Ockham's Razor, Empirical Complexity, and Truth-finding Efficiency

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

May 9, 2007

Abstract

The nature of empirical simplicity and its relationship to scientific truth are long-standing puzzles. In this paper, empirical simplicity is explicated in terms of empirical effects, which are defined in terms of the structure of the inference problem addressed. Problem instances are classified according to the number of empirical effects they present. Simple answers are satisfied by simple worlds. An efficient solution achieves optimum worst-case cost over each complexity class with respect to such costs such as the number of retractions or errors prior to convergence and elapsed time to convergence. It is shown that always choosing the simplest theory compatible with experience and hanging onto it while it remains simplest is both necessary and sufficient for efficiency.

1 The Simplicity Puzzle

Machine learning, statistics, and the philosophy of science 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, whose work exemplified a similar tendency. But in spite of its intuitive appeal, how could Ockham's razor possibly help us find the true theory? For if we already know that the simplest theory is true or probably true, we don't need Ockham's razor to infer that it is. And if we don't know that the simplest theory is true or probably true, how do we know that simplicity steers us in the right direction?

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), minimization of predictive risk (Akaike 1973), or minimum description length (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).

Over-fitting arguments based upon minimization of predictive risk might seem to be an exception (cf. Wasserman 2004, Mitchell 1997, Forster and Sober 1994). Such arguments show that using a complex model for predictive purposes in the presence of random noise can increase the expected deviation from the truth of estimates based upon the complex model.[1] But that is no less true when you know in advance that the truth is complex, so over-fitting arguments concern accuracy of prediction rather than theoretical truth.[2]

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.[3] Simulation studies, in which simplicity-biased methods are applied to problems randomly generated according to a simplicity-biased sampling distribution over problem instances, are circular in precisely the same way.

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 philosophers and even some machine learning texts have replaced Providence with an equally benevolent, evolutionary etiology.

> Why should one prefer simpler hypotheses? ... If the species of agents employs a learning algorithm whose inductive bias is Occam's razor, then we expect evolution to produce internal representations for which Occam's razor is a successful strategy. The essence of the argument here is that evolution will create internal representations that make the learning algorithm's inductive bias a self-fulfilling prophecy, simply because it can alter the representation easier than it can alter the learning algorithm (Mitchell 1997, p. 66; cf. also Duda et al. 2000, pp. 464-465).

[1] Deviation is measured, for example, in terms of Kullback-Leibler divergence and the estimates based upon the model are maximum likelihood estimates. It false that the standard methods of model selection in any sense minimize worst-case expected distance of maximum likelihood estimates from the truth; they merely minimize an unbiased, empirical estimate of that distance.

[2] Often, the former is practically "good enough", but not always, as when one is learning how to intervene in a causal system from non-experimental data (Spirtes et al. 2000). In such applications, causal theories with arbitrarily good predictive ability can lead to arbitrarily bad policy recommendations because arbitrarily small correlations can indicate reversals in the causal order among variables (Spirtes and Zhang 2003).

[3] Some standard Bayesian prior probabilities in machine learning (e.g., the prior probability whose posterior is approximated by the Bayes' information criterion or BIC; cf. Wasserman 2004.) impose equal prior probabilities on all answers, whether they be simple or complex. But imposing equal probabilities on answers implies a strong bias against complex possible worlds (parameter settings), since the parameter space of the complex theory has a higher dimension than that of the simple theory and the distribution over paramters is usually assumed to be flat. If the simple theory happens to explain the data accurately, then the complex theory does so only over a narrow interval of its parameters, which carries very low prior probability compared to the simple theory because the prior probability of the complex theory is spread uniformly over its possible parameter values. Updating the probabilities of the two theories by Bayes' theorem simply passes along the prior bias in favor of the best-fitting settings of the parameters in the simple theory over the best-fitting settings of the parameters in the complex theory, so the argument is still circular when both theories are assigned equal prior probabilities.

Maybe. But even if the adaptationist story is true, it explains 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 uniformity (simplicity) of experience and, hence, to Ockham's razor, which is another example of circular reasoning.

Circularity aside, it is awkward for computer scientists to find themselves pressed to the extremity of paleontological arguments in behalf of the most recent machine learning procedures. One routinely expects computer scientists to justify a procedure by demonstrating its efficiency at finding the right answer, whatever the right answer might be (e.g., Aho et. al. 1974, Garey and Johnson 1979). Even if Providence or evolution does 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, mathematical 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 paper presents such an argument.[4] 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 be hoped to achieve 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 is shown to do. The argument is presented first for a particular example and is then generalized to a broad class of possible inference problems.

2 Empirical Effects

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

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

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.[5] 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

[4]The approach is based on concepts from computational learning theory. For a survey of related ideas, cf. (Jain et al., 1999) and (Kelly 1996).

[5]The idea is to approximate the usual situation in statistics: increasing the sample size tightens interval estimates of the values of the function at each argument. Here, the shrinking intervals are guaranteed to catch the truth exactly, rather than just with high probability. The analogy is sketched in greater detail in the conclusion.

