A New Solution to the Puzzle of Simplicity

4623 words

Abstract

Explaining the connection, if any, between simplicity and truth is among the deepest problems facing the philosophy of science, statistics, and machine learning. Say that an *efficient* truth-finding method minimizes worst-case costs *en route* to converging to the true answer to a theory choice problem. Let the costs considered include the number of times a false answer is selected, the number of times opinion is reversed, and the times at which the reversals occur. It is demonstrated that (1) always choosing the simplest theory compatible with experience and (2) hanging onto it while it remains simplest is both necessary and sufficient for efficiency.

1 The Puzzle of Simplicity

Philosophy of science, statistics, and machine learning all recommend the selection of simple theories or models on the basis of empirical data, where simplicity has something to do with minimizing independent entities, principles, causes, or equational coefficients. This intuitive preference for simplicity is called Ockham's razor, after the fourteenth century theologian and logician William of Ockham. But in spite of its intuitive appeal, how *could* Ockham's razor help us find the true theory? For if we already know that the truth is simple, we don't need Ockham's help. And if we don't already know that the truth is simple, what entitles us to assume that it is?

It doesn't help to say that simplicity is associated with other virtues such as testability (Popper 1968), unity (Friedman 1983), better explanations (Harman 1965), higher "confirmation" (Carnap 1950, Glymour 1980), or minimum description length (Li and Vitanyi and Li 2000), since if the truth weren't simple, it wouldn't have these nice properties either. To assume otherwise is to engage in wishful thinking (vanFraassen 1981).

Overfitting arguments (Akaike 1973, Forster and Sober 1994) show that using a complex model for predictive purposes in the presence of random noise can increase the expected squared error of predictions. But that is still the case when you know in advance that the truth is complex, so overfitting arguments concern accuracy of prediction rather than finding the true theory. Furthermore, if one is interested in predicting the causal outcome of a policy on the basis of non-experimental data, the prediction could end up far from the mark because the counterfactual distribution after the policy is enacted may be quite different from the distribution sampled (Spirtes and

Zhang 2003). Finally, such arguments work only in statistical settings, but Ockham's razor seems no less compelling in deterministic ones.

Nor is Ockham's razor explained by a prior probabilistic bias in favor of simple possibilities, for the propriety of a systematic bias in favor of simplicity is precisely what is at issue. The argument remains circular even if complex and simple theories receive equal prior probabilities, for theories with more free parameters can be true in more "ways", so each way the complex theory might be true ends up carrying less prior probability than each of the ways the simple theory might be true, and that prior bias toward simple possibilities is merely passed through Bayes' theorem (e.g., Rosenkrantz 1983 and the discussion of the Bayes information criterion in Wasserman 2004).

There are non-circular, relevant arguments for Ockham's razor, if one is willing to grant speculative premises. G. W. Leibniz (1714) appealed to the Creator's taste for elegance. More recently, some "naturalistic" philosophers and machine learning researchers have replaced Providence with an equally benevolent, evolutionary etiology (e.g., Mitchell 1997, p. 66; cf. also Duda et al. 2000, pp. 464-465). But even if these adaptationist speculations were were true, they explain the truth-finding efficacy of Ockham's razor only in dealings with matters of pre-historic survival. How does simplicity continue to track the truth in the vastly expanded linguistic and experiential realm of contemporary science? To respond that what was successful in prehistorical applications will continue to succeed in future situations is an appeal to the simple uniformity of nature and, hence, to Ockham's razor, which is another circle.

Even if Providence or evolution did arrange the truth of simple theories in a way that we can never know without begging the question, it would surely be nice, in addition, to have a clear, normative argument to the effect that Ockham's razor is the most efficient possible method for finding the true theory when the problem involves theory choice. This note presents just such an argument.[1] The idea is that it is hopeless to provide an *a priori* explanation how simplicity points at the truth immediately, since the truth may depend upon subtle empirical effects that have not yet been observed or even conceived of. The best that Ockham's razor could guaranteed to achieve *a priori* is to keep us on the straightest possible path to the truth, allowing for unavoidable twists and turns along the way as new effects are discovered—and that is just what it guarantees. Readers who wish to cut to the chase may prefer to peek immediately at theorem 1 in section 5 prior to reviewing the relevant definitions.

