Preferential Attachment and the Search for Successful Theories

J. McKenzie Alexander*

Multiarm bandit problems have been used to model the selection of competing scientific theories by boundedly rational agents. In this article, 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 maximize 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 neighbors. 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

*To contact the author, please write to: Department of Philosophy, Logic and Scientific Method, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, United Kingdom; e-mail: j.alexander@lse.ac.uk.

Philosophy of Science, 80 (December 2013) pp. 769–782. 0031-8248/2013/8005-0048$10.00
Copyright 2013 by the Philosophy of Science Association. All rights reserved.

https://doi.org/10.1086/674080 Published online by Cambridge University Press

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 behavior 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). In this case, it ensures 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, 284) observes that "in 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."

In contrast to multiarm bandits, arm-acquiring bandits provide a better model of choosing among competing scientific theories because new theories can be introduced over time. 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-

1. A Pólya urn model consists of an urn containing an initial assortment of colored balls. At each time step, a ball is sampled with replacement, and then another ball of that color is added to the urn. Now suppose that the ball colors 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 color to the urn. This is the probabilistic reinforcement.

2. This formal result nicely aligns with Peirce's (1992, 189) statement that "the opinion which is fated to be ultimately agreed to by all who investigate, is what we mean by the truth."

https://doi.org/10.1086/674080 Published online by Cambridge University Press

acquiring bandit problem requires agents to negotiate between exploration, exploitation, and innovation.

Adding the extra dimension of innovation to the problem would matter less if it was guaranteed that new theories (new bandit arms) always performed better than the current collection of theories (bandit arms). There is no reason to assume this. Although science advances, with theories of ever-greater empirical adequacy, this overlooks the fact that many candidate theories are discarded before they are 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 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 (1984). 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 periodically invent new theories as well.

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

Consider the following model. Suppose that an agent begins with a Hoppe-Pólya urn containing a single black ball, a zero-arm bandit, and a big bag containing infinitely many arms. Each arm in the bag has a randomly assigned probability of winning, where this unknown probability is drawn from some distribution over $(0, 1)$.[4] At the beginning of each iteration, the agent draws a ball, with replacement. If the black ball is drawn, the agent selects an arm from the bag and attaches it to the bandit. The agent then pulls the arm. If a win occurs, the agent assigns the arm a new color and adds a ball of that color to the urn. If a win did not occur, the agent detaches the arm and returns it to the bag. After time, the urn will contain colored balls in addition to the black ball. If a colored ball is drawn from the urn, the agent simply pulls the appropriately colored handle of the bandit and, if a win occurs, reinforces by adding one ball of that color 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 iden-

3. Apologies to Beckett.

4. See the appendix for how probabilities are assigned to new arms.

https://doi.org/10.1086/674080 Published online by Cambridge University Press

tify 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 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. She still faces the task of identifying the best one to play.

The second point to note is that, in the limit, each bandit acquires infinitely many arms. (This follows trivially from the Borell-Cantelli theorem.) There will also be arms whose probability of winning is arbitrarily close to 1. In the case where there are only a fixed number of arms, 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. The probability assigned to new candidate arms was determined as described in the appendix using a Gamma distribution with parameters $\alpha = 3/2$ and $\beta = 2$. After 10,000 iterations, the bandit 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 to favoring a suboptimal arm. Eventually the agent started to play the higher probability arms more frequently—but note how slowly! After 2 billion iterations, only a small amount of probability has been shifted toward the better arms.

Simulations suggest that boundedly rational agents gradually learn to favor higher probability arms, but very slowly. What happens if we introduce social structure, allowing agents to learn from the discoveries of others?

3. Preferential Attachment and the Social Structure of Science. Publication databases provide a rich source of information regarding the social structure of science. Typically one looks at the coauthorship graph, where nodes represent authors and two authors are connected by an edge if they have coauthored a paper. Barabási et al. (2002) examine "all relevant journals" in mathematics and neuroscience 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: from bibliographical databases in biology, physics, and mathematics, he concludes that the networks 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 differs from a power law, preferential attachment may explain the network structure.

