Preferential Attachment and the Search for Successful Theories

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

February 26, 2012

Abstract

Multiarm bandit problems have been used to model the selection of competing scientific theories by boundedly rational agents. In this paper, I define a *variable-arm* bandit problem, which allows the set of scientific theories to vary over time. I show that Roth-Erev reinforcement learning, which solves multiarm bandit problems in the limit, cannot solve this problem in a reasonable time. However, social learning via preferential attachment combined with individual reinforcement learning which discounts the past, does.

1. Introduction.

When faced with competing scientific theories of unknown value, how should rational agents decide which one to use? Following Zollman (2007, 2010), we can think of this in terms of what is known in Economics as a multiarm bandit problem. One faces a slot machine with N arms attached, each arm having a different and unknown probability of winning. The arms represent the competing theories, and the probability of winning corresponds to the probability of making a correct prediction. If you play the bandit repeatedly, what should you do to maximise your expected earnings over the long run?

Zollman assumes the population to be composed of Bayesians, whose prior probability function is given by a beta distribution and who update via sampling. In addition, he treats the agents as situated within a socially structured environment, so that an agent learns about the success and theory used by her neighbours. In this framework, Zollman derives a nicely counterintuitive result

that, sometimes, more information is harmful with respect to identifying the better theory.

In what follows, I extend the multiarm bandit model of competing scientific theories in two directions, each one attempting to address a limiting assumption of Zollman's model. The first concerns the number of scientific theories over which an agent must deliberate. Instead of taking this number to be fixed, the model developed in section 2 allows agents to introduce and eliminate scientific theories from consideration. The second concerns the structure and nature of the underlying social network. Instead of assuming the structure to be fixed, I allow agents to *preferentially attach* to others based on their past success. Each agent is thus free to stop learning from someone whose past performance has been poor in order to start learning from someone else with a better track record.

2. Reinforcement learning and variable-arm bandits.

Instead of assuming Bayesian rational agents, let us assume that agents use reinforcement learning. In particular, they use reinforcement learning of the kind introduced by Roth and Erev (1995) to explain the behaviour of experimental subjects. As Skyrms (2010) notes, Roth-Erev reinforcement learning in a multiarm bandit problem is equivalent to a Pólya urn model with probabilistic reinforcement.[1] One nice feature of Roth-Erev reinforcement learning is that, in the limit, it converges upon playing the arm with the greatest expected payoff (see Beggs, 2005). Or, to put the point in the language of scientific theories: Roth-Erev reinforcement learning ensures that the most empirically adequate theory will be adopted in the limit.[2]

Although most discussions of multiarm bandit problems assume a fixed number of arms, not all do. Whittle (1981, pg. 284) observes that "[i]n medicine, agricultural and technological applications one can expect that new projects will be added as time goes on, as new compounds, technical possibilities, etc., become available for investigation"; a better model would allow the number of arms, here representing projects, to vary as a function of time. When the number of arms increases over time, we have what Whittle called an "arm-acquiring bandit".

[1] A Pólya urn model consists of an urn containing an initial assortment of coloured balls. At each time step a ball is sampled with replacement, and then another ball of that colour is added to the urn. Now suppose that the ball colours represent handles of the multiarm bandit, and when a ball is sampled from the urn, with replacement, the player pulls the corresponding bandit arm. If, and only if, a 'win' occurs, does the player add another ball of that colour to the urn. This is the probabilistic reinforcement.

[2] This formal result nicely aligns with Peirce's statement that "[t]he opinion which is fated to be ultimately agreed to by all who investigate, is what we mean by the truth" (Peirce, 1992, pg. 189).

In contrast to multiarm bandits, arm-acquiring bandits provide a better model of choosing among competing scientific theories because new theories can be introduced along the way. In doing so, the learning problem becomes more difficult. Whereas the original multiarm bandit problem required agents to negotiate between *exploration* and *exploitation*, the arm-acquiring bandit problem requires agents to negotiate the tradeoff between exploration, exploitation, and *innovation*.

Adding the extra dimension of innovation to the problem would not matter so much if it was guaranteed that new theories (new bandit arms) always performed better — even if only marginally — than the current collection of theories (bandit arms). There is no reason to assume this is so. Although we like to think of science as progressing, with theories of ever-greater empirical adequacy, this overlooks the fact that, along the way, many candidate theories are discarded before they ever reach the stage of being seriously considered in the marketplace of ideas. Before failing better, one must first try again, and fail again.[3]

Whittle assumed an exogenous stochastic process determined whether a bandit acquired a new arm. We can endogenize this process, while still treating agents as reinforcement learners, using a modification of Roth-Erev reinforcement learning. In 1984, Fred Hoppe introduced what he called "Pólya-like urns" in order to study neutral evolution in biology. Skyrms (2010) used Hoppe-Pólya urns to model signal invention in sender-receiver games. Following in this tradition, we can use Hoppe-Pólya urns to model boundedly rational agents who not only choose between competing scientific theories, but who periodically invent new theories as well.