2 Illustration: Empirical Effects

Suppose that you are interested in the form of an unknown polynomial law

$$f(x) = \sum_{i=0}^{n} a_i x^i.$$

[1]The approach is based on concepts from computational learning theory. For a survey of related ideas, cf. (Jain et al., 1999) and (Kelly 1996). Earlier versions of the following argument may be found in (Schulte 1999, Kelly 2002, Kelly 2004, Kelly and Glymour 2004, and especially, Kelly 2005 and Kelly 2006).

It seems that laws involving fewer monomial terms are simpler, so Ockham's razor favors them. Suppose that patience and improvements in measurement technology allow one to obtain ever tighter open intervals around $f(x)$ for each specified value of x as time progresses.[2] Suppose that the true degree is zero, so that f is a constant function. Each finite collection of open intervals around values of f is compatible with degree one (linearity), since there is always a bit of wiggle room within finitely many open intervals to tilt the line. So suppose that the truth is the tilted line that fits the data received so far. Eventually you can obtain data from this line that refutes degree zero. Call such data a (first-order) *effect*. Any further, finite amount of data collected for the linear theory is compatible (due to the remaining, minute wiggle room) with a quadratic law, etc. The truth is assumed to be polynomial, so the story must end, eventually, at some finite set A of effects. Thus, determining the true polynomial law amounts, essentially, to determining the finite set A of all monomial effects that one will ever see.

So conceived, empirical effects have the property that they never appear if they don't exist but may appear arbitrarily late if they do exist.[3] To reduce the curve-fitting problem to its essential elements, let E be a denumerable set of *potential effects* and assume that at most finitely many of these effects will ever occur. Assume that your lab merely reports the finite set of effects that have been detected so far, so a *world* or *input sequence* is an a sequence of finite subsets of E that converges to some finite subset of E. An *input stream* or *empirical world* is an infinite input sequence. Let the effects presented in input sequence s be denoted $\epsilon(s)$. The true answer to the effect accounting problem in empirical world w is then just $\epsilon(w)$. Call this more abstract problem the *effect accounting problem.* The effect accounting problem subsumes a number of naturally posed inference problems, such as determining the set of independent variables a dependent variable depends upon, determining quantum numbers from a set of reactions (Schulte 2000), and causal inference (Spirtes et al. 2000), in addition to the polynomial inference problem already mentioned.

A *strategy* for the effect accounting responds to an arbitrary input sequence either with a finite set of effects or with '?', indicating a refusal to choose. Strategy σ *solves* the effect accounting problem iff σ converges to the true set of effects $\epsilon(w)$ in each empirical world $w \in K$. One obvious solution to the effect accounting problem is the strategy $\sigma_0(e) = \epsilon(e)$, which guesses exactly the effects it has seen so far. If the possibility of infinitely many effects were admitted, then the effect accounting problem would not be solvable at all, due to a classic result by E. Gold (1978).

Ockham's razor is the principle that one should never output an informative answer unless that answer is among the simplest answers compatible with experience. In the effect accounting problem, there is a uniquely simplest answer compatible with

[2] In statistics, the situation is analogous: increasing the sample size reduces the interval estimates of the values of the function at each argument. The analogy is sketched in greater detail in the conclusion.

[3] In typical statistical applications, something similar is true: effects probably do not appear at each sample size if they don't exist and probably appear at some sample size onward if they do exist. The data model under discussion may be viewed as a logical approximation of the statistical situation, if one thinks of samples accumulating through time.

experience e, namely, the set $\epsilon(e)$ of effects reported so far along e. Thus, strategy σ is *Ockham* at e if and only if σ produces either $\epsilon(e)$ or '?' in response to finite input sequence e.