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

https://doi.org/10.1086/674080 Published online by Cambridge University Press

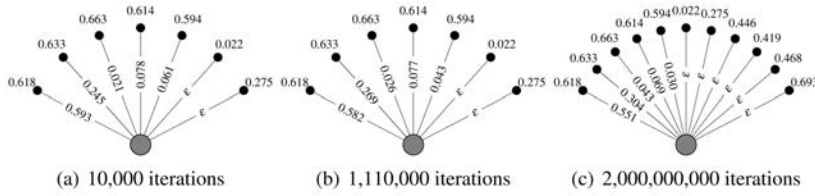


Figure 1. Learning to use the most empirically adequate theory. The gray node represents the agent, smaller black nodes the arms of the bandit. The number above the arm nodes is the probability of that arm winning when pulled. Numbers on the edges indicate the probability of the agent choosing to pull that arm. Probabilities close to zero are denoted by "ε."

1. New researchers join the network at a constant rate and link to existing researchers via preferential attachment.
2. Existing researchers 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 a population of N researchers. Each researcher begins with an arm-acquiring bandit with no arms and a Hoppe-Pólya urn containing just the mutator ball. Define the efficiency of a researcher to be the number of nonblack balls in their urn divided by the number of iterations since she started experimenting. A researcher's efficiency provides a measure of her tendency to make successful predictions. Since we are modeling researchers who aim at the truth, our agents preferentially attach with probability proportional to an agent's efficiency.

When do agents preferentially attach? The reflection rate r determines how often one agent might communicate with another. If a researcher X visits another researcher Y whose efficiency exceeds X's, then X attaches to Y. Attachment generates a one-way communication link where X observes Y and copies his reinforcement,[5] adding arms, if necessary.[6] Once

5. 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.

6. Suppose that X is attached to Y, X's bandit arms are colored red, green, and blue, and that Y's bandit arms are colored purple and magenta. Suppose X reinforces by adding a blue ball and Y reinforces by adding a purple ball. It makes no sense for X to blindly add

https://doi.org/10.1086/674080 Published online by Cambridge University Press

connected, X remains attached to Y as long as Y's efficiency is greater than X's. If Y's efficiency drops below X's, when X next reflects, he breaks the connection and possibly attaches to another.

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 had 50 agents, with new agents introduced with probability .03 and existing agents eliminated with probability .02. Note that the plot shows the actual expected accuracy of a researcher: the sum of the probability of playing each arm 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 theories can spread throughout the population. Without social learning, each agent must discover for herself 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. Modeling the context of discovery pessimistically, using an exponential distribution, lowers the overall empirical accuracy achieved through social learning considerably.

We should not worry too much about this dependency. In the eighth edition of *A System of Logic* (1904), 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 great people,[7] the real question becomes: Does preferential attachment enable the spread of successful theories introduced by a Newton-like figure? There is little

---

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.

7. Mill (1904) 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 *A System of Logic*, chap. 11, bk. 6, p. 612.)

https://doi.org/10.1086/674080 Published online by Cambridge University Press

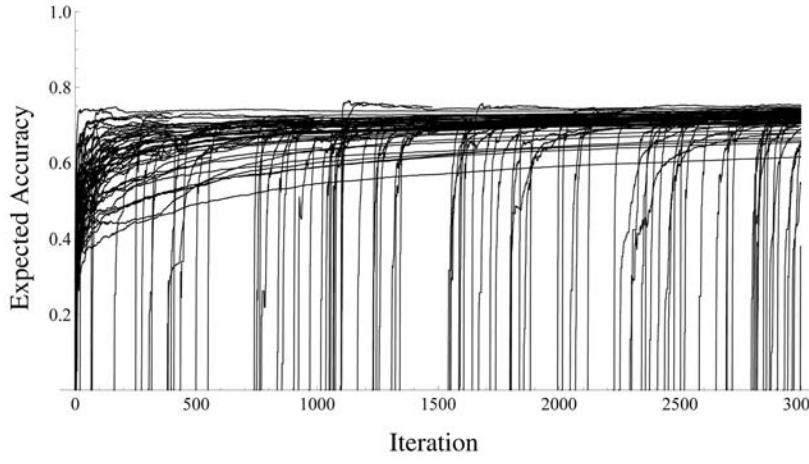


Figure 2. Gradual adoption of empirically adequate theories. Each line plots the time evolution of the expected accuracy of an agent. New lines sprouting from the x-axis correspond to new agents introduced into the population. The probability assigned to new bandit arms was derived from a Gamma distribution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

point attempting to model *how* genius appears; rather, the interesting question is what happens *after* genius appears.