open intervals to tilt the line. So suppose that the truth is a 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 finite amount of data collected for the linear theory is compatible (due to the remaining wiggle room) with a quadratic law. Further data collected from the quadratic law eventually refutes linearity (a second-order effect), and so forth. 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. 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 stream* is an infinite, increasing, sequence of finite subsets of E that converges to some finite subset of E. Let K denote the set of all such worlds. If $w \in K$, then let $w|n$ denote the initial segment of w of length n. Then the set of all finite input sequences compatible with K is given by:

$$K_{\text{fin}} = \{w|n : w \in K \text{ and } n \in \omega\}.$$

Given $s, s' \in K \cup K_{\text{fin}}$, let $s \leq s'$ indicate that s is an initial segment of s' and let $s \subseteq s'$ indicate that s is a sub-sequence of s' (not necessarily an initial segment). Then $<$ and $\subset$ denote the strict versions of these respective relations. If $e \in K_{\text{fin}}$, then the set of all worlds in K compatible with e is given by:

$$K_e = \{w \in K : e < w\}.$$

Let $s \in K \cup K_{\text{fin}}$. Then let the *effects presented* in s be defined by:

$$\epsilon(s) = \bigcup_{i \in \text{dom}(s)} s(i).$$

Let T_A denote the proposition that A is the set of all effects that one will ever see:

$$T_A = \{w \in K : \epsilon(w) = A\}.$$

Then the *true answer* to the effect accounting problem in world w is then just $T_{\epsilon(w)}$ and the *potential answers* are given by the partition of K:

$$\Pi = \{T_{\epsilon(w)} : w \in K\}.$$

Let the *effect accounting problem* be the pair (K, Π), where K is the problem's *presupposition* and Π is the *question* posed by the problem. The effect accounting problem captures the essential character of a number of natural 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 problem maps finite input sequences in K_{fin} to potential answers or to '?', which indicates refusal to choose an answer. Strategy σ *solves* the effect accounting problem iff[6] for each $w \in K$,

$$\lim_{i \to \infty} \sigma(w|i) = T_{\epsilon(w)}.$$

One obvious solution to the effect accounting problem is the strategy $\sigma_0(e) = T_{\epsilon(e)}$, 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 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, it seems that there is a uniquely simplest answer compatible with experience e, namely, $T_{\epsilon(e)}$, where $\epsilon(e)$ is the set of all effects presented along e. Thus, strategy σ is *Ockham* at e iff for each $e \in K_{\text{fin}}$:

$$\sigma(e) = T_{\epsilon(e)} \text{ or } \sigma(e) = \text{'?'}.$$

As stated, Ockham's razor is compatible with suspension of judgment at any time. 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 is empty). Say that solution σ is *stalwart* at e iff for each $e \in K_{\text{fin}}$:

$$\sigma(e) = T_{\epsilon(e)} \text{ if } \sigma(e_-) = T_{\epsilon(e)}.$$

The intuition behind stalwartness is that there is no better explanation than the simplest one, so why drop it after selecting it?[7] 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 $e \in K_{\text{fin}}$).

The obvious solution $\sigma_0(e) = T_{\epsilon(e)}$ is both stalwart and Ockham. A more plausible sort of stalwart, Ockham solution suspends judgment for some time before plumping for the simplest answer and then hangs onto it until it is dethroned by experience. But as obvious as such strategies seem, neither Ockham's razor nor stalwartness is necessary

[6] i.e, if and only if

[7] Stalwartness can be counter-intuitive if the problem under consideration is modeled too coarsely. Suppose that you are watching a cannon factory that may produce a cannon that may produce a shot. After watching the factory fail to produce a cannon for a long time, you might reasonably come to conclude that it will never succeed, so you will never see one of its cannons produce a shot. But then when the factory finally rolls out its first cannon, you might very well retract your conclusion that you will never see one of their cannons shoot, since you have no prior experience with cannon built by this factory. If the rolling out of the cannon and the shot of the cannon are both effects, then the loss of confidence when the cannon appears satisfies stalwartness. If only the shot is an effect, however, then the loss of confidence violates stalwartness.

for solving the effect accounting problem. For example, one could start with answer T_A where $A \neq \emptyset$ and retract back to $\emptyset$ if no effect appears after by stage 1000. Or one could spontaneously retract $T_{\epsilon(e_-)}$ at e even though no new effect has been seen at e and then return to set $T_{\epsilon(e)}$ thereafter. In either case, one would still converge to the true answer 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, developed in detail below, is that the nuances of scientific method are similarly grounded in the efficiency of empirical inquiry.

3 Costs of Inquiry

Consider some plausible measures of the complexity or cost of converging to the true answer to the effect accounting problem. An obvious cost is the number of times the 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.[8] The uninformative output '?' is not, properly speaking, a conclusion, so dropping '?' in favor of a conclusion does not count as a retraction. 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 and provide a fair picture of the overall quality or efficiency of inquiry so far as finding the truth is concerned. 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.

Minimizing these costs jointly can occasion some hard choices (e.g., vacillating between answer A and '?' for two hundred times results in one hundred more retractions than sticking with A for two hundred times, but may also commit one hundred more errors if A is false). Happily, it turns out that the hard choices are irrelevant to

[8] H. Putnam (1965) proposed and analyzed the idea of bounding retractions in inference. Retractions are called *mind-changes* in computational learning theory (cf. Jain et al. 1999) and *contractions* in the literature on belief revision (Gärdenfors 1988). The former literature has focused on problem complexity rather than upon justifying particular strategies like Ockham's razor (Freivalds and Smith 1993) and the latter on minimizing the amount of information retracted in one step in the face of inconsistency. A general perspective on complexity measures for inductive inference which includes mind-changes as a special case is developed in (Daley and Smith 1986).

the argument that follows: the only comparisons that matter are thos in which one cost sequence is as good as or better than another in every respect. Accordingly, let $(q, (r_1, \ldots, r_k)) \leq (q', (r'_0, \ldots, r'_{k'}))$ iff

1. $q \leq q'$ and
2. 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 $\mathbf{v}' \leq \mathbf{v}$. Relation $\leq$ may be referred to as *Pareto comparison* and relation $<$ is called *weak Pareto dominance* with respect to retractions, retraction times, and errors. So to rephrase the point, the only comparisons that will be made about convergent performance of empirical strategies are the uncontroversial, Pareto comparisons among cost pairs.