If the inputs currently received are $e = (e_0, \ldots, e_{n+1})$, then let the previous evidential state be $e_- = (e_0, \ldots, e_n)$ (where e_- is stipulated to denote the empty sequence if e does). Say that solution σ is *stalwart* at e if and only if $\sigma(e) = \sigma_{e_-}$ if $\sigma(e_-) = \epsilon(e)$. The intuition behind stalwartness is that there is no better explanation than the simplest one, so why drop it? One may speak of stalwartness and/or Ockham's razor as being satisfied from e onward (i.e., at each extension e' of e compatible with K) or always (i.e., at each e compatible with K).

The simplicity puzzle now arises because neither Ockham's razor nor stalwartness is necessary for solving the effect accounting problem. For example, one could start with answer $A \neq \emptyset$ and retract back to $\emptyset$ if no effect appears after by stage 1000. Or one could spontaneously retract the set $A \neq \emptyset$ of effects seen so far at stage n even though no new effect has been seen and then return to set A at stage $n + 1$. In either case, one would still converge to the true number of effects in the limit. The trouble is that there are infinitely many ways to solve the effect accounting problem, just as there are infinitely many algorithmic solutions to a solvable computational problem. The nuances of programming practice—the very stuff of textbook computer science—are derived not from solvability, but from efficiency or computational complexity (e.g., the time or storage space required to find the right answer). The proposal is that Ockham's razor is similarly grounded in the efficiency of empirical inquiry.

3 Costs of Inquiry

An obvious, doxastic cost of inquiry is the total number of times one's strategy produces a false answer prior to convergence to the true one, since error is obviously to be avoided if possible. Another is the number of times a conclusion is "taken back" or *retracted* prior to convergence, which corresponds to the degree of "straightness" of the path followed to the truth.[4] One might also wish to minimize the respective times by which these retractions occur, since there is no point "living a lie" longer than necessary or allowing subsidiary conclusions to accumulate prior to being "flushed" when the retraction occurs. Taken together, these statistics concern the accuracy, bumpiness, and timeliness of one's route to the truth. For a given strategy σ and infinite input stream w, let the *loss* or complexity of σ in w be represented by the pair

$$\lambda(\sigma, w) = (q, (r_1, \ldots, r_k)),$$

where q is the total number of errors or false answers output by σ in w, k is the total number of retractions performed by σ in w, and r_i is the stage of inquiry at which the ith retraction occurs.

[4]Retractions are called *mind-changes* in computational learning theory (cf. Jain et al. 1999) and *contractions* in the literature on belief revision (Gärdenfors 1988).

Happily, it turns out that the hard (apples and oranges) comparisions along these dimensions are irrelevant to the argument that follows: one need only consider comparisons in which one cost sequence is as good as or better than another in each dimension. Such comparisons are called *Pareto* comparisions. Accordingly, let $(q, (r_1, \ldots, r_k)) \leq (q', (r'_0, \ldots, r'_{k'}))$ iff $q \leq q'$ and there exists a sub-sequence $(u_0, \ldots, u_k)$ of $(r'_0, \ldots, r'_{k'})$ such that for each i from 1 to k, $r_i \leq u_i$. Then for cost pairs $\mathbf{v}, \mathbf{v}'$, define $\mathbf{v} < \mathbf{v}'$ iff $\mathbf{v} \leq \mathbf{v}'$ but $\mathbf{v}' \leq \mathbf{v}$. With respect to Pareto comparison, each set S of cost pairs has a unique supremum $\sup(S)$, of form $(q, (r_1, \ldots, r_k, \ldots))$, in which q and r_i may assume the first infinite ordinal ω as values.