The simplest Hoppe-Pólya urn model begins with an urn containing only a single black ball known as the "mutator". Whenever the mutator is drawn, a ball of a new colour is inserted into the urn. If a coloured ball is drawn, that ball is returned to the urn along with another ball of the same colour. The key difference between this urn model and Pólya's urn is that new colours are introduced over time.

Consider the following model. Suppose that an agent begins with a Hoppe-Pólya urn containing a single black ball, a 0-arm bandit, and a big bag containing infinitely many arms. Each arm in the bag has a randomly assigned probability of winning, where this probability is drawn from some distribution[4] over $(0, 1)$, and the probability of winning cannot be discerned by inspecting the arm. At the beginning of each iteration, the agent reaches into the urn and draws a ball (with replacement). If the black ball is drawn — which always happens in the first round of play — the agent draws an arm from the bag and attaches it the

[3] Apologies to Beckett.

[4] See appendix A for a discussion of how probabilities are assigned to new arms.

bandit. The agent then pulls the arm. If a win occurs, the agent colour-codes the arm with a new colour, and adds a ball of that colour to the urn. If a win did not occur, the agent detaches the arm and returns it to the bag. Given this, after some time the urn will contain coloured balls in addition to the black ball. If a coloured ball is drawn from the urn, the agent simply pulls the appropriately coloured handle of the bandit and, if a win occurs, reinforces by adding one ball of that colour to the urn.

This gives a model of reinforcement learning for a multiarm bandit problem where the number of arms increases over time. It also provides a model of competing scientific theories where a boundedly rational agent tries to identify the most empirically adequate theory as the number of theories grows over time.

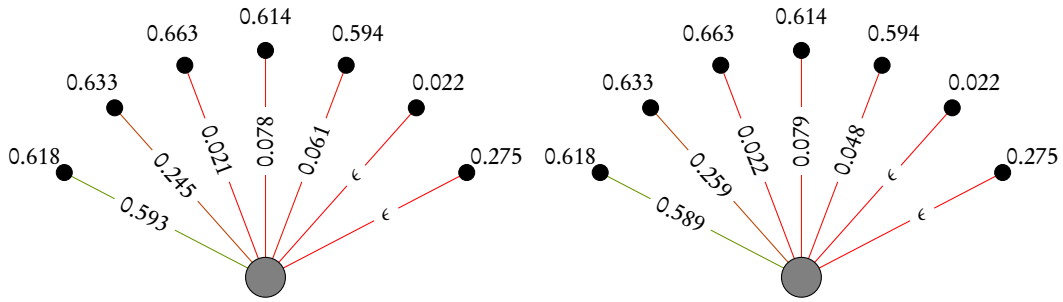
The first point to note is that, although the agent is cautious in attaching arms to the bandit (a new arm needs to win on the first pull to be kept), the arms will still vary in quality. The agent may have simply been lucky on the first pull — stopped clocks being right twice a day, and all of that. The agent still faces the task of identifying the best one to play.

The second point to note is that, in the limit, this model of arm-acquiring bandits will end up with infinitely many arms. (A proof is given in appendix B.) This also means that there will be arms whose probability of winning is arbitrarily close to 1. In the case where there are only a *fixed* number of arms, as noted earlier (Beggs, 2005), Roth-Erev reinforcement learning identifies the best arm to play. What happens here?

We can investigate this via simulation. Figure 1 illustrates a representative outcome for a single researcher faced with an arm-acquiring bandit problem. The probability assigned to new candidate arms was determined as described in appendix A using a Gamma distribution with parameters $\alpha = \frac{3}{2}$ and $\beta = 2$. After 10,000 iterations, the bandit has acquired seven arms. However, since the first arm the bandit acquired issued a win approximately 61.8% of the time and it was some time before a better arm was attached, reinforcement learning led the agent to favour an arm which later turned out to be suboptimal. Over time, the agent gradually started to learn to play the higher probability arms more frequently — but note how slowly the transition occurs! After two *billion* iterations only a very modest amount of probability has been shifted towards the better arms.