The next step is to define and to compare worst-case bounds on sets of cost vectors. Let ω denote the first infinite ordinal number. A *potential cost bound* is a pair (q, γ) where $q \leq \omega$ and γ is a finite or infinite, non-descending sequence of entries $\leq \omega$ in which no finite entry occurs more than once. Then if $\mathbf{v}$ is a cost vector and $\mathbf{b}$ is a cost bound, $\mathbf{v} \leq \mathbf{b}$ can be defined just as for cost vectors, themselves. Cost bounds $\mathbf{b}, \mathbf{c}$ may now be compared as follows:

$$\begin{aligned}
\mathbf{b} \leq \mathbf{c} &\quad \text{iff} \quad \text{for each cost pair } \mathbf{v}, \text{ if } \mathbf{v} \leq \mathbf{b} \text{ then } \mathbf{v} \leq \mathbf{c};\\
\mathbf{b} \equiv \mathbf{c} &\quad \text{iff} \quad \mathbf{b} \leq \mathbf{c} \text{ and } \mathbf{c} \leq \mathbf{b};\\
\mathbf{b} < \mathbf{c} &\quad \text{iff} \quad \mathbf{b} \leq \mathbf{c} \text{ and } \mathbf{c} \not\leq \mathbf{b}.
\end{aligned}$$

It follows, for example, that $(4, (2)) < (\omega, (2, \omega)) < (\omega, (0, 1, 2, \ldots)) \equiv (\omega, (\omega, \omega, \omega, \ldots))$.

Now each set C of cost vectors has a unique (up to equivalence) least upper bound $\sup(C)$ among the potential upper bounds, computed as follows. Define:

$$\begin{aligned}
\mathrm{Err}_C &= \{c : \text{ there exists } \tau \text{ such that } (c, \tau) \in C\};\\
\mathrm{Ret}_C &= \{\tau : \text{ there exists } c \text{ such that } (c, \tau) \in C\}.
\end{aligned}$$

Then

$$\sup(C) = (\sup(\mathrm{Err}_C), \sup(\mathrm{Ret}_C)),$$

where $\sup(\mathrm{Ret}_C)$ is defined as follows. Let R be an arbitrary set of ascending, finite sequences of natural numbers. If there is no finite bound on the lengths of the sequences in R, then, evidently, there is no finite γ such that for each $\tau \in R$, $\tau \leq \gamma$. Hence, $\sup(R)$ must have infinite length. But for each pair $(q, \gamma), (q, \gamma')$ of potential bounds such that γ, γ' are both infinite, $(q, \gamma) \equiv (q, \gamma')$, so one may represent all such bounds by (q, ∞). If, on the other hand, there is a maximum length m on the length of the sequences in R, then $\sup(R)$ can be constructed in the following way: list the (countably many) elements of R in a right-justified column and then take the supremum $\leq \omega$ in each column to arrive at sequence $\sup(R)$ whose length is m.

Suppose that finite input sequence e has already been seen. Then one knows that possibilities incompatible with ϵ cannot happen, so define the *worst-case cost* of σ given e as:

$$\lambda_e(\sigma) = \sup_{w \in K_e} \lambda(\sigma, w).$$

4 Empirical Complexity and Efficiency

As in typical, computational problems, 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 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. Since there are only finitely many inputs of each size, worst case bounds exist necessarily for each size and these can be compared across algorithms. In the case of empirical problems, the input from the environment is potentially infinite, so input length no longer distinguishes among problem instances, but in this case there is another plausible measure of instance complexity, namely, the total number of effects presented. If $s \in K \cup K_{\text{fin}}$, then let the *empirical complexity* of s be given by

$$c(s) = |\epsilon(s)|.$$

If $w \in K$ and $e \in K_{\text{fin}}$, let the *conditional empirical complexity* of w *at* e be given by:

$$c(w, e) = c(w) - c(e).$$

Then the nth *empirical complexity class* 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 you have been following strategy σ and that the input sequence you have received so far is e and that you have not yet produced an output in response to e. Then since the past cannot be altered, there is no point considering problem instances incompatible with e or alternative strategies that do not agree with σ along e_-. Accordingly, say that σ' *agrees with* σ along e_- (abbreviated $\sigma \asymp_{e_-} \sigma'$) iff for each $e' < e$, $\sigma(e') = \sigma'(e')$.

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

$$\sigma \leq_e \sigma' \quad \text{iff} \quad (\forall n)\ \lambda_e(\sigma, n) \leq \lambda_e(\sigma', n);$$

$$
\begin{aligned}
\sigma \prec_e \sigma' \quad &\text{iff} \quad (\forall n)\ C_e(n) \neq \emptyset \Rightarrow \lambda_e(\sigma, n) < \lambda_e(\sigma', n);\\
\sigma <_e \sigma' \quad &\text{iff} \quad \sigma \leq_e \sigma' \text{ and } \sigma' \not\leq_e \sigma;\\
\sigma \equiv_e \sigma' \quad &\text{iff} \quad \sigma \leq_e \sigma' \text{ and } \sigma' \leq_e \sigma;\\
\sigma \amalg_e \sigma' \quad &\text{iff} \quad \sigma \not\leq_e \sigma' \text{ and } \sigma' \not\leq_e \sigma.
\end{aligned}
$$