4 Empirical Complexity and Efficiency

No solution to the effect accounting problem achieves a non-trivial cost bound over the whole problem, since each theory can be overturned by future effects in the arbitrarily remote future. Computational complexity theory (cf. Aho et al.) has long since sidestepped that difficulty by partitioning *problem instances* (inputs) into respective *sizes* and then then examining worst-case resource consumption as instance size increases. In empirical problems, each input stream w has infinite length, but another plausible measure of input stream complexity is the total number $c(s) = |\epsilon(s)|$ of empirical effects presented in w. Then the *conditional empirical complexity* of w *at* e be given by: $c(w, e) = c(w) - c(e)$ and the nth *empirical complexity cell* at e is defined by: $C_e(n) = \{w \in K_e : c(w, e) = n\}$. Let σ be an arbitrary solution to the effect accounting problem. Define the *worst-case loss* of solution σ over complexity class $C_e(n)$ as: $\lambda_e(\sigma, n) = \sup_{w \in C_e(n)} \lambda(\sigma, w)$, where the supremum is understood in the sense of the preceding section.

Suppose that input sequence e has just been received and the question concerns the efficiency of one's strategy σ. Since the past cannot be altered, the only relevant alternatives are strategies that produce the same answers as σ along e_-. Say that such a strategy *agrees with* σ along e_- (abbreviated $\sigma \asymp_{e_-} \sigma'$).

Given solutions σ, σ', the following, natural, worst-case performance comparisons can be defined at e:

$$\begin{aligned} \sigma \leq_e \sigma' \quad &\text{iff} \quad (\forall n)\ \lambda_e(\sigma, n) \leq \lambda_e(\sigma', n); \\ \sigma \prec_e \sigma' \quad &\text{iff} \quad (\forall n)\ C_e(n) \neq \emptyset \Rightarrow \lambda_e(\sigma, n) < \lambda_e(\sigma', n). \end{aligned}$$

These comparisons give rise to two natural concepts, efficiency and being strongly beaten with respect to worst-case cost over empirical complexity classes.

$$\begin{aligned} \sigma \text{ is } \textit{strongly beaten} \text{ at } e \quad &\text{iff} \quad (\exists \text{ solution } \sigma' \asymp_{e_-} \sigma)\ \sigma' \prec_e \sigma; \\ \sigma \text{ is } \textit{efficient} \text{ at } e \quad &\text{iff} \quad (\forall \text{ solution } \sigma' \asymp_{e_-} \sigma)\ \sigma' \geq_e \sigma. \end{aligned}$$

A solution that is strongly beaten does worse than some solution in worst-case performance over *each* non-empty, empirical complexity cell. An efficient solution is as good as an arbitrary solution in worst-case performance over *each* empirical complexity cell.

Since efficiency can be reassessed at each time, one may speak of being efficient from e onward or always. There may also be situations in which one method does better over some complexity cells and worse over others but, remarkably, these messy comparisons are irrelevant to the argument that follows.

5 The New Solution

Here is the proposed efficiency argument for Ockham's razor. The proof is in the appendix.

Theorem 1 (efficient = unbeaten = Ockham + stalwart) *Let the costs be Pareto-comparison of the total number of errors, the total number of retractions, and the respective times of the retractions. Let σ solve the effect accounting problem. Let e be a finite input sequence.*

Then, the following statements are equivalent:

1. *σ is stalwart and Ockham from e onward;*
2. *σ is efficient from e onward;*
3. *σ is never strongly beaten from e onward.*

So the set of all solutions to the effect accounting problem is cleanly partitioned at e into two groups: the solutions that are stalwart, Ockham, and efficient from e onward and the solutions that are strongly beaten at some stage $e' \geq e$ due to future violations of the stalwart, Ockham property. As promised, the argument is *a priori*, normative, truth-directed, and yet non-circular. The argument presumes no prior bias of any kind, so there is no question of a circular appeal to a simplicity bias, as in Bayesian arguments. The argument is driven only by efficient convergence to the truth, so there is no bait-and-switch from truth-finding to some other aim. There is no confusion between "confirmation" and truth-finding, since the concept of confirmation is never mentioned. There is no wishful presumption that the truth must be testable or nice in any other way. There is no appeal to the hidden hands of Providence or evolution.