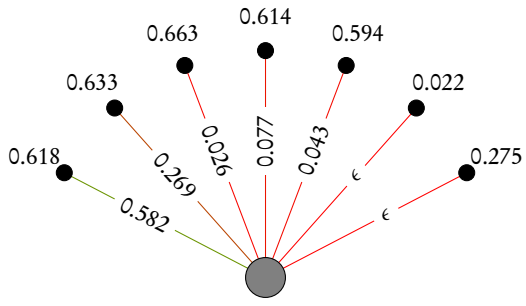
Simulations suggest that boundedly rational agents will gradually learn to favour higher probability arms, but this happens very slowly. Any significant convergence will not occur in a reasonable amount of time. What happens if we introduce social structure, allowing agents to learn from the discoveries of others?

Figure 1 Learning to use the most empirically adequate theory. The gray node represents the agent, the smaller black nodes the arms of the bandit. The number displayed above the arm nodes is the probability of that arm winning when pulled. Numbers displayed on the *edges* indicate the probability of the agent choosing to pull that arm by drawing a ball from the Hoppe-Pólya urn. Probabilities extremely close to zero are denoted by 'ϵ'.

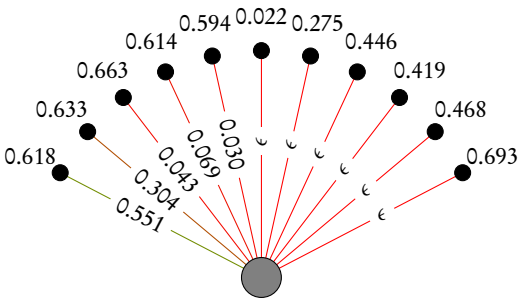


(a) 10,000 iterations

(b) 110,000 iterations



(c) 1,110,000 iterations



(d) 2,000,000,000 iterations

3. Preferential attachment and the social structure of science.

Quantitative analysis of publication databases provides a rich source of information regarding the social structure of science. Most typically one looks at the co-authorship graph, where nodes represent authors and two authors are connected by an edge if they have written a paper together. Barabási et al. (2002) examine "all relevant journals" in mathematics and neuroscience over the eight-year period from 1991 to 1998 and find that "the network is scale-free, and that the network evolution is governed by preferential attachment". Newman (2004) dissents slightly: examining bibliographical databases in biology, physics and mathematics, he concludes that the network's are not truly scale-free as the degree distribution does not follow a power-law. Investigating *international* scientific collaboration, Wagner and Leydesdorff (2005) find that, although the degree distribution does differ from a power-law, preferential attachment can still be used to explain the resulting network structure.

Regardless of whether real scientific networks are scale-free or not, let us assume that the underlying assortment mechanism is that of preferential attachment. The basic model of Barabási et al. (2002, pg. 602) assumes that scientific collaboration networks "evolve" as follows:

1. New researchers join the network at a constant rate, and link to researchers already present via preferential attachment.

2. Researchers already present in the network form new internal links via preferential attachment.

3. No age effects exist: once introduced, nodes continue to be present, initiating and receiving links.

If we are to apply this dynamic model to our social learning problem, we need to define the mechanism of preferential attachment.

Suppose we have an initial population consisting of N researchers. Each researcher begins with an arm-acquiring bandit with no arms, initially, and a Hoppe-Pólya urn containing just the mutator ball. Define the *efficiency* of a researcher to be the number of non-black balls in their urn divided by the number of iterations which have passed since she began experimenting.[5] A solo researcher's efficiency thus provides a measure for his or her overall frequency of making successful predictions. Whereas Barabási et al. (2002) used the *degree* of a node when choosing to whom one should attach, since we are modelling

[5]It is worth distinguishing "efficiency" from "accuracy" because a researcher with low probability arms on his bandit may nevertheless have a lucky streak in which a number of wins occurred.

researchers who aim at the truth, our agents preferentially attach with probability proportional to an agent's *efficiency*.

How do agents preferentially attach to another? The *reflection rate* r determines how often one agent might choose to visit, speak, or communicate with another. If a researcher X visits another researcher Y whose efficiency exceeds X's, then X will attach to Y. Attachment generates a one-way communication link where X observes Y and copies his reinforcement,[6] including adding copies of bandit arms, if necessary.[7] Once connected, an agent X will remain attached to Y as long as Y's efficiency continues to be greater than X's. If Y's efficiency drops below X's, the next time X reflects upon whether to visit another agent, he will break his connection and possibly attach to another.

