Efficient Convergence Implies Ockham's Razor

Kevin T. Kelly
Department of Philosophy
Carnegie Mellon University
kk3n@andrew.cmu.edu

March 6, 2005

Abstract

A finite data set is consistent with infinitely many alternative theories. Scientific realists recommend that we prefer the simplest one. Anti-realists ask how a fixed simplicity bias could track the truth when the truth might be complex. It is no solution to impose a prior probability distribution biased toward simplicity, for such a distribution merely embodies the bias at issue without explaining its efficacy. In this note, I argue, on the basis of computational learning theory, that a fixed simplicity bias is necessary if inquiry is to converge to the right answer efficiently*, whatever the right answer might be. Efficiency is understood in the sense of minimizing the least fixed bound on retractions or errors prior to convergence.*

Keywords: learning, induction, simplicity, Ockham's razor, realism, skepticism

1 Introduction

There are infinitely many alternative theories compatible with any finite amount of experience, so choosing the right one on the basis of the evidence alone seems hopeless. Scientfic realists "justify" such choices by invoking other methodological virtues like simplicity, unity, uniformity of nature, or minimal causal entanglement to narrow the field. These appeals to "Ockham's razor" smack of wishful thinking, however, for *how could* a fixed bias toward simple theories *possibly* facilitate finding the right answer? A *fixed* bias of any kind can no more indicate the right answer than a stopped clock can indicate the time. This concern is a principal motive for anti-realism.

Here is a bad explanation: impose a prior probability distribution biased toward simplicity and then argue, on the basis of this distribution, that a simplicity bias probably leads to the right answer. Whatever this tight circle "justifies", it does not *explain* how such a bias could facilitate finding the right answer because it presupposes the very bias to be explained.[1] My

[1] Sometimes this bias is hidden by the Bayesian apparatus. For example, the realist's "miracle" argument amounts to this. "Surely an open-minded anti-realist who keeps the door open to all the skeptical possibilities should 'leave the door open' to the unified theory when choosing between a unified theory and a disunified one. But since the disunified theory has more parameters to set in order to yield the predictions, the likelihood of the data given the disunified theory is infinitesimal with respect to the likelihood with respect to the unified theory. Since the priors of the two theories are real- valued, by Bayes theorem, the posterior of the disunified theory is infinitesimal compared to that of the unified theory. So it would be a 'miracle' if the disunified theory were true." The trick is that the "fair-minded" distribution over the two *theories* forces an infinite bias against disunified *worlds*, so the anti-realist is actually offered a sucker's choice between a dogmatic bias against the unified theory and a dogmatic bias against disunified worlds (in which only the former bias is explicit). The story I am about to tell does not refer to prior probability at all, so there is no place for such a bias to hide.

aim in this note is to provide the anti-realist with a fairly generally applicable explanation that does not beg the question by presupposing a simplicity bias.

2 The Main Results

Let's review the scientific realist's options. One can't explain how Ockham's razor helps us find the truth by presupposing a bias toward simplicity, since Ockham's razor is just such a bias: it would be like explaining how opium works in terms of a "dormative virtue". Nor can one argue, without presupposing such a bias, that simplicity points at the truth, because it points in the wrong direction in complex worlds. Hence, the realist must show how simplicity could help us find the truth without pointing at the truth.

So what could such "help" amount to? Here is an analogy. The ideal automobile may not be as fast or as beautiful as we would like, but any increase in one direction would entail a more than compensating loss elsewhere. Maybe Ockham's razor is like that: deviating from it may reduce epistemic costs (e.g., errors or retractions) in some complex worlds, but the improvement may entail still greater numbers of errors and retractions in other worlds, reducing one's efficiency overall. In other words, it may turn out that violating Ockham's razor reduces one's overall **epistemic efficiency**, which requires that one achieve the least achievable, fixed bound on epistemic costs over all possible worlds satisfying the presuppositions of the empirical problem at hand.[2]

An anti-realist should be impressed to learn that commitment to Ockham's razor in one's inductive strategy minimizes errors and retractions prior to convergence to the right answer. She should be still more impressed to see it explained how such a bias is also *necessary* for retraction or error efficiency. That is just what I claim. Readers who prefer examples to principles may prefer to skim section 3 at this point.