Furthermore, the argument in favor of Ockham's razor is *diachronically stable* in the sense that it always makes sense to return to the Ockham fold no matter how many times you violated Ockham's razor in the past. Not only do you become efficient as soon as you return to the stalwart, Ockham fold—you are strongly beaten each time you stray, no matter what you have done in the past, so the entire argument is stable in spite of past deviations. That is important, for Ockham violations are practically unavoidable in real science because the simplest theory cannot always be formulated in time to forestall acceptance of a more easily conceived but more complex alternative (e.g., Ptolemaic astronomy *vs.* Copernican astronomy, Newtonian optics *vs.* wave optics, Newtonian kinematics *vs.* relativistic kinematics, and special creation *vs.* natural selection). So although it has been urged that scientific revolutions are extra-rational events governed only by the vagaries of scientific politics (Kuhn 1975),

revision to the simpler theory when it is discovered has a clean explanation in terms of truth-finding efficiency.

The preceding result is proved only for problems structurally identical to the effect counting problem. One can also define problems very generally and then define simplicity in terms of problem structure so that simpler worlds are worlds in which nature loses fewer opportunities to force retractions of answers from an arbitrary, convergent method. Simpler answers are satisfied by simpler worlds and Ockham's razor requires, plausibly, that one never produce an answer that is not among the simplest compatible with experience. According to this approach, simplicity does *not* depend upon mere description (it is invariant under grue-like translations of the inputs provided to the method). It is then possible to show a more general version of theorem 1 once for all for a broad class of inference problems (Kelly 2006).

The proposed solution to the simplicity puzzle does not accomplish the impossible: Ockham's razor cannot be shown, without circularity, to point at the truth immediately, like an occult divining rod; there is not even an *a priori* bound on the number of times an Ockham method might produce the wrong answer or reverse its conclusion in the worst case. The justification for using a method with such weak properties is, however, straightforward and compelling: Ockham's razor is demonstrably the uniquely most efficient strategy and the best possible strategy had best be good enough. That is the routine argument for algorithms generally in computer science and all that has been done is to extend that form of argument to the vindication of Ockham's razor. Of course, weak efficiency arguments of this sort can readily be overturned by genuine background knowledge concerning the nature of the world one faces. The puzzle about simplicity is not to incorporate existing knowledge into rational choices—Bayesian update does that, after a fashion. The puzzle is to justify our default preference for simplicity when such knowledge is entirely lacking, and that is just what the argument does. By way of loose analogy, the argument can be thought of as doing the same thing for Ockham's razor that Dutch book arguments do for coherence, itself. In both cases, more structural recommendations are invoked to vindicate aspects of Bayesian reasoning when the aspects in question are laid aside, temporarily, for the sake of the argument.

One might hope for a guarantee stronger than minimization of worst-case, accrued doxastic costs, but the obvious candidates fail. (1) One cannot establish weak dominance for Ockham methods with respect to all problem instances jointly, because anticipation of unseen effects might be vindicated immediately, saving retractions that the Ockham method would have to perform when the effects appear. (2) Nor can one show that Ockham's razor does best in terms of a global worst-case bound over all problem instances (minimax theory), for such worst-case bounds on errors and retractions are trivially infinite for all methods at every stage. (3) Nor can one show a decisive advantage for Ockham's razor in terms of expected retractions. For example, if the question is whether one will see at least one effect, then the expected retractions of the obvious strategy $\sigma(e) = \epsilon(e)$ are less than those of an arbitrary Ockham violator only if the prior probability of the simpler answer is at least one half, so that if more than one complex world carries nonzero probability, no complex world is as probable as the

simplest world, which begs the question in favor of simplicity.[5] If the prior probability of the simple hypothesis drops below 0.5, the advantage lies not only with violating Ockham's razor, but with violating it more rather than less. So Bayesians must either beg the question or rule strongly against Ockham.

6 References

Aho, A., Hopcroft, J., and Ullman, J. (1974). *The Design and Analysis of Computer Algorithms*. New York: Addison-Wesley.

Daley, R. and Smith, C. (1986) "On the complexity of inductive inference", *Information and Control* 69: pp. 12-40.

Duda, R., Stork, D., and Hart, P. (2000). *Pattern Classification*, 2nd. ed., v. 1. New York: Wiley.

