Efficient Convergence Implies Ockham's Razor

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

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. I propose, 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 retractions or errors prior to convergence to the right answer.*

Keywords: simplicity, Ockham's razor, realism, skepticism, computational learning theory

1 Introduction

There are infinitely many alternative theories compatible with any finite amount of experience. How do we choose the right one? Scientific realists justify such choices on the basis of simplicity, unity, uniformity of nature, or minimal causal entanglement. 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 "track" or indicate the right answer than a stopped clock can indicate the time.

Here is a bad explanation: impose personal probabilities biased toward simplicity on the problem and then argue, on the basis of this bias, that a prior bias toward simplicity probably helps us find the truth. Whatever this tight circle "justifies", it does not *explain* how a simplicity bias could facilitate finding the right answer because it presupposes the very bias to be explained.[1] Nor does it help to say that if the bias is in error, it will "wash out" in light of future data: so would a prior bias toward complexity. The question is how a simplicity bias could help us find the truth, not how we could eventually recover from its ill effects.

The realist must explain how simplicity could help us find the truth without pointing at or indicating the truth— but isn't that double-talk? Not quite. The ideal automobile may not be as fast or as beautiful as we would like, but any gain in one direction implies a more than compensating loss elsewhere. Maybe Ockham's razor is like that: deviating from it reduces epistemic costs (e.g., errors or retractions) in some complex worlds, but the improvement subtly entails still greater num-

[1] Sometimes this bias is hidden by the Bayesian apparatus. Consider two competing theories, $T, T[\alpha]$, where α is an adjustable, real-valued parameter. It would be "unfair" to start out with an infinite bias against the simple theory T, so suppose that $T, T[\alpha]$ both have non-extremal, real-valued prior probabilities. But $T[\alpha]$ can be true in continuum many ways (one for each setting of α), so each parameter setting of $T[\alpha]$ has infinitesimal prior probability with respect to the unique parameter setting of T. This prior bias against parameter settings of $T[\alpha]$ is why T ends up much better confirmed than $T[\alpha]$ on data explained by T but only under particular parameter settings of $T[\alpha]$. We say that it would "be a miracle" if $T[\alpha]$ were true. But the miracle is in ourselves, not the world, since it is a reflection of our infinite, prior bias against worlds in which $T[\alpha]$ is true.

bers of errors and retractions in other worlds, so that one's *overall* (i.e., worst-case) epistemic costs are increased. In other words, Ockham's razor may optimize overall, epistemic *efficiency* even though it points in the wrong direction in some (or even most) of the possibilities under consideration.[2] That is just what I claim.

In a nutshell, the idea is this. Complex worlds present signs ("anomalies") that witness their complexity, whereas simple worlds do not. So if you conclude that an "anomaly" will appear prior to seeing one, Nature is free to withhold anomalies until you concede that none will appear (on pain of converging to the wrong answer). Thereafter, Nature is free to present one, forcing you to change your mind yet again, for a total of two retractions (note the further embarrassment that they lead in a circle back to where you started). Had you sided against anomalies until seeing one, you would have succeeded with at most one retraction (without a cycle of opinions), for Nature cannot "take back" an anomaly once it has been presented.

This simple idea extends to cases in which there is no finite bound on the number of anomalies that might occur. The extension is necessary for deriving Ockham's razor from error-efficiency, for whenever the problem of induction arises, there is no finite bound on errors prior to convergence to the truth (e.g., Nature can delay presenting the white raven arbitrarily long after we have concluded that all ravens are black).

The general results are stated in the next section. Readers who prefer examples to principles may prefer to read section 3 first.

2 Results

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 refusal to choose an answer. A method *solves* an empirical problem just in case it stabilizes, eventually, to a correct answer to the question in each world compatible with the problem's presupposition. R. Freivalds and C. Smith [4] have devised an ingenious definition of solving a problem *under a transfinite retraction bound* that generalizes the more straightforward concept of finitely bounded retractions introduced in [9]. The following results are based on a refinement of Freivalds' and Smith's idea.