An **empirical problem** consists of a set of mutually exclusive **possible answers** that jointly exhaust the problem's **presupposition**, which is the set of possible worlds over which a corrrect answer must be given. Each world affords a potentially infinite sequence of **inputs** to the learner. A **learning method** responds to each finite sequence of inputs with a potential answer or with '?', which indicates "not ready to choose". A method **solves** an empirical problem just in case it stabilizes, eventually, to a correct answer to the question in each possibility compatible with the problem's presupposition. R. Freivalds and C. Smith (1993) have devised an ingenious definition of solving a problem **under a transfinite retraction bound** that is much more general than the more obvious concept of success with finitely bounded retractions introduced in (Putnam 1965). The following results are based on a refinement of Freivalds and Smith's idea.

The notion of an **Ockham answer** relative to the current inputs can be defined in a language-invariant manner that reflects intuitive simplicity in particular cases (cf. section 3 below for examples) so that the following are mathematical theorems.[3]

Proposition 1 *A solution to a problem is uniformly efficient with respect to errors just in case it never outputs an informative answer other than the (unique) Ockham answer for the current data.*

[2]The idea of counting retractions prior to convergence was appealed to for purely logical ends by Hilary Putnam in (1965). Counting retractions as a definition of the intrinsic difficulty or complexity of an empirical problem has seen a great deal of study in computational learning theory. A good reference and bibliography may be found in (Jain et al. 1999). What follows is heavily indebted to (Freivalds and Smith 1993) and generalizes the topological perspective on learning developed in (Kelly 1996).

[3]The proofs are based on what I call "surprise complexity": a topological invariant that generalizes both Cantor-Bendixson rank and Kuratowski's (transfinite) difference hierarchy (Kechris 1991).

Proposition 2 *A solution to a problem is uniformly efficient with respect to retractions or errors only if it never outputs an informative answer other than the (unique) Ockham answer for the current data.*

It is familiar wisdom that convergence in the limit is compatible with any crazy behavior in the short run (e.g., Earman 1992). The results in this paper show that the wisdom is wrong (with a vengeance) when we require efficient convergence, for then one's only choice is to accept the unique, Ockham answer or to hold one's tongue.
The conditions for retraction efficiency impose even tighter restrictions on short-run acceptance behavior. Skeptical retreat from an informative answer to '?' counts as a retraction, but not as an error. So minimizing retractions must impose some restriction on skeptical retreats as well as on which informative answer one chooses. It is possible to define the notion of an **anomaly** so that the following is a mathematical theorem.

Proposition 3 *A solution to a problem is uniformly efficient in terms of retractions just in case it never retracts an informative answer until a corresponding anomaly has occurred.*

Proposition 3 provides an anti-realist explanation of another methodological reflection of realist attitudes: never drop an accepted theory unless there is a concrete, empirical anomaly. Propositions 3 and 2 have a surprising corollary.

Proposition 4 *If a solution to a problem never retracts an informative answer unless an anomaly occcurs, then it never outputs any informative answer other than the (unique) Ockham answer.*

The surprise is that a constraint on when to drop what you accepted could entail a constraint on what to accept in the first place. The explanation is that if you accept a needlessly complex answer , the constraint on retraction will prevent you from ever dropping it, so you won't converge to the right answer. Thus, simplicity and resolution are essentially bound to one another by the concept of convergent success.
There is an escape hatch for the anti-realist: if efficiency is not achievable at all, then efficiency implies Ockham's razor only in the trivial sense that it implies everything. But the escape comes with a cost, for it is available only if the presuppositions of the problem are empirically inscrutable even in the ideal limit of inquiry.