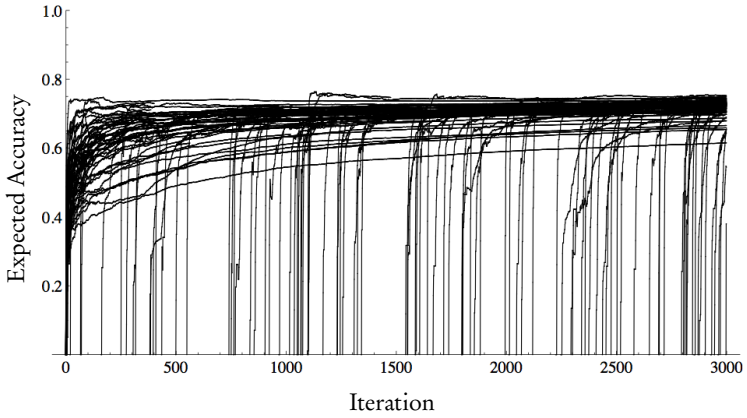
How much does preferential attachment help individuals identify the most empirically accurate theory? Figure 2 illustrates the outcome of 3,000 iterations of reinforcement learning, both individual and social, with preferential attachment. The initial population size was 50 agents, with new agents introduced with probability 0.03 and existing agents eliminated with probability 0.02. Note that the plot shows the time-evolution of the actual expected *accuracy* of a researcher: the sum of the probability of playing each arm (determined by the researcher's urn contents) multiplied by the probability of that arm winning when pulled. Although preferential attachment occurs by selection on an agent's *efficiency*, enough correlation exists between efficiency and accuracy to lead to an overall, albeit slight, increase in accuracy.

Preferential attachment helps because a better theory discovered by a single agent can spread throughout the population. Without a process of social learning, each agent must discover through her own hard labour incrementally improved theories. Given this, one may suspect that the real work is being done by the model of the context of discovery: the method of assigning probabilities to new bandit arms. As figure 3 illustrates, there is considerable truth to this. Modelling the context of discovery pessimistically, using an exponential distribution as

[6]This means the efficiency of an agent no longer corresponds to his past frequency of successful predictions. In fact, an agent's efficiency may well exceed 1. This does not matter, though, as agents choose whom to preferentially attach to by converting efficiencies into probabilities by normalization.

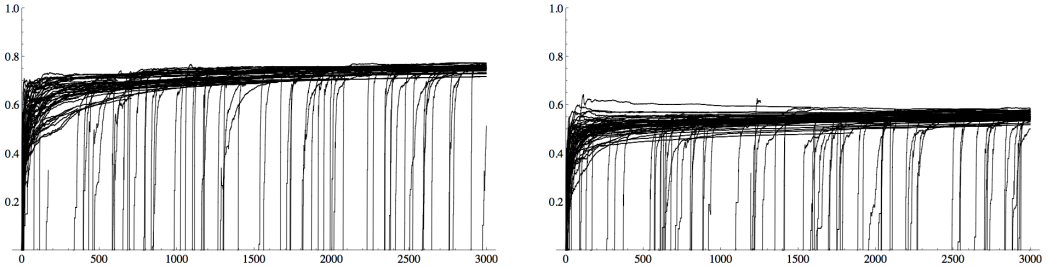
[7]Suppose that agent X is attached to agent Y, X's bandit arms are colour coded red, green, and blue, and that Y's bandit arms are coloured purple and magenta. Suppose that X reinforces by adding a blue ball and Y reinforces by adding a purple ball. It makes no sense for X to blindly add a purple ball to his urn unless he adds a purple arm to his bandit, so X creates a copy of Y's purple-coded bandit arm (with the same probability of winning) and adds it to his bandit. Although this sounds strange when phrased in terms of arm-acquiring bandits — how can you clone an arm without knowing the probability of it winning? — this makes perfect sense given that the arms represent *theories*. X can copy Y's *theory*, represented by the purple-coded arm, without knowing the probability of that theory making correct predictions.

Figure 2 The gradual adoption of empirically adequate theories. Each line plots the time-evolution of the expected accuracy of an agent. New lines sprouting from the x-axis correspond to new agents introduced into the population at that iteration. The probability assigned to new bandit arms was derived from a Gamma distribution with $\alpha = \frac{3}{2}$ and $\beta = 2$.



per appendix A, lowers the overall empirical accuracy achieved through social learning considerably.

Figure 3 The influence of the model of the context of discovery on population-wide increases in expected accuracy, using preferential attachment.



(a) The Gamma distribution ($\alpha = 1, \beta = 2$) (b) The exponential distribution ($\lambda = 0.3$)

We should not worry too much about this dependency. In the 8th edition of *A System of Logic*, Mill, commenting on an essay by Lord Macaulay, wrote: "I believe that if Newton had not lived, the world must have waited for the Newtonian philosophy until there had been another Newton or his equivalent. No ordinary man, and no succession of ordinary men, could have achieved it." If we adopt the Millian view that scientific progress depends upon great

theories introduced by a great person,[8] the real question becomes: how well does preferential attachment enable the spread of successful theories once introduced by a Newton-like figure? There is little point attempting to model *how* genius appears; rather, the interesting question is what happens *after* genius appears.