Worlds are *simpler* (roughly) insofar as they present "fewer" anomalies. The *Ockham* answer is (quite roughly) the unique answer satisfied by the simplest worlds compatible with the current inputs.[3] This concept reflects intuitive simplicity in particular cases (cf. section 3) and also permits one to prove the following, mathematical theorem.[4]

Proposition 1 *If a solution to a problem is error or retraction-efficient and the method outputs an informative answer, then the answer is the (unique) Ockham answer given the current inputs.*

In the case of error-efficiency, the converse also obtains.

[2] The idea of counting retractions prior to convergence was first invoked (for purely logical ends) by Hilary Putnam [9]. 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 [6]. What follows is heavily indebted to Freivalds and Smith [4] and refines the topological perspective on learning developed in [14].

[3] An answer uniquely satisfied by the simplest worlds compatible with the inputs is Ockham and the Ockham answer is an answer for which the least upper bound on the simplicities of the worlds that satisfy it is minimum. Neither converse holds, however (since simplicity admits of transfinite degrees, the supremum of the simplicities of worlds satisfying a hypothesis may exceed the complexity of any such world).

[4] 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 [7].

Proposition 2 *The error-efficient solutions to a problem are exactly the solutions that never output an informative answer other than the (unique) Ockham answer for the current data.*

The converse doesn't hold for retraction-efficiency. 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 one's choice among informative answers.

Proposition 3 *The retraction-efficient solutions to a problem are exactly the solutions that never retract an informative answer until a corresponding anomaly has occurred.*

Proposition 3 explains another realist attitude: the mere possibility of an alternative explanation doesn't undermine the current scientific consensus unless the alternative explanation is simpler.

Propositions 2 and 3 have a surprising corollary.

Proposition 4 *If a solution to a problem never retracts an informative answer unless an anomaly occcurs, then it never adopts an answer unless that answer is the (unique) Ockham answer in light of the current inputs.*

The surprise is that a constraint on when to drop what you accepted could entail a constraint on which answer to adopt 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 a possible 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 converges to '?' when the presupposition of the problem is false, then the problem has an error and retraction-efficient solution, so the preceding results apply non-trivially.*

It is familiar wisdom that convergence in the limit is compatible with any crazy behavior in the short run [2]. The preceding results reverse this wisdom (with a vengeance) when we require efficient convergence. An error-efficient method is forced to side with the Ockham hypothesis if it sides with any hypothesis at all, and a retraction-efficient method is also forced to hang on to its simple theory until an anomaly signals that it may be dropped.

3 Illustrations

The precise definitions underlying the preceding results cannot be developed in this brief note. Instead, I will illustrate them with some simple examples. It should be kept in mind, however, that the general results apply as long as the problem at issue is solvable in the limit and its presupposition is decidable in the limit (proposition 5); a condition satisfied by empirical problems infinitely more complex than any of the following examples.

Uniformity of nature. 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 (an "anomaly") 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 grue$_{t,t'}$ whose inputs are green through t, blue through t' and then green thereafter, grue$_{t,t',t''}$, and so forth, as long as there is a finite bound on the number of "anomalies" [10] [11].

The point of Nelson Goodman's [5] grue$_t$ construction was to show that uniformity is relative to description and that 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, rather, on a description-independent, *topological* asymmetry in the branching structure of the possible input streams compatible with the presuppositions of the problem. 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 Goodman's translation into the grue$_t$/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.[5] The proposed conception of simplicity is contextual, in the sense that the same world can be simple or complex, depending on the problem we face [1]. 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 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 syntactically "uniform" answer "forever green" is available? Because in the spine world in which "forever grue$_{9,t}$" is true, the intrinsic difficulty of the problem never drops, no matter how much experience one receives. In all the alternative worlds, there is a time after which some answer is determinately verified, so that finding the right answer becomes trivial, so the structure of the problem one faces is fundamentally altered. This idea can be generalized by transfinite recursion to yield non-trivial, infinite degrees of simplicity.

Conservation laws. 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 [3][17]. 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 [12].

Curve fitting. In the context of curve fitting, simplicity is often identified with the polynomial degree of a 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 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.