These comparisons give rise to some natural, worst-case concepts of efficiency and inefficiency:

$$
\begin{aligned}
\sigma \text{ is } \textit{strongly beaten} \text{ at } e \quad &\text{iff} \quad (\exists \text{ solution } \sigma' \succeq_{e\_} \sigma)\ \sigma' \prec_e \sigma;\\
\sigma \text{ is } \textit{weakly beaten} \text{ at } e \quad &\text{iff} \quad (\exists \text{ solution } \sigma' \succeq_{e\_} \sigma)\ \sigma' <_e \sigma;\\
\sigma \text{ is } \textit{unbeaten} \text{ at } e \quad &\text{iff} \quad (\forall \text{ solution } \sigma' \succeq_{e\_} \sigma)\ \sigma' \not<_e \sigma;\\
\sigma \text{ is } \textit{efficient} \text{ at } e \quad &\text{iff} \quad (\forall \text{ solution } \sigma' \succeq_{e\_} \sigma)\ \sigma' \geq_e \sigma.
\end{aligned}
$$

The preceding properties are ordered from the least desirable to the most desirable. A solution that is strongly beaten does worse than some solution over each problem instance size. A solution that is weakly beaten does as poorly as some solution over each problem instance size and worse over some problem instance size. A solution is unbeaten if it is as good as or merely incomparable with an arbitrary solution. Efficiency, on the other hand, is a rather strong recommendation, so far as worst-case recommendations go, for an efficient solution is as good as an arbitrary solution in each problem instance size. Since efficiency can be reassessed at each time, one may speak of being efficient from e onward or always.

5 Efficiency Implies Ockham's Razor

Now it is easy to argue for the efficiency of stalwart, Ockham solutions to the effect accounting problem. In spite of all the hard choices between errors and retractions that particular pairs of solutions might occasion, each stalwart, Ockham solution is as good as an arbitrary solution in the worst case over each problem instance size. Indeed, something stronger is true: it is never too late for prodigal methods to repent and return to the Ockham fold, for born-again Ockham solutions are efficient from their moment of rebirth onward. One may say, therefore, that stalwart, Ockham solutions are not merely efficient, but *stably* efficient, in the sense that past deviations from efficiency (for whatever reason) do not undermine the efficiency argument.

Proposition 1 (efficiency of stalwart, Ockham solutions) *Let the cost be Pareto-comparison of error, retractions, and retraction times and let σ be a solution to the effect accounting problem that is stalwart and Ockham from finite input sequence e onward. Then σ is an efficient solution to the effect accounting problem from e onward.*

Proof. Let σ be a solution that is stalwart and Ockham from e' onward. Let $e \geq e'$ have length j. Let σ' be an arbitrary solution such that $\sigma' \succeq_{e\_} \sigma$. 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 $T_{\epsilon(e)}$ along e_-. Let $w \in C_e(0)$.

Consider the hard case in which σ retracts at e. 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:

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

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:

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

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 arbitrarily large natural number i be given and let $A_0 = \epsilon(e)$. Since σ' is a solution, σ' eventually converges to T_{A_0} 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 T_{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 the set of effects $A_1 = A_0 \cup \{x_0\}$). Then σ' must converge to T_{A_1} in w_1 and, therefore, produces T_{A_1} successively at least i times after the end of e_0 along some initial segment e_1 of w. 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,

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

In the easy case in which σ does not retract at e, the argument is similar to that in the preceding case except that the retraction at j is deleted from all the bounds. $\dashv$

Efficiency is only half of the argument for Ockham's razor. The other half is that non-Ockham solutions and non-stalwart solutions are not merely inefficient, but are *strongly beaten* at *each* violation of *either* principle. So not only do you become efficient as soon as you return, permanently, 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.

Proposition 2 (efficiency implies the stalwart, Ockham property) *Let the costs be as in proposition 1 and let σ solve the effect accounting problem. If σ is not Ockham or is not stalwart at e, then σ is strongly beaten at e.*[9]

Proof. 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 $A \subseteq E$ such that $A \neq \epsilon(e)$, $\sigma(e) = T_A$. Let $w \in C_e(0)$. Then T_A is false in w and $T_{\epsilon(e)}$ is true in w. Let q denote the number of times both σ and σ' produce an answer other than $T_{\epsilon(e)}$ along e_-. Since σ' produces the true answer $T_{\epsilon(e)}$ at e in w and continues to produce $T_{\epsilon(e)}$ 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 T_A is false in w_0 and σ is a solution, σ retracts T_A in w_0 at some stage greater than j, so

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

Suppose that $C_e(n+1) \neq \emptyset$. Since σ' produces $T_{\epsilon(e')}$ 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 T_A is false in w_0, so since σ is a solution, σ eventually converges to T_{A_0} such that $A_0 = \epsilon(e)$ in w_0, so there exists e_0 properly extending e by which σ has produced T_{A_0} successively at least i times after the end of e and σ revises T_A to T_{A_0} no sooner than stage $j+1$. Now continue according to the recipe described in the proof of proposition 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,

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

[9]Proposition 2 remains true if total elapsed time to convergence to the truth (i.e., *modulus of convergence*) is included among the other costs. Proposition 1 does not, since convergence time efficiency demands, rather counter-intuitively, that one leap for the uniquely simplest answer as soon as it exists. Therefore, a more intuitive theory results if convergence time is eliminated from the list of costs.

Next, consider the case in which σ violates stalwartness at e. So $\sigma(e_-) = T_{\epsilon(e)}$ but $\sigma(e) \neq T_{\epsilon(e)}$. Let $w \in C_e(0)$. Let q denote the number of errors committed in w by both σ and σ' along e_-. Since $\sigma'(e_-) = T_{\epsilon(e)}$, 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 for each $e' \geq e$, σ' produces $T_{\epsilon(e')}$ at 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 1 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,

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

$\dashv$

The proof of proposition 2 entails some extra information. Ockham violators are strongly beaten with respect to timed retractions alone and are weakly beaten in terms of errors alone. Solutions that violate stalwartness are strongly beaten in terms of retractions but are not even weakly beaten in terms of errors. It is also worth mentioning that solutions that over-count are strongly beaten in terms of retractions alone at the first violation of Ockham's razor. Solutions that cling to a refuted answer "save" one retraction, and hence are strongly beaten only in the sense that they can be forced to skip the missed retraction later than an "honest" solution would have performed it. So the beating argument is not stable with respect to retractions alone.