4. Newton, forgetting, and the usefulness of discounting.

Taking Mill seriously, consider a variant model: a population of reinforcement learners (modelled using Hoppe-Pólya urns and arm-acquiring bandits) invent scientific theories and use preferential attachment to learn from others. After a period of time, a new researcher appears with a theory that always yields correct predictions. That is, after a period of time we introduce a new researcher (let's call him "Newton") with an arm-acquiring bandit already containing an arm with probability 1 of winning. Will boundedly rational agents learn to benefit from the appearance of Newton?

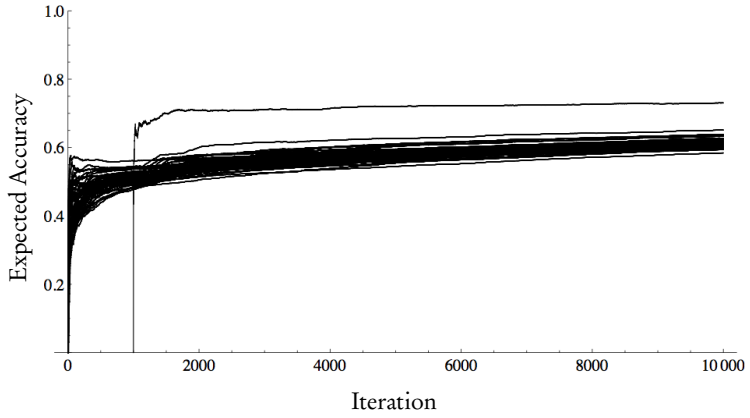
Simulations indicate the answer is often no. As figure 4 shows, the rounds of play before Newton appears (1,000 iterations, in this case) tend to generate significant lock-in to suboptimal theories among the existing researchers. Even though Newton has a theory which always yields successful predictions when used, the boundedly rational reinforcement learners are incapable of appreciating this point. Furthermore, because Newton himself doesn't appreciate just how good his theory is, he is subject to having his ideas polluted by preferentially attaching to others, taking onboard their suboptimal theories as competitors.[9]

One difference between figure 4 and earlier simulation results was that the population portrayed in figure 4 was static, save for the sudden appearance of Newton. Does it make a difference if we allow new researchers to enter the population, and existing ones to die out? Sometimes it does: although lock-in to suboptimal theories still occurs among existing researchers, new researchers are more likely to preferentially attach to Newton. Because new researchers are not locked in to suboptimal theories, social learning has a greater effect. Thus they are more likely to acquire and use Newton's highly successful theory, increasing the chance that that theory spreads throughout the population. As old researchers die out, the overall expected accuracy of the population thereby increases more rapidly than in static populations. Provided, of course, that the random culling

[8] Mill qualifies this point somewhat by saying that a succession of persons might have been able to reproduce the work of Newton, in stages. But even so, "the least of those steps required a man of great intellectual superiority." (This, and the previous quote, are taken from chapter 11 of Book 6 of *A System of Logic*.)

[9] Perhaps it is worth noting that the real Isaac Newton pursued studies in alchemy, heretical interpretations of Christianity, and the occult.

Figure 4 Reinforcement learning generates lock-in to suboptimal theories.



of individuals does not remove Newton before he has successfully promulgated his theory.

As a story about the spread and adoption of competing scientific theories, there are some nice structural parallels between this dynamic population account and one interpretation of Kuhn on competing paradigms. Old scientists, recalcitrant in their beliefs and incapable of appreciating the potential of the new theory, need to be replaced by the next generation before that theory's full potential may be realised. That said, as much as a boundedly rational agent like myself may be comforted by knowing that my children will benefit from better theories, even if I will not, it would be nice to have a model by which boundedly rational agents can generally benefit from social learning in the presence of Newton.

4.1. Forgetting

Lock-in to suboptimal outcomes occurs in other contexts where reinforcement learning is applied. Skyrms (2010) notes that, for Lewis sender-receiver games, reinforcement learning can fall into partial pooling equilibrium where inefficient communication occurs.[10] In a later paper, Alexander, Skyrms, and Zabell (2011) show that a certain model of reinforcement learning with *forgetting* avoids partial pooling traps by reducing the detrimental effects of lock-in.

Let us incorporate the model of forgetting of Alexander et al. (2011) as follows: periodically, a researcher examines his or her urn and selects a non-black colour at random. Having selected a colour, the researcher then discards one ball of that colour from the urn. If the last ball of a certain colour is discarded, then the researcher also removes the corresponding arm from her bandit. This changes

[10] A partial pooling equilibrium is one where the same signal is used for more than one state, leading to imperfect communication.