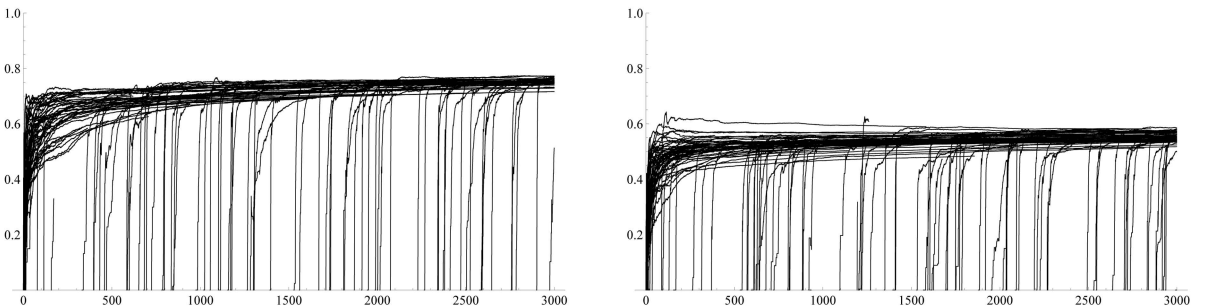
4. Newton, Forgetting, and the Usefulness of Discounting. Taking Mill seriously, consider a variant model: a population of reinforcement learners invent scientific theories and use preferential attachment to learn from others. At some point, a new researcher appears with a theory that always yields correct predictions. That is, after a period of time, we introduce a new researcher (call him "Newton") with an arm-acquiring bandit 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. In figure 4, the rounds of play before Newton appears (1,000 iterations, in this case) generate lock-in to suboptimal theories among the existing researchers. Even though Newton has an optimal theory, the boundedly rational reinforcement learners are incapable of appreciating this point. Furthermore, because Newton himself doesn't know how good his theory is, he may adopt suboptimal theories through preferential attachment.[8]

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

https://doi.org/10.1086/674080 Published online by Cambridge University Press

https://doi.org/10.1086/674080 Published online by Cambridge University Press



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

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

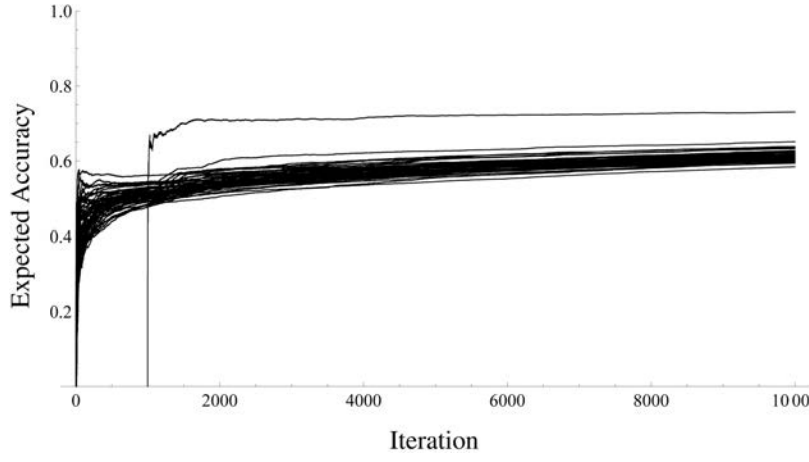


Figure 4. Reinforcement learning generates lock-in to suboptimal theories.

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 new researchers enter the population and existing ones die out? Sometimes: 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 into suboptimal theories, social learning has a greater effect. They are more likely to acquire and use Newton's 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 Newton isn't removed before his theory spreads.