Together, propositions 1 and 2 yield the following, crisp characterization of efficiency.

Corollary 1 (efficiency characterization) *Let the costs be as in proposition 1 and 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 weakly beaten from e onward;*

4. σ is never strongly beaten from e onward.

Proof. Implications (2) $\Rightarrow$ (3) $\Rightarrow$ (4) are immediate from the definitions. Implication (4) $\Rightarrow$ (1) is by proposition 2. Implication (1) $\Rightarrow$ (2) is from proposition 1. $\dashv$

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. Furthermore, it follows immediately that if σ, σ' are both efficient at e then $\sigma \equiv_e \sigma'$ and that if σ is efficient and σ' is not, then $\sigma \succ_e \sigma'$. Hence, the awkward situation in which $\sigma \amalg_e \sigma'$ arises only if both σ and σ' are both strongly beaten at e.

The idea that one should select answer $T_{\epsilon(e)}$ in light of e is so natural that one feels there *must* be some simpler explanation than the one just given; but the obvious candidates fail, which helps to explain the usual recourse to circles or Providence in standard explanations of Ockham's razor. (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.[10] 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.

6 Stochastic Methods

The preceding arguments still work when empirical strategies are stochastic. In that case, convergence is replaced with convergence in probability and retractions are re-

[10]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 zero. Let a denote the probability that no effect occurs, let b denote the probability that an effect occurs no later than stage i and let c denote the probability that an the first effect occurs after stage i. Then, *a priori*, the expected retractions of 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 iff $b < .5$. One can make c arbitrarily small by increasing i, 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.

placed with retractions in probability, where a retraction in probability is any drop in the chance of producing an answer from stage e_- to stage e. Then the best worst-case upper bounds on retractions in chance in each empirical complexity class are no better than those for deterministic methods. Let r denote an arbitrarily small but non-zero real number and let k be an arbitrary, empirical complexity. Nature can present no effects until stochastic solution σ achieves chance $> 1 - r/2k$ of producing answer $T_{\{\}}$. Now nature can present only effect x_0 until stochastic solution σ achieves chance $> 1 - r/2k$ of producing $T_{\{x_0\}}$. Nature can continue this way until k distinct effects have been presented. The total retractions in probability incurred by σ in complexity class $C_e(k)$ are, therefore, $> k(1 - r)$. Since r is arbitrarily low, the best worst-case bound on retractions in probability over $C_e(k)$ is k, as in the deterministic case. Also, any nonzero chance of violating Ockham's razor or stalwartness at e gets added to the subsequent retractions nature can force from e onward, as in the deterministic argument, so efficiency precludes any non-zero chance of such violations.

7 A More General Setting for the Argument

The preceding argument for Ockham's razor merely stipulates what counts as an empirical effect and complexity is defined in terms of such effects. The argument would be strengthened by a general definition of empirical effects that applies to a broad range of problems and that still supports the preceding argument over that range of problems. The balance of the paper presents just such a generalization.

An *empirical inference problem* is a pair (K, Π), where *presupposition* K is a set of infinite sequences of inputs called *worlds* and *question* Π partitions K into *potential answers*. Define K_{fin} and K_e as before. There is no longer any stipulation about what an empirical effect is or when it is presented. If $w \in K$, then let $T_w \in \Pi$ denote the unique answer true in w. A *strategy* is a mapping from finite sequences of inputs to answers in $\Pi \cup \{\text{`?'}\}$. A *solution* is a strategy such that for each $w \in K$, $\lim_{i \to \infty} \sigma(w|i) = T_w$. To keep the text readable, references to (K, Π) are suppressed as far as possible.

8 Empirical Effects and World Complexity

There have been various attempts to define empirical complexity in terms of descriptive or computational syntax (e.g. Vitanyi and Li 2000), but these proposals are implausible as efficient guides to the truth, because such notions are not invariant under recoding of the data, making Ockham's razor a mere matter of evidential description (Goodman 1983). Here is a plausible proposal that is invariant under recoding of inputs and that has the further advantage of preserving the efficiency argument for Ockham's razor.

In the effect accounting problem, nature can force an arbitrary solution through arbitrary, finite, sub-sequences of answer sequence:

$$b = (T_{\{\}}, T_{\{x_0\}}, T_{\{x_0,x_1\}}, \ldots)$$

according to the argument for proposition 1: after presenting effect x_0, nature can force only finite sub-sequences of the truncated sequence $b' = (T_{\{x_0\}}, T_{\{x_0,x_1\}}, \ldots)$, and so forth. Hence, the presentation of effect x_0 leaves an indelible, structural "footprint" among the sequences of answers forcible by nature from that point onward, in the sense that there is a non-empty, finite, forcible sequence u of answers such that for each sequence of answers b' forcible after effect x_0 is presented, the sequence $b = u * b'$ is forcible before x_0 is presented:

$$b = (\underbrace{T_{\{\}}}_{u}, \underbrace{T_{\{x_0\}}, T_{\{x_0,x_1\}}, \ldots}_{b'}).$$

It remains to express the preceding idea with mathematical precision. The following definitions are all relative to a fixed problem (K, Π). An *answer pattern* is a finite sequence of answers in which no answer occurs immediately after itself. Let g be an answer pattern. Say that g is *forcible* given e of length n just in case for each solution σ there exists k and $e' > e$ of length $n + k$ such that g is a sub-sequence of $(\sigma(e'|n), \ldots, \sigma(e'|(n + k)))$. Say that answer pattern g is *backwards-maximally forcible* at e iff g is forcible given e and for each forcible answer pattern g' given e, if $g \subseteq g'$ then $g \leq g'$ (i.e., it is impossible to extend g to a forcible sequence at e other than by adding further answers to the end of g, so g is extendable but has no "gaps"). Let Δ_e denote the set of all answer patterns that are backwards-maximally forcible at e.