Whittle's model of arm-acquiring bandits to a model of *variable-arm* bandits, where the number of arms can fluctuate up and down over time. As before, this occurs in a context featuring both individual and social learning.

Figure 5 shows three outcomes which occur under reinforcement learning with forgetting. The first shows forgetting can hurt as well as help: Newton may foolishly discard his theory shortly after appearing. Second, even if Newton doesn't discard his theory, researchers who preferentially attach to Newton may themselves discard the theory. Why? Suppose a researcher whose bandit has three arms — red, green, and blue — preferentially attaches to Newton. Suppose that the researcher acquires Newton's theory by observation, assigning it the colour violet. Now suppose that the researcher preferentially attaches to someone other than Newton (whom does not have Newton's theory). Since forgetting first selects a colour by *type* to deinforce, there is a one in four chance that the researcher will select a violet-coloured ball to throw away. If substantial lock-in occurred to the red, green, and blue arms *before* attaching to Newton, this method of forgetting may well end up removing the violet arm from the bandit before any others. In the last case of figure 5(c), although clearly an improvement over that of figure 4, Newton's influence only improves half the population at a modest rate.

4.2. Discounting

A more realistic model of forgetting can be obtained from theories of *discounting* used in economics and finance, where one assumes that rational agents trade off present amounts against uncertain future gains. In addition, others (see Charles Wolf, 1970; Caplin and Leahy, 2004) have suggested that rational agents discount the *past* as well.

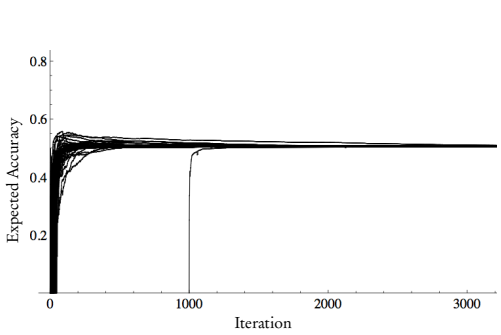
In a separate paper, I have shown (—, 2012) how urn-based models of reinforcement learning with *discounting*,[11] rather than forgetting, work surprisingly well at handling a variety of problems in dynamic sender-receiver games.[12] What happens if our boundedly rational reinforcement learners both discount the past and use preferential attachment to learn from others?

Figure 6 illustrates the typical simulation outcome when individual and social learning, with discounting, are exposed to a new, highly successful theory. The

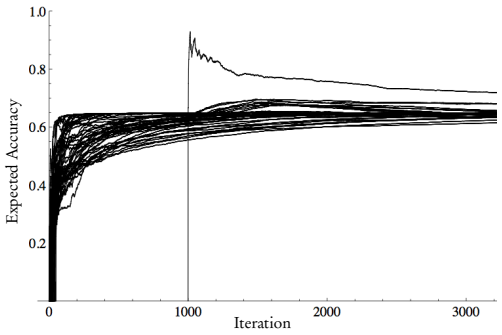
[11]Instead of thinking of colours in an urn having integral weights (i.e., a certain number of balls), think of the colours in the urn as having real-valued weights. Discounting occurs at the end of each iteraction by multiplying all of the real-valued weights, except the mutator, by a discount factor $0 < \beta < 1$. A cutoff threshold τ is assumed to exist, so that when a weight drops below τ that colour is removed from the urn and the corresponding bandit arm removed.

[12]The traditional Lewisian sender-receiver game assumes a fixed mapping between the action to perform and the state of the world. In a dynamic sender-receiver game, the correct action to perform, given the state of the world, can vary over time.

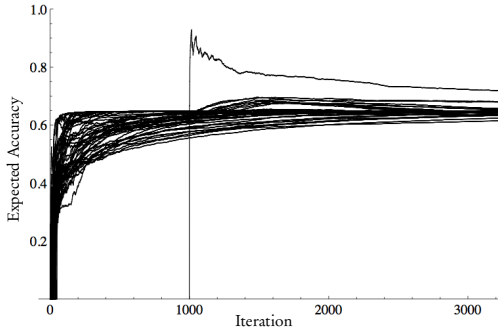
Figure 5 Preferential attachment and the spread of scientific theories. Three possible outcomes with reinforcement learning and forgetting, with Newton.



(a) Newton forgets his theory.



(b) Newton's theory doesn't spread.