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

Forster, M. and Sober, E. (1994). How to Tell When Simpler, More Unified, or Less Ad Hoc Theories will Provide More Accurate Predictions. *The British Journal for the Philosophy of Science* 45: 1 - 35.

Gärdenfors, P. (1988). *Knowledge in Flux*. Cambridge: M.I.T. Press.

Garey, M. and Johnson, D. (1979). *Computers and Intractability*, New York: Freeman.

Gold, E. (1978). "Language identification in the limit", *Information and Control* 10: 447 - 474.

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

Glymour, C. (1980). Theory and Evidence, Princeton: Princeton University Press.

Harman, G. (1965). The Inference to the Best Explanation, *Phil Review* 74: 88-95.

[5]Let σ_i be a non-Ockham strategy that starts by guessing answer ≥ 1 until no effect is seen by stage i, at which point σ_i returns 0. If the effect is ever seen, σ returns answer ≥ 1. Consider the competing, Ockham method σ that always guesses 0 until the effect is seen, at which time σ returns answer ≥ 1. Consider probabilities at stage 0. Let a denote the probability that no effect occurs, let b denote the probability that an effect occurs no later than stag i and let c denote the probability that an effect occurs after stage i. Then, *a priori*, the expected retractions σ_i are given by $a + 2c$, whereas the expected retractions of σ are $b + c$. So the Ockham strategy σ does better when $a + c > b$. Since $a + c + b = 1$, this is true if and only if $b < .5$. By increasing i, one can drive c arbitrarily small (by countable additivity), so if the Ockham strategy is to beat the expected retractions of an arbitrary σ_i, then $a \geq b$. That implies that each of the several (complex) possibilities over which mass b is distributed receives less probability than the simple world carrying probability b. This bias increases with i and with the number of ways the complex theory can be true.

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.

Kelly, K. (2002). "Efficient Convergence Implies Ockham's Razor", *Proceedings of the 2002 International Workshop on Computational Models of Scientific Reasoning and Applications*, Las Vegas, USA, June 24-27.

Kelly, K. (2004a) "Justification as Truth-finding Efficiency: How Ockham's Razor Works", *Minds and Machines* 14: pp. 485-505.

Kelly, K. (2004b)"Uncomputability: The Problem of Induction Internalized," *Theoretical Computer Science* 317: pp. 227-249.

Kelly, K. (2006). "How Simplicity Helps You Find the Truth Without Pointing at it", forthcoming, V. Harazinov, M. Friend, and N. Goethe, *Philosophy of Mathematics and Induction*, Dordrecht: Springer.

Kelly, K. and Glymour, C. (2004). "Why Probability Does Not Capture the Logic of Scientific Justification", C. Hitchcock, ed., *Contemporary Debates in the Philosophy of Science*, Oxford: Blackwell, pp. 94-114.

Leibniz, G. W. (1714) *Monadologie*, in *Die Philosophischen Schriften von G. W. Leibniz*, vol. IV. Berlin: C. J. Gerhardt, 1875, pp. 607-23.

Kuhn, T. (1970). *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press.

Mitchell, T. (1997). *Machine Learning.* New York: McGraw-Hill.

Popper, K. (1968). *The Logic of Scientific Discovery*, New York: Harper.

Rosenkrantz (1983). "Why Glymour is a Bayesian", in *Testing Scientific Theories*, Minneapolis: University of Minnesota Press.

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

Schulte, O. (2000). "Inferring Conservation Principles in Particle Physics: A Case Study in the Problem of Induction", *The British Journal for the Philosophy of Science* , 51: 771-806.

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

Spirtes, P., and Zhang, J. (2003). "Strong Faithfulness and Uniform Consistency in Causal Inference", *Proceedings of the 19th Conference in Uncertainty in Artificial Intelligence*, August 7-10 2003, Acapulco, Mexico. San Mateo: Morgan Kaufmann. pp. 632-639.

van Fraassen, B. (1981). *The Scientific Image*. Clarendon Press: Oxford.