Effects are now definable in terms of Δ_e along the lines discussed informally above. Let e be a non-empty input sequence in K_e and let u be a non-empty answer pattern. Then *effect* u is presented at the end of e iff:

$$(\forall b \in \Delta_e)\ u * b \in \Delta_{e^-}.$$

In the effect accounting problem, the number of effects presented at the end of e corresponds to the length of u in the preceding definition, so one might define the complexity of world w as the total length of the effect sequences encountered along w. Alternatively, one might simply count the number of times that at least one effect is presented. It turns out not to matter for the following efficiency theorems which course is adopted, so the latter, simpler path is adopted. If s is a sequence in $K \cup K_{\text{fin}}$, $w \in K$, $e \in K_{\text{fin}}$, and $A \in \Pi$, define empirical complexity of worlds and answers as follows:

$$\begin{aligned}
c(s) &= |\{e' \leq s : e' \text{ is an effect}\}|; \\
c(w, e) &= c(w) - c(e); \\
c(A) &= \min_{w \in A} c(w); \\
c(A, e) &= c(A) - c(e); \\
C_e(n) &= \{w \in K_e : c(w, e) = n\}.
\end{aligned}$$

Empirical complexity, so defined, is invariant under 1-1 recoding of the inputs but is highly dependent upon the structure of the problem (K, Π), in the sense that the same world w could be arbitrarily simple or arbitrarily complex depending upon the

problem under consideration. Indeed, if Ockham's razor is to have anything to do with the objective efficiency in problem of finding the true answer, simplicity of answers *must* depend upon the structure of the problem because efficiency, itself, does. For example, if $\Pi = \{\{w\}, K - \{w\}\}$ then $c(\{w\}, e) = 0$, whichever world w is singled out as the "point null hypothesis" $\{w\} \in \Pi$. That is intuitive, since it is common practice to favor the simple point hypothesis as the "null" hypothesis in pair-wise decisions. In the present setup, that practice is another instance of Ockham's razor, a striking connection that is impossible to see if empirical complexity is viewed as a problem-independent property of worlds.

9 Ockham's Razor and Stalwartness

Answer T is a *simplest* answer at e iff

$$c(T, e) = \min_{T' \in \Pi} c(T', e).$$

Equivalently, answer T is simplest at e iff $c(T, e) = 0$ (cf. lemma 5 in the appendix). Since there may be several simplest answers at e, Ockham's razor has two versions, strong and weak. Solution σ satisfies the the *weak* version of Ockham's razor at e iff $\sigma(e)$ is a simplest answer at e if $\sigma(e) \neq$'?'. Solution σ satisfies the *strong version* of Ockham's razor at e iff $\sigma(e)$ is the uniquely simplest answer at e if $\sigma(e) \neq$ '?'. *Stalwartness* at e requires that if the scientist's output T at e_- is uniquely simplest at e, then the scientist continues to select T at e.

Ockham's razor may be expressed in terms of simplicity rather than complexity, using a standard rescaling trick familiar from information theory. Define *conditional simplicity* as:

$$s(T, e) = \exp(-c(T, e)).$$

The preceding definition reveals an interesting connection between Ockham's razor and Bayesian updating, for it follows immediately from the definition of $c(T, e)$ that:

$$c(T, e) = c(T \cap K_e) - c(K_e).$$

Applying the definition of $s(T, e)$ to both sides of the preceding equation yields:

$$s(T, e) = \frac{s(T \cap K_e)}{s(K_e)},$$

which is the usual definition of Bayesian updating.[11] Thus, updating a prior bias toward simplicity follows from Ockham's razor and, ultimately, from efficiency. The standard, Bayesian story simply assumes that one updates from a prior simplicity bias, without deriving either the updating rule or the bias from truth-finding performance.

[11] However, $s(T, e)$ is not a probability measure, since several, mutually incompatible propositions can carry unit simplicity.

10 Some Regularity Assumptions

Several regularity assumptions govern the results that follow. In some pathological problems, no answer is forcible "first", so Δ_e is empty even though arbitrarily long sequences of answers are forcible.[12] Such problems are excluded from consideration.

Axiom 1 (forcibility is well-founded) *If pattern b is forcible at e, then there exists $b' \in \Delta_e$ such that $b \subseteq b'$.*

As a simplifying assumption, problems in which the same answer occurs more than once in a forcible pattern are also excluded from consideration. When this assumption is violated, one can typically choose a natural way to refine the problem so as to satisfy it. For example, if the question is whether the total number of effects is even or odd, one can refine the question to ask for the total number of effects.[13]

Axiom 2 (forcibility is acyclic) *If pattern b is forcible at e, then no answer occurs more than once in b.*

A further simplifying assumption is that each answer in a backwards-maximally forcible sequence can be decremented individually, without the intrusion of new, intermediate answers that weren't already in the sequence.[14]

Axiom 3 (monotonicity) *Let $T * T' * b \in \Delta_e$. Then there exists finite $e' > e$ such that exactly one effect occurs along e' after e and $T' * b \in \Delta_{e'}$.*

There exist problems that have no strongly Ockham solution. In such problems, waiting for simplicity to determine uniquely what one should say does not suffice to solve the problem—extra choices that break the symmetry among simplest answers must be made. For example, suppose that the problem is to say when each effect occurs, in addition to just listing them. In that problem, only answer patterns of unit length are forcible and there are always forcible unit-length patterns, so there are no effects and every answer compatible with experience is simplest at each stage, so the strong version of Ockham's razor precludes all answers except for '?'. Such problems are excluded from consideration.

