in a 2 situation, 2 signal, 2 act signaling game. Where subjects were only given own payoffs as feedback, experimental results were consistent with reinforcement learning.

References

Argiento, R. R. Pemantle, B. Skyrms and S. Volkov (2009) "Learning to Signal: Analysis of a Micro-Level Reinforcement Model" (2009) *Stochastic Processes and their Applications* 119: 373-319.

Barrett, J. A. (2007) "Dynamic Partitioning and the Conventionality of Kinds" *Philosophy of Science* 74:526-546.

Barrett, J. A. and K. Zollman (2009) "The Role of Foregetting in the Evolution and Learning of Language" *Journal of Experimental and Theoretical Artificial Intelligence* 21: 293-309.

Beggs, A. (2005) "On the Convergence of Reinforcement Learning" *Journal of Economic Theory* 122:1-36.

Blume, A., D. V. DeJong, Y-G Kim and G. B. Sprinkle (1998) . "Experimental Evidence on the Evolution of the Meaning of Messages in Sender-Receiver Games." *American Economic Review* 88:1323-340.

Blume, A., D. V. DeJong, G. R. Neumann and N. E. Savin (2001) "Learning and Communication in Sender-Receiver Games:An Econometric Investigation" *Journal of Applied Econometrics* 17:225-247.

Donaldson, M., M. Lachmann and C.T. Bergstrom (2007) "The Evolution of Functionally Referential meaning in a Structured World" *Journal of Theoretical Biology* 246: 225-233.

Erev, I. and A. Roth (1998) "Predicting How People Play Games: Reinforcement Learning in Games with Unique Mixed-Strategy Equilibria" *American Economic Review* 88: 848-881.

Fudenberg, D. and D. Levine (1998) *A Theory of Learning in Games.* Boston: MIT Press.

Herrnstein, R. J. (1970) "On the Law of Effect" *Journal of the Experimental Analysis of Behavior* 13:243-266.

Hofbauer, J. and S. Huttegger (2008) "Feasibility of Communication in Binary Signaling Games" *Journal of Theoretical Biology* 254: 843-849.

Hopkins, E. and M. Posch (2005) "Attainability of boundary points under Reinforcement Learning" *Game and Economic Behavior* 53: 110-125.

Hu, Y. (2010) *Essays on Random Processes with Reinforcement* PhD Thesis St. Anne's College, Oxford University.

Hu, Y., B. Skyrms and P. Tarrés "Reinforcement Learning in Signaling Games" (In preparation).

Huttegger, S. and B. Skyrms "Emergence of a Signaling Network with *Probe and Adjust*" (forthcoming) In *Cooperation, Complexity, and Signaling* Ed. B. Calcott, R. Joyce and K. Sterelney. Cambridge, Mass.: MIT Press.

Kimbrough, S. O. and F. H. Murphy (2009) "Learning to Collude Tacitly on Production Levels by Oligopolistic Agents" *Computational Economics* 33:47-78.

Lewis, D. K. (1969) *Convention: A Philosophical Study*. Cambridge, Mass.: Harvard University Press.

Marden, J. P., H. P. Young, G. Arslan and J. S. Shamma. "Payoff-based dynamics for Multiplayer Weakly Acyclic Games." *SIAM Journal on Control and Optimization* 48: 373-396, 2009.

Pemantle, R. (2007) "A Survey of Random Processes with Reinforcement" *Probability Surveys* 4: 1-79.

Roth, A. and I. Erev (1995) “Learning in Extensive Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term” *Games and Economic Behavior* 8: 164-212.

Skyrms, B. (2010) *Signals: Evolution, Learning and Information.* London and New York: Oxford University Press.

Young, H. P. (2004) *Strategic Learning and its Limits* London and New York: Oxford University Press.

Young, H. P. (2009) “Learning by Trial and Error.” *Games and Economic Behavior* 65: 626-643.