As a story about the spread and adoption of competing scientific theories, there is one parallel 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 realized.

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.[9] In a later

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

https://doi.org/10.1086/674080 Published online by Cambridge University Press

paper, Alexander, Skyrms, and Zabell (2012) show that a certain model of reinforcement learning with forgetting avoids partial pooling traps by reducing the detrimental effects of lock-in.

Suppose that a researcher periodically examines his or her urn and selects a nonblack color at random. Having selected a color, the researcher then discards one ball of that color from the urn. If the last ball of a certain color is discarded, then the researcher also removes the corresponding arm from her bandit. This changes 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 two outcomes that 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 color violet. Now suppose that the researcher preferentially attaches to someone other than Newton (who does not have Newton's theory). Since forgetting first selects a color by type to deinforce, there is a one in four chance that the researcher will select a violet-colored 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.

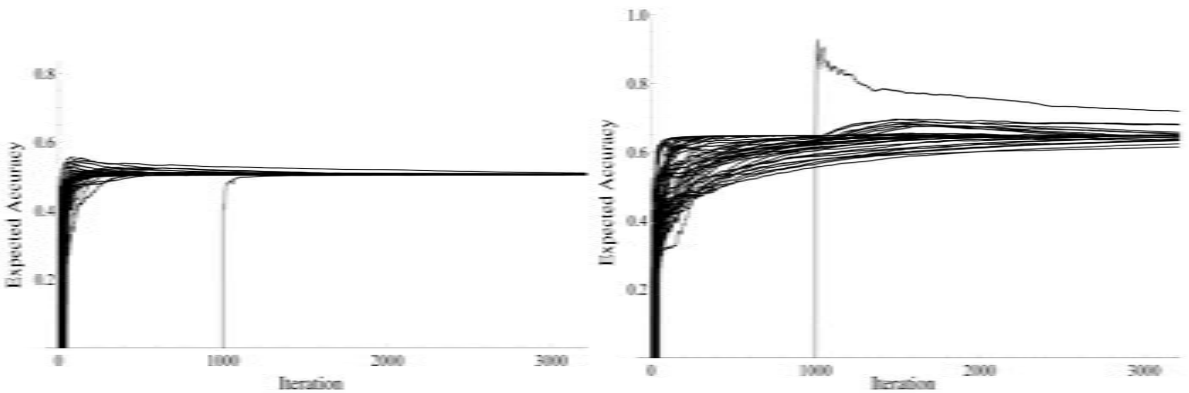
4.2. Discounting. A more realistic model of forgetting can be obtained from theories of discounting used in economics and finance, where one assumes that rational agents trade off present amounts against uncertain future gains. In addition, others (e.g., Wolf 1970; Caplin and Leahy 2004) have suggested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

In a separate paper (Alexander, forthcoming), I have shown how urn-based models of reinforcement learning with discounting,[10] rather than forgetting, work surprisingly well at handling a variety of problems in dynamic sender-

10. Instead of thinking of colors in an urn having integral weights (i.e., a certain number of balls), think of the colors in the urn as having real-valued weights. Discounting occurs at the end of each iteration 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 color is removed from the urn and the corresponding bandit arm removed.

https://doi.org/10.1086/674080 Published online by Cambridge University Press

https://doi.org/10.1086/674080 Published online by Cambridge University Press



(a) Newton forgets his theory. (b) Newton's theory doesn't spread.