[12] E.g., suppose that tomorrow you may see any number of effects and that any of the effects may disappear at any time thereafter. The problem is to count the total number of effects. At the outset, each finite, descending sequence of effect counts is forcible, so each forcible pattern can be extended at the beginning to a forcible pattern.

[13] There may be many, materially inequivalent such refinements. The even/odd problem just described is, indeed, a coarsening of the problem of counting effects. But, following the strategy of (Goodman 1983), say that a *neffect* is an effect at any stage but stage 1, at which time it is a non-effect. Then finitely many effects are the same as finitely many neffects, so K is preserved, and evenly many effects are the same as oddly many neffects, so the answers "even" and "odd" are preserved. Hence, neffect counting is also a refinement of the even/odd problem, but neffect counting is a materially distinct problem from effect counting.

[14] Suppose that the task is to report the total number of effects n you will see if $n \neq 2$ and to report (n, k), where k is the time of appearance of the second effect otherwise. Then $\Delta_{()} = \{(0, 1, 3, 4, \ldots)\}$. But upon receipt of input sequence $(\{x_1, x_2\})$, $\Delta_e = \{(2, 1), 3, 4, \ldots)\}$, so there is no finite pattern b such that $b * (0, 1, 2, \ldots, n) \in \Delta_{()}$. This example violates even the weaker requirement that $b \in \Delta_{(\{x_1, x_2\})}$ is a sub-pattern of some pattern in $\Delta_{()}$.

Axiom 4 (strong Ockham solvability) *The problem under consideration is strongly Ockham solvable.*

Say that a problem is *nested* if there exists a uniquely simplest answer at each e compatible with K. For example, the problem of accounting for effects is nested, but that is no longer the case if new effects are announced by a horn prior to their appearance. The nesting property is not assumed as an axiom, but it does yield stronger results, so the case of nested problems is considered separately.

11 General Argument for Ockham's Razor

The following assumptions govern all of the results that follow (including those presented in the appendices). Proofs of the subsequent propositions are presented in the appendix.

1. (K, Π) is an empirical problem satisfying axioms 1, 2, 3, and 4;
2. Empirical complexity is defined in terms of forcibility, in the structural manner just discussed;
3. Efficiency and strong and weak beating are defined in terms of empirical complexity as in the discussion of the effect accounting problem;
4. σ is a solution;
5. $e \in K_{\text{fin}}$.

To begin with, stalwart, Ockham solutions to nested problems are stably efficient and stalwart, strongly Ockham solutions are efficient in general, albeit not necessarily stably so.

Proposition 3 (strong Ockham and stalwartness imply efficiency)

1. *If the problem is nested and σ is stalwart and Ockham solution from e onward, then σ is efficient from e onward.*
2. *If σ is always stalwart and Ockham, then σ is always efficient.*

Furthermore, violators of Ockham's razor or stalwartness are not merely inefficient, but strongly beaten at each violation, so Ockham's razor has a stable sanction.

Proposition 4 (efficiency stably implies Ockham's razor and stalwartness) *If σ violates Ockham's razor or stalwartness at e, then σ is strongly beaten at e.*

In general, violations of the strong version of Ockham's razor incur only a weak beating at the first violation, so the strong version of Ockham's razor has a weaker motivation

than does the weak one.[15] The beating occasioned by a violation of the strong Ockham principle may be strong, however, even when several answers are simplest: for example when effects are announced by a horn.

Proposition 5 (efficiency implies strong Ockham) *If σ violates the strong Ockham principle or stalwartness at e, then σ is weakly beaten at the first moment e' along e at which the strong Ockham principle is violated.*

The following corollary extends corollary 1 to all nested problems satisfying axioms 1-4.

Corollary 2 (efficiency characterization, nested case) *If the problem is nested, then the following statements are equivalent.*

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

In general, stalwart, strongly Ockham solutions are still uniquely best, but the argument is weaker: the beating incurred by violators may be weak or unstable.

Corollary 3 (efficiency characterization, general case) *The following statements are equivalent.*

1. *σ is always efficient.*
2. *σ is never weakly beaten.*
3. *σ is always stalwart and strongly Ockham.*

Since the problem is assumed to be strongly Ockham solvable (axiom 4), there exists a strongly Ockham solution and, hence:[16]

Corollary 4 *There exists a stalwart, strongly Ockham solution.*

Hence, given axioms 1-4, one ought to use a stalwart, Ockham solution and one should not use any other sort of strategy.

[15] For example, suppose at e that a curtain will be opened tomorrow that reveals either a marble emitter or nothing at all. The question is whether there is an emitter behind the curtain and, if so, how many marbles it will emit. The no-emitter world and the marble-free emitter world are both simplest in this example, so the strong Ockham principle requires that one suspend judgment between the corresponding answers until the curtain is opened. Suppose that you guess that you are in the marble-free emitter world, in spite of the strong Ockham principle. You are not strongly beaten, because you do as well as possible in each class $C_e(k)$ such that k exceeds zero.

[16] For a strongly Ockham solution converges to the uniquely simplest answer in each world and is not prevented from doing so by hanging onto a uniquely simplest answer until it is no longer uniquely simplest.