Proposition 5 *If a problem has a solution that also converges in the limit to '?' when the presupposition of the problem is false, then the problem has error-efficient and retraction-efficient solutions.*

3 Some Illustrations

The exact definitions underlying the preceding results cannot be motivated in this brief note. Instead, I will illustrate the results with some examples. Suppose that, for whatever reason, the possibilities on the table are worlds in which all inputs are green and in which all inputs are grue_t, where a grue_t observation is green up to and including t and blue thereafter. The question is which kind of world we are in. The "natural" approach is to eventually become sure that the world is "uniformly" green and to retract to grue_t only after a blue input is received at t. This approach retracts at most once (when the first blue input is received). But if one were ever to project grue_t prior to receipt of a blue input, Nature would be free to continue presenting green inputs until one retracts to "all inputs are green", on pain of converging to the wrong answer when all inputs are green. Thereafter, Nature could present

all blue inputs, exacting two retractions in a problem that could be solved under a unit retraction bound. By a similar argument, projecting "forever green" minimizes errors, but the least feasible error bound is ω. The results generalize if we add worlds of type $\text{grue}_{t,t'}$ whose inputs are green through t, blue through t' and then green thereafter, $\text{grue}_{t,t',t''}$, and so forth, as long as there is a finite bound on the number of "surprises" (Schulte 1999a, b).

The point of Nelson Goodman's (1983) grue_t construction was to show that uniformity is relative to description and that perfect definitional symmetry blocks any attempt to favor one description over another on syntactic grounds. The preceding argument does not appeal to uniformity relative to a description or to syntactic definitional form, however. It hinges on a topological asymmetry in the underlying branching structure of the problem, for the "forever green" input stream is the unique input stream for which distinct input streams compatible with the problem "veer off" infinitely often (no input streams compatible with the problem veer off of "forever grue_t" after stage t). This property is preserved under translation into the $\text{grue}_t/\text{bleen}_t$ language, for the translation is just a one-to-one relabeling of the inputs along each input stream, which evidently leaves branching structure of the problem intact.[4] This conception of simplicity is contextual, in the sense that the same world can be simple or complex, depending on the problem we face (Chart 2000). For example, we can make the "forever grue_9" world into the spine by considering only the worlds "forever grue_1", ..., "forever grue_9", "forever green", and "forever $\text{grue}_{9,t}$", for all $t > 9$. Why should one say that "forever grue''_9 is the simplest or most uniform answer in this problem when the answer "forever green" is available? Because in either problem, the spine world is simpler or more uniform in the methodological sense that the empirical problem one faces never gets easier, no matter how much experience one receives. In all the alternative worlds, there is a time after which some answer is determinately verified. This idea can be generalized by transfinite recursion to yield non-trivial, infinite degrees of simplicity.

Ockham's razor is (roughly) a matter of presuming that the actual world is among the simplest worlds compatible with the current inputs. The principle accords with a surprising variety of "simplicity" judgments. For example, a familiar policy in particle physics is to posit the most restrictive conservation laws compatible with reactions that are known to have occurred (Ford 1963). Here, the "spine" world is one in which only the known reactions are possible and "veering" occurs when a new type of reaction that is not permitted by the earlier conservation laws is observed. If there are at most n particles, all of which are observable, then by an argument like the preceding one, achievement of the least feasible retraction bound in each subproblem demands that one never choose a conservation theory compatible with a non-observed reaction (cf. Schulte 2001).

In the context of curve fitting, simplicity is often identified with the polynomial degree of the curve's equation. Suppose we wish to know the degree of an empirical curve from evidence gathered with error $< \epsilon$ and it is known that the true degree is n. If we were to guess a degree higher than k when k is the least degree compatible with the inputs, Nature is free to make it appear that the true degree is k until we take the bait (on pain of converging to the wrong answer). Thereafter, Nature is free to choose a curve of properly higher degree that remains compatible with the inputs presented so far and to present inputs from it until we retract. Nature can force another retraction in this way for each further degree $< n$, for a total of $n - k + 1$ when $n - k$ would have sufficed.

Copernican astronomy, Newtonian physics, the wave theory of optics, evolutionary theory, and chaos theory all won their respective revolutions by providing unified, low-parameter explanations of phenomena for which their competitors required many. Suppose that there is

[4] In mathematical jargon, grue-like translations are just continuous automorphisms of the problem with respect to the "branching" or Baire space topology restricted to the problem's presupposition.

a series of logically independent, empirical laws $L_0, \ldots, L_n$ and a corresponding series of mutually exclusive theories such that T_i entails $L_0, \ldots, L_i$. Suppose we were to accept any theory other than L_0 a priori. Then by a similar argument, Nature could exact $n+1$ retractions from us, whereas if we had always assumed the "most unified" theory compatible with experience, and if Nature promises to show us evidence refuting any false law eventually, then Nature could have exacted at most n retractions.[5]