Vitanyi, P. and Li, M. (2000) "Minimum Description Length Induction, Bayeisanism, and Kolmogorov Complexity", *IEEE Transactions on Information Theory* 46: 446-464.

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. New York: Springer.

7 Appendix

Proof of theorem 1. $(2 \Rightarrow 3)$, is immediate from the definitions. For $(3 \Rightarrow 1)$, suppose that σ violates Ockham's razor or stalwartness at finite input sequence e. Let σ be a solution that is stalwart and Ockham from e' onward. Let $e \geq e'$ have length j. Then σ is Ockham and stalwart from e onward. Let σ' be an arbitrary solution such that $\sigma' \asymp_{e_-} \sigma$. Let $r_1, \ldots, r_k$ be the retraction times for both σ and σ' along e_-. Let q denote the number of times σ produces an answer other than $\epsilon(e)$ along e_-. Let $w \in C_e(0)$. In w, σ never retracts after e (but may do so at e) and σ produces only the true answer $\epsilon(e)$ after e. Hence:

$$\lambda_e(\sigma, 0) \leq (q, (r_1, \ldots, r_k, j)).$$

Consider the hard case in which σ retracts at e. There exists $w_0 \in C_e(0)$ (just extend e by repeating $\epsilon(e)$ forever). Then $\sigma(e_-) = \sigma'(e_-)$ is false in w_0. So since σ' is a solution, σ' converges to the true answer $\epsilon(e)$ in w_0 at some point after e_-, which implies a retraction at some point no sooner than e. Hence:

$$\lambda_e(\sigma', 0) \geq (q, (r_1, \ldots, r_k, j)) \geq \lambda_e(\sigma, 0).$$

If $C_e(n+1) = \emptyset$, then every method succeeds under the trivial bound $(0, ())$, so suppose that $C_e(n+1) \neq \emptyset$. Since σ is a stalwart, Ockham solution, σ retracts at most once at each new effect, so

$$\lambda_e(\sigma, n+1) \leq (\omega, (r_1, \ldots, r_k, j, \underbrace{\omega, \ldots, \omega}_{n+1 \text{ times}})).$$

Let arbitrary natural number i be given. Since σ' is a solution, σ' eventually converges to $A_0 = \epsilon(e)$ in w_0, so there exists e_0 such that $e \leq e_0 < w_0$ by which σ' has retracted the false answer $\sigma'(e_-)$ and has produced the true answer A_0 successively at least i times after the end of e, so σ' retracts at least as late as e in e_0. Then there exists $w_1 \in C_e(1)$ such that $e_0 < w_1$ (since $C_e(n+1) \neq \emptyset$, nature can choose some $x_0 \in E - A_0$ and extend e_0 forever with answer $A_1 = A_0 \cup \{x_0\}$). Again, σ' must converge to A_1 in w_1 and, therefore, produces A_1 successively at least i times by some initial segment e_1 of w that extends e_0. Continuing in this manner, construct $w_{n+1} \in C_e(n+1)$. Then

$$\lambda_e(\sigma', w_{n+1}) \geq (i, (r_1, \ldots, r_k, j, j+1i, j+2i, \ldots, j+(n+1)i)).$$

Since i is arbitrary,

$$\lambda_e(\sigma', n+1) \geq (\omega, (r_1, \ldots, r_k, j, \underbrace{\omega, \ldots, \omega}_{n+1 \text{ times}})) \geq \lambda_e(\sigma, n+1).$$

Now consider the easy case in which σ does not retract at e. Then the argument is similar to that in the preceding case except that the retraction at j is dropped from all the bounds.

For the proof of $(1 \Rightarrow 2)$, let σ be a solution that violates either Ockham's razor or stalwartness at e of length j. Let σ' return $\epsilon(e')$ at each $e' \in K_{\text{fin}}$ such that $e' \geq e$ and let σ' agree with σ otherwise. Then $\sigma' \asymp_{e_-} \sigma$ by construction and σ' is evidently a solution. Let $r_1, \ldots, r_k$ be the retraction times for both σ and σ' along e up to but not including the last entry in e.