Theoretical unification. Copernican astronomy, Newtonian physics, the wave the-

[5] 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.

ory 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 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 $L_0, \ldots, L_i$. Let the presupposition of the problem be that one of the theories T_i is true and that if T_i is true, then some counterexamples to $L_{i+1}, \ldots, L_n$ will appear, for otherwise, the total inputs for eternity would not distinguish T_i from T_{i+1}. Suppose we were to accept T_i a priori, where $i < n$. Then Nature could withhold counterinstances to $L_0, \ldots, L_n$ until we revise to L_n (on pain of converging to the wrong answer). As soon as we do so, she is free to feed a counterexample to L_n to force us to retract to T_{n-1}, and so forth, for a total of $n + 1$ retractions when the obvious, Ockham method would have succeeded with at most n.[6]

Causal simplicity. Ockham's razor is often understood as a bias toward fewer causes. Our understanding of causal inference has improved considerably in recent years [13] [8]. Instead of "reducing" causation to probabilistic or modal relations, the idea is to axiomatize the connection between probability and causation.[7] 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).[8]

4 Piece-meal Efficiency

The anti-realist may find some comfort in the following result, which offsets the optimism suggested by proposition 5.

Proposition 6 *If a problem is error or retraction-efficiently solvable, then some answer is eventually verified in some world satisfying the presuppositions of the problem.*

For example, if we presuppose that the curve we are fitting has at most degree n, then eventually all degrees $< n$ are refuted so degree n is verified. If there were no such a priori bound, however, no polynomial degree would ever be verified so by the preceding result, no method is an efficient solution and the proposed reason for preferring simple theories is trivial. The same point applies to uniformity of nature if

[6] 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. Nor does it explain why we should choose a unified theory over a complex competitor when it is possible that the data could be the same for eternity, regardless of which is true.

[7] One of these assumptions, called *faithfulness* [13], essentially says that if the world is causally complex, eventually we will see data (i.e., a sufficiently large sample) in which a causally simple story looks bad. This is similar to the assumption I invoked in the preceding example.

[8] Here, I neglect the small probability of a mistaken rejection. For a more literal learning theoretic analysis of statistical tests, cf. [14], chapter 3. For an explicitly statistical treatment of issues dealing with related themes, cf. [16].

no finite, a priori bound on "breaks in uniformity" is presupposed.

The approach may be extended to such problems in a *piece-meal* sense [12]. Say that problem P' is a *coarsening* of a given problem P if each answer to P' is a (possibly infinite) disjunction of answers to P. A method that outputs answers to the original problem P is not charged for an error (or retraction) in the coarsening P' as long as it outputs (or continues to output) a disjunct of the right answer in P'. Then a method is *piece-meal* efficient in the original problem P if it is efficient in each coarsening of P that has an efficient solution. As one would hope, accordance with Ockham's razor is necessary for piece-meal error and retraction-efficiency in each of the preceding examples when the finite bounds are relaxed.

5 Acknowledgements

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

References

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

[2] Earman, J. *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory.* Cambridge: MIT Press, 1992.

[3] Ford, K. *The World of Elementary Particles.* New York: Blaisdell, 1963.

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

[5] Goodman, N. *Fact, Fiction, and Forecast*, fourth edition. Cambridge: Harvard University Press, 1983.

[6] Jain, S., D. Osherson, J. Royer, and A. Sharma. *Systems that Learn*, second edition. Cambridge: M.I.T. Press, 1999.

[7] Kechris, A. *Classical Descriptive Set Theory.* New York: Springer, 1991.

[8] Pearl, J. *Causality.* Cambridge: Cambridge University Press, 2000.

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

[10] Schulte, O. "The Logic of Reliable and Efficient Inquiry". *The Journal of Philosophical Logic*, 28:399-438, 1999.

[11] Schulte, O. "Means-Ends Epistemology". *The British Journal for the Philosophy of Science*, 50: 1-31, 1999.

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

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

[14] Kelly, K. *The Logic of Reliable Inquiry.* New York: Oxford, 1996.

[15] Martin, E. and D. Osherson *Elements of Scientific Inquiry*, Cambridge: M.I.T. Press, 1998.

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

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