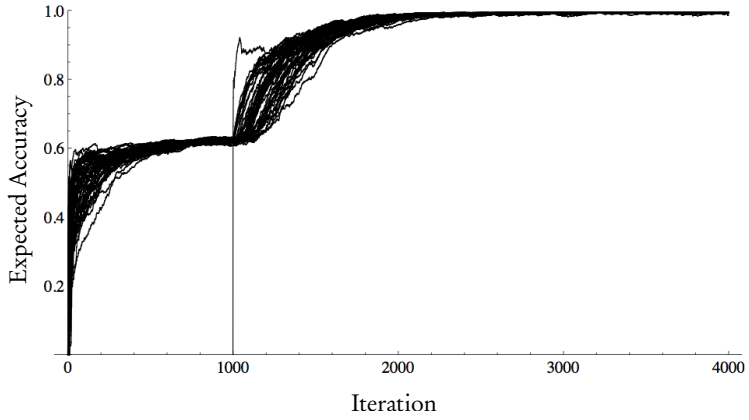


(c) Newton has some influence over people.

discount factor used was $\beta = 0.99$, ensuring that the maximum possible weight attained by reinforcement learning, for any given colour, was 100. When Newton appears after 1,000 iterations, his theory rapidly spreads through the population of 50 researchers until, by iteration 2,000, all researchers have an actual expected accuracy of 95% or greater.

There are three reasons this result is interesting: first, we do not need a dynamic population in which old agents are eliminated in order for the new theory to spread. Second, the adoption of the new theory occurs without needing to invoke theoretical virtues such as simplicity, elegance, beauty, explanatory strength, and so on. All that matters is brute-force empirical adequacy. Finally, the learning rule employed by the boundedly rational agents is an extraordinarily crude one, incapable of recognising success rates.

Figure 6 Reinforcement learning, with discounting, combined with preferential attachment, enables the rapid spread of successful scientific theories.



5. Conclusion.

Multiarm bandit problems have been used to model the spread of competing scientific theories in populations of boundedly rational agents. Although it has been shown that Roth-Erev reinforcement learning will correctly identify the optimal scientific theory (bandit arm) in the limit, the time required to move even approximately towards correct identification in variable-arm bandit problems can be very long indeed, even in the presence of social learning.

However, the combination of reinforcement and social learning (via preferential attachment) with discounting-the-past is seen to be remarkably successful at modelling the rapid spread of successful theories. Since there are good empirical reasons to think that people do, in fact, interact socially via preferential attachment and do, in fact, discount the past, these results tentatively suggest that these two behavioural traits may not be an evolutionary, or socially conditioned, accident. Rather, both may be important contributors to our epistemic success as boundedly rational agents.

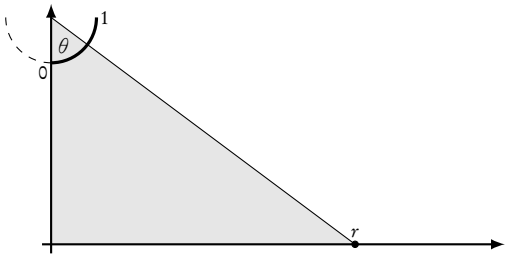
A. Assigning probabilities to new arms.

Any model of assigning probabilities to new bandit arms will, of necessity, be speculative, as it amounts to formalizing that black-box known as the "context of discovery". Nevertheless, there are still worse or better ways to proceed. To begin, consider the naïve approach which selects the probability p from the uniform distribution over $[0, 1]$. This is clearly undesirable, as it means that an agent who invents a new theory is just as likely to obtain one of high empirical

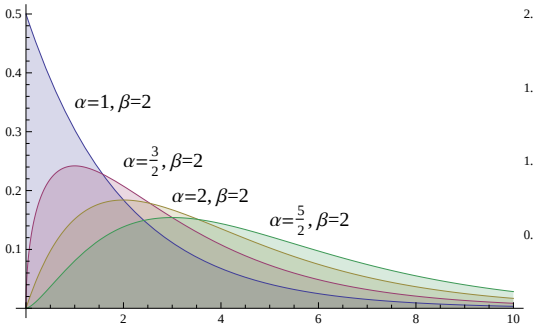
adequacy ($0.99 \leq p \leq 1$) as one of low empirical adequacy ($0 \leq p \leq 0.01$). Yet, as we all know too well, false theories are more easily obtained.[13]

Alternatively, consider the following proposal: suppose that a random real r is chosen from the half-open interval $[0, \infty)$. We then transform r into a probability via the map $\frac{2}{\pi}\arctan(r)$. This transformation simply determines the angle θ whose projection onto the positive reals would yield r, as per figure 7(a), and then rescales θ so as to be between 0 and 1. Arbitrary? Yes. Unmotivated? Not entirely. The advantage of this scheme is that, if the random real r is generated using either a gamma or exponential distribution, the new arm probability will more likely be low than high — reflecting the intuition that bad theories are more easily generated than good theories. If you have the intuition that it is hard to generate *really* bad theories if you honestly seek at the truth, then you might prefer the gamma distribution, as this distribution tends to put the majority of the probability mass on mediocre theories: neither too low, nor too high.