Consider the case in which σ violates Ockham's razor at e. So for some $A \subseteq E$, $\sigma(e) = A \neq \epsilon(e)$. Let $w \in C_e(0)$. Then A is false in w and $\epsilon(e)$ is true in w. Let q denote the number of times both σ and σ' produce an answer other than $\epsilon(e)$ along e_-. Since σ' produces the true answer at e in w and continues to produce it thereafter:

$$\lambda_e(\sigma', 0) \leq (q, (r_1, \ldots, r_k, j)).$$

There exists w_0 in $C_e(0)$ (just extend e forever with $\epsilon(e)$). Since A is false in w_0 and σ is a solution, σ retracts A in w_0 at some stage greater than j, so

$$\lambda_e(\sigma, 0) \geq \lambda(\sigma, w_0) \geq (q+1, (r_1, \ldots, r_k, j+1)) > \lambda_e(\sigma', 0).$$

As in the proof of $(3 \Rightarrow 1)$, it suffices to consider the case in which $C_e(n+1) \neq \emptyset$. Since σ' produces $\epsilon(e')$ at each $e' \geq e$,

$$\lambda_e(\sigma', n+1) \leq (\omega, (r_1, \ldots, r_k, j, \underbrace{\omega, \ldots, \omega}_{n+1 \text{ times}})).$$

Let $i \in \omega$. Answer $A = \sigma(e)$ is false in w_0, so since σ is a solution, σ eventually converges to $A_0 = \epsilon(e)$ in w_0, so there exists e_0 properly extending e by which σ has produced A_0 successively at least i times after the end of e and σ retracts A back to A_0 no sooner than stage $j+1$. Now continue according to the recipe described in the proof of $(3 \Rightarrow 1)$ to construct $w_{n+1} \in C_e(n+1)$ such that:

$$\lambda(\sigma, w_{n+1}) \geq (i, (r_1, \ldots, r_k, j+1, j+1i, j+2i, \ldots, j+(n+1)i)).$$

Since i is arbitrary,

$$\lambda_e(\sigma, n+1) \geq (\omega, (r_1, \ldots, r_k, j+1, \underbrace{\omega, \ldots, \omega}_{n+1 \text{ times}})) > \lambda_e(\sigma', n+1).$$

Next, consider the case in which σ violates stalwartness at e. So $\sigma(e_-) = \epsilon(e)$ but $\sigma(e) \neq \epsilon(e)$. Let $w \in C_e(0)$. Let q denote the number of errors committed in w by

both σ and σ' along e_-. Since $\sigma'(e_-) = \epsilon(e)$, it follows that σ' does not retract in w from j onward, so:

$$\lambda_e(\sigma', 0) \leq (q, (r_1, \ldots, r_k)).$$

Again, there exists w_0 in $C_e(0)$. Since σ retracts at j,

$$\lambda_e(\sigma, 0) \geq (q, (r_1, \ldots, r_k, j)) > \lambda_e(\sigma', 0).$$

Let $C_e(n+1) \neq \emptyset$. Since σ' produces $\epsilon(e')$ at each $e' \geq e$,

$$\lambda_e(\sigma', n+1) \leq (\omega, (r_1, \ldots, r_k, \underbrace{\omega, \ldots, \omega}_{n+1 \text{ times}})).$$

Let arbitrary natural number i be given. Since σ retracts at j, one may continue according to the recipe described in the proof of proposition **??** to construct w_{n+1} extending e in $C_e(n+1)$ such that:

$$\lambda(\sigma, w_{n+1}) \geq (i, (r_1, \ldots, r_k, j, j+1i, j+2i, \ldots, j+(n+1)i)).$$

Since i is arbitrary,

$$\lambda_e(\sigma, n+1) \geq (\omega, (r_1, \ldots, r_k, j, \underbrace{\omega, \ldots, \omega}_{n+1 \text{ times}})) > \lambda_e(\sigma', n+1).$$

$\dashv$