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

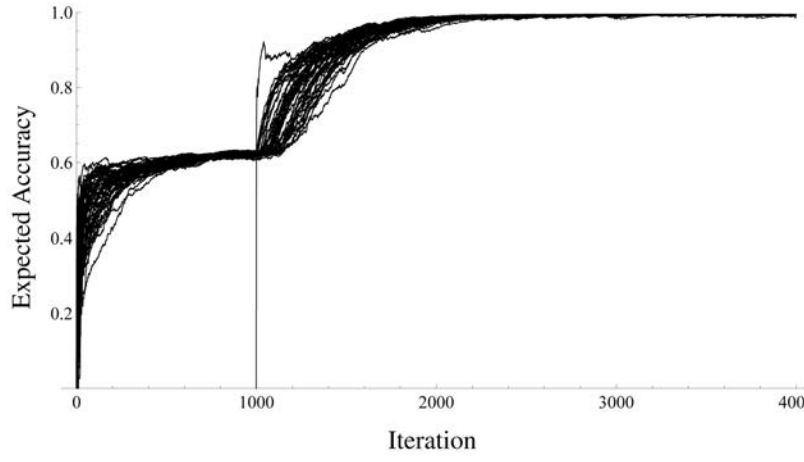


Figure 6. Reinforcement learning, with discounting, combined with preferential attachment, enables the rapid spread of successful scientific theories.

receiver games.[11] 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 discount factor used was $\beta = 0.99$, ensuring that the maximum possible weight attained by reinforcement learning, for any given color, 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 recognizing success rates.

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

11. 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.

https://doi.org/10.1086/674080 Published online by Cambridge University Press

will correctly identify the optimal scientific theory (bandit arm) in the limit, the time required to move even approximately toward 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 modeling 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 behavioral traits may not be an evolutionary, or socially conditioned, accident. Rather, both may be important contributors to our epistemic success as boundedly rational agents.

Appendix

Assigning Probabilities to New Arms

Any model of assigning probabilities to new bandit arms will, of necessity, be speculative, as it formalizes the black box known as the "context of discovery." Nevertheless, there are still worse or better ways to proceed. To begin, consider the nave approach that selects the probability p from the uniform distribution over $[0, 1]$. This is clearly undesirable, as an agent who invents a new theory is just as likely to obtain one of high empirical adequacy $(.99 \leq p \leq 1)$ as one of low empirical adequacy $(0 \leq p \leq .01)$. As we all know too well, false theories are more easily obtained.[12]

Alternatively, consider the following proposal: suppose that a random real r is chosen from the half-open interval $[0, \infty)$.We convert r into a probability via the map $(2/\pi)\arctan(r)$. This transformation simply determines the angle θ whose projection onto the positive reals from the point $(0, 1)$ (with $\theta = 0$ pointing straight down) would yield r 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

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

https://doi.org/10.1086/674080 Published online by Cambridge University Press

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.

REFERENCES

Alexander, J. McKenzie. Forthcoming. "Learning to Signal in a Dynamic World." *British Journal for the Philosophy of Science*.

Alexander, J. McKenzie, Brian Skyrms, and Sandy Zabell. 2012. "Inventing New Signals." *Dynamic Games and Applications* 2 (1): 129–45.

Barabási, A. L., H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. 2002. "Evolution of the Social Network of Scientific Collaborations." *Physica* A 311:590–614.

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

Caplin, Andrew, and John Leahy. 2004. "The Social Discount Rate." *Journal of Political Economy* 112 (6): 1257–68.

Hoppe, Fred M. 1984. "Pólya-like Urns and the Ewens' Sampling Formula." *Journal of Mathematical Biology* 20:91–94.

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

Newman, Mark E. J. 2004. "Coauthorship Networks and Patterns of Scientific Collaboration." *Proceedings of the National Academy of Sciences* 101:5200–5205.

Peirce, Charles S. 1992. *The Essential Peirce*. Vol. 1. Bloomington: Indiana University Press.

Roth, Alvin E., and Ido 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, Brian. 2010. *Signals: Evolution, Learning, and Information*. Oxford: Oxford University Press.

Wagner, Caroline S., and Loet Leydesdorff. 2005. "Network Structure, Self-Organization, and the Growth of International Collaboration in Science." *Research Policy* 34:1608–18.

Whittle, P. 1981. "Arm-Acquiring Bandits." *Annals of Probability* 9 (2): 284–92.

Wolf, Charles, Jr. 1970. "The Present Value of the Past." *Journal of Political Economy* 78 (4): 783–92.

Zollman, Kevin J. S. 2007. "Network Epistemology." PhD diss., University of California, Irvine.

———. 2010. "The Epistemic Benefit of Transient Diversity." *Erkenntnis* 72 (1): 17–35.

https://doi.org/10.1086/674080 Published online by Cambridge University Press