Ockham's razor is often understood as a bias toward fewer causes. Recent years have seen a considerable increase in our understanding of causal inference (Spirtes et al. 2000, Pearl 2000). Instead of "reducing" causes to probabilistic or modal relations, the idea is to axiomatize the connection between probability and causation. A consequence of these axioms is that there is a direct, causal connection between two variables (one way or the other) just in case the two variables are probabilistically dependent conditional on each subset of the remaining variables. One then says that the two variables are **d-connected**. Otherwise, they are **d-separated**. The methodological question is what to infer now, from the available data. Spirtes et al. have proposed the following method (which I oversimplify). For each pair of variables X, Y, perform a standard statistical test of independence of X and Y conditional on each subset of the remaining variables. If every such test results in rejection of the null hypothesis of independence, conclude that X and Y are d-connected and add a direct causal link between X and Y (without specifying the direction). Otherwise, provisionally conclude that there is no direct causal connection. In other words, assume the smallest number of causes compatible with the outcomes of the tests. By an argument analogous to those already given, one must follow such a procedure or Nature could elicit more retractions than necessary (at most n retractions are required by the algorithm proposed by Spirtes et al., one for each possible direct causal connection among the variables under consideration).[6]

4 Acknowledgements

This note benefits from discussions with Oliver Schulte, Joseph Ramsey, Peter Spirtes, Richard Scheines, and Julia Pfeifer. It was also improved by the remarks of the referees.

5 References

Chart, D. (2000) "Schulte and Goodman's Riddle", *The British Journal for the Philosophy of Science*, 51: 147-149.

Earman, J. (1992), *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory, Cambridge: MIT Press.*, M.I.T. Press.

Ford, K. (1963) *The World of Elementary Particles*, New York: Blaisdell.

Freivalds, R. and C. Smith (1993) "On the Role of Procrastination in Machine Learning", *Information and Computation* 107: pp. 237-271.

Goodman, N. (1983) *Fact, Fiction, and Forecast*, fourth edition, Cambridge: Harvard University Press.

[5] The argument doesn't recommend the choice of a unifying theory over the conjunction of the unified laws, however, as these alternatives are not mutually exclusive.

[6] Here, I neglect the small probability of a mistaken rejection. For a more literal learning theoretic analysis of statistical tests, cf. (Kelly 96, chapter 3). For an explicitly statistical treatment of related issues, cf. (Robins et al. 2000).

Jain, S., Osherson, D., Royer, J. and Sharma A. (1999) *Systems that Learn* 2nd ed., Cambridge: M.I.T. Press.

Kechris, A. (1991) *Classical Descriptive Set Theory*, New York: Springer.

Pearl, J. (2000) *Causality*, Cambridge: Cambridge University Press.

Putnam, H. (1965) "Trial and Error Predicates and a Solution to a Problem of Mostowski", *Journal of Symbolic Logic* 30: 49-57.

Schulte, O. (1999a) "The Logic of Reliable and Efficient Inquiry", *The Journal of Philosophical Logic*, 28:399-438.

Schulte, O. (1999b), "Means-Ends Epistemology", *The British Journal for the Philosophy of Science*, 50: 1-31.

Schulte, O. (2001) "Inferring Conservation Laws in Particle Physics: A Case Study in the Problem of Induction", *The British Journal for the Philosophy of Science*, Forthcoming.

Spirtes, P., Glymour, C., and Scheines, R. (2000) *Causation, Prediction and Search*, 2nd ed., Cambridge: M.I.T. Press.

Kelly, K. (1996) *The Logic of Reliable Inquiry*, New York: Oxford.

Martin, E. and D. Osherson (1998). *Elements of Scientific Inquiry*, Cambridge: M.I.T. Press.

Robins, J., Scheines, R., Spirtes, P., and Wasserman, L. (2000) "Uniform Consistency in Causal Inference", Carnegie Mellon University Department of Statistics Technical Report 725.

Valdes-Perez and Erdmann (1994) "Systematic Induction and Parsimony of Phenomenological Conservation Laws", *Computer Physics Communications* 83: 171-180.