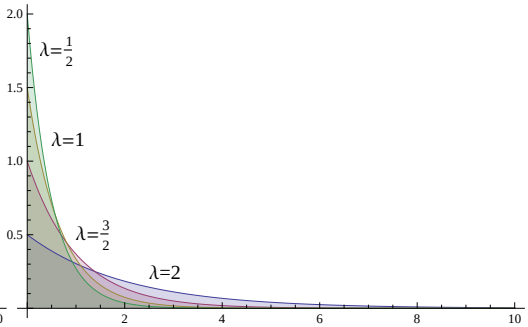
Figure 7 Assigning probabilities to new bandit arms.



(a) Transforming a real into a probability.



(b) The gamma distribution



(c) The exponential distribution

[13] Assuming, of course, proper standards of precision and empirical content. The theory which states "At time t, something will happen" makes correct predictions all the time, but it is useless given the lack of empirical content.

B. Proof that the basic arm-acquiring bandit model has infinitely many arms in the limit.

Let $\langle E_n \rangle_{n=1}^{\infty}$ denote a sequence of random variables representing whether an arm is attached to the bandit in the basic arm-acquiring bandit model of section 2. Let $\langle F_n \rangle_{n=1}^{\infty}$ denote a sequence of random variables of an alternate model where there is simply a $\frac{1}{n+1}$ chance of attempting to attach an arm to the bandit in iteration n. Notice that $\Pr(F_n) \leq \Pr(E_n)$, since the number of balls in the Hoppe-Pólya urn for the basic arm-acquiring bandit model does not always increase each iteration. We will now show that the sequence $\langle F_n \rangle$ will end up, in the limit, attaching an infinite number of arms to the bandit, and hence so will the basic arm-acquiring bandit model.

Now consider the sequence $\langle F_n \rangle$. If the probability associated with a candidate arm is drawn from a Gamma distribution with parameters $\alpha, \beta > 0$, then the expected probability associated with a typical candidate arm is $\frac{2\arctan(\alpha\beta)}{\pi}$, as the mean of the Gamma distribution is $\alpha\beta$. Thus the probability of attaching an arm to the bandit in iteration n is simply $\frac{1}{n+1} \cdot \frac{2\arctan(\alpha\beta)}{\pi}$, as this is the probability of attempting to attach an arm multiplied by the expected probability of that arm winning. Since $\sum_{n=1}^{\infty} \Pr(F_n) = \infty$, by the second Borel-Cantelli theorem it follows that an infinite number of arms will be attached to the bandit. The same holds if the probability associated with a candidate arm is drawn from an exponential distribution with parameter λ, since the expected probability of a typical candidate arm is just a positive constant which does not affect the divergence of the sum $\sum \Pr(F_n)$. More generally, if the probability associated with a candidate arm attempted to be attached in iteration n is p_n (i.e., it can vary as a function of time), and $\sum_{n=1}^{\infty} \frac{p_n}{n+1} = \infty$, then the basic arm-acquiring bandit model will, in the limit, end up with infinitely many arms.

References

J. McKenzie Alexander. Learning to signal in a dynamic world. Working paper., February 2012.

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

A. L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A*, 311:590–614, 2002.

A. Beggs. On the convergence of reinforcement learning. *Journal of Economic Theory*, 122:1–36, 2005.

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.

Fred M. Hoppe. Pólya-like urns and the ewens' sampling formula. *Journal of Mathematical Biology*, 20:91–94, 1984.

John Stuart Mill. *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*. London: Longmans, Green, 8th edition edition, 1904.

M. E. J. Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Science*, 101:5200–5205, 2004.

C. S. Peirce. *The Essential Peirce*, volume 1. Bloomington: Indiana University Press, 1992.

Alvin E. Roth and Ido Erev. Learning in extensive form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior*, 8:164–212, 1995.

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

Caroline S. Wagner and Loet Leydesdorff. Network structure, self-organization, and the growth of international collaboration in science. *Research Policy*, 34: 1608–1618, 2005.

P. Whittle. Arm-acquiring bandits. *The Annals of Probability*, 9(2):284–292, 1981.

Kevin J. S. Zollman. *Network Epistemology*. PhD thesis, University of California-Irvine, 2007.

Kevin J. S. Zollman. The epistemic benefit of transient diversity. *Erkenntnis*, 72 (1):17–35, 2010.