12 Conclusion

A fairly general explanation of Ockham's razor and related methodological principles has been given in terms of truth-finding efficiency. In nested problems, in which a uniquely simplest answer always exists, the result is that:

$$\text{stable efficiency} = \text{Ockham's razor} + \text{stalwartness}.$$

In general:

$$\text{efficiency} = \text{strong Ockham's razor} + \text{stalwartness}.$$

The underlying notion of efficiency is similar in spirit to that employed in standard, worst-case complexity analyses of algorithm performance, with a few alterations suitable to the unending character of empirical inquiry. The costs considered are the number of errors and retractions prior to convergence together with the elapsed time to each retraction. The argument works without the need to broach the awkward question of trading errors against retractions or retraction times. The standard, computer science concept of "problem instance size" is replaced with a mathematical explication of the intuitive notion of empirical complexity of a possible world relative to the inference problem at hand. Efficiency means doing as well as an arbitrary solution in the worst case over each problem instance size. Not only are violators of Ockham's razor inefficient in this sense, they are strongly dominated with respect to worst-case bounds over instance sizes. Thus, Ockham's razor does have an objective, complexity-theoretic, connection to truth-finding that works without either changing the question or begging it with prior probabilistic biases toward simple worlds or simple theories.

The price of this objective non-circularity is that the explanation does not imply that the simpler answer is true or has a high chance of being true. Ockham's razor keeps you on the straightest possible path to the truth, but the path may still have arbitrarily many twists and bends that steer you away from your goal from time to time—or even for most of the time—prior to convergence. To demand more is to demand the impossible.

Many projects and questions remain. (1) Axioms 1-4 are neither principles of rationality nor universal features of empirical problems; they are nothing but restrictions on the scope of the efficiency theorems and should, therefore, be weakened as far as possible.

(2) It has not been shown that the proposed definition $c(w, e)$ of empirical complexity is the only definition for which the preceding propositions can be established. Indeed, the following, more sophisticated definition of empirical complexity looks promising. Assume that for each world $w \in K$, the limit

$$\Delta_w = \lim_{i \to \infty} \Delta_{w|i}$$

exists, in the sense that there exists n such that for each $i \geq n$, $\Delta_{w|i} = \Delta_{w|n}$. Now, think of $\Gamma_e = \{\Delta_w : w \in K_e\}$ as a set of *states* and let $\Gamma = \Gamma_{()}$. Define *accessibility* between states in Γ as follows: $D \leq D'$ iff for each e' such that $D = \Delta_e$ there exists

$e' \geq e$ such that $D' = \Delta_{e'}$. The result of the construction, in many natural problems, is a partial order $(\Gamma, \leq)$ whose ascending paths are all of order type $\leq \omega$. Let $\Gamma_e^{\min}$ denote the set of all states in Γ_e that are minimal in the $\leq$ order. Let $\pi(\Delta)$ denote the set of all $\leq$-paths in $\Gamma|e$ that originate in $\Gamma_e^{\min}$ and terminate with Δ. Define $c'(w, e) = \sup\{\text{length}(p) : p \in \pi(\Delta_w)$. The definition of $c'(w, e)$ yields intuitive judgments about simplicity over a broader range of problems than does the concept $c(w, e)$ involved in the preceding proofs. Also, $c'(w, e)$ allows for a broader range of strongly-Ockham-solvable problems than does $c(w, e)$.[17]

(3) Worlds are modeled as concrete input streams in the preceding development. Real scientific questions, insofar as they are formalized at all, are usually formulated in terms of real-valued parameter spaces. Since the forcing and convergence concepts employed in the theory all make sense in metrizable, separable topological spaces, it is both promising and desirable to lift the preceding results to problems (K, Π) in which K is such a space. In this more general topological setting, an input stream can be modeled as an infinite, descending sequence of open neighborhoods whose intersection is a subset of some answer to the problem. Efficiency and empirical complexity can then be defined much as in the preceding development.

(4) The theory does not yet apply literally to statistical problems, but that is the ultimate aim. Here is a tentative sketch of how it might go (cf. also Kelly 2004a, Kelly and Glymour 2004). In statistical problems, as in the case of stochastic methods discussed above, the aim is to converge in probability to the true theory in a way that minimizes retractions in probability and the times at which they occur. Retractions in chance are ubiquitous in statistics. For example, consider a standard statistical test of the null hypothesis that the mean of a normal distribution is identically zero. If the true value of the mean is non-zero but very small, then in small samples the test will very probably accept the null hypothesis and in large samples it will very probably reject. No possible method that converges in chance to the true answer to the binary question as sample size increases can avoid the retraction in chance (accept, reject). Furthermore, favoring rejection of the point hypothesis rather than acceptance at small sample sizes risks, in the worst case, an extra retraction in chance (reject, accept, reject). So favoring the null hypothesis in the usual way by setting a "significance" level minimizes statistical retractions *en route* to the truth. Optimization of a statistical procedure's "power" can then be explained in terms of minimizing the time (i.e. sample size) to each retraction in chance, since a sloppy statistical test will require a larger sample to reject the null hypothesis. But whereas the logic of statistical testing is applied only to binary questions, the idea of minimizing retractions in chance applies more broadly

[17]For example, suppose, in the effect accounting problem, that world w presents first the disjunctive information $\{x\} \vee \{y, z\}$, which is interpreted as saying that either x or both y, z will be seen, and that presents disjunct $\{x\}$ followed by $\emptyset$ thereafter. Then for each e presenting no effects, $c(w, e) = 0$. Hence, a strongly Ockham method must withhold judgment forever if the truth is $T_{\{x\}}$ and, therefore, never converges to the truth, so the problem is not strongly Ockham-solvable, violating axiom 4. On the other hand, $c'(w, e) = 1$ so, assuming that all disjunctive reports are resolved, eventually, in favor of some disjunct (or, possibly, for both), the problem is strongly Ockham solvable according to the new definition and the strong version of Ockham's razor plausibly demands that one suspend judgment between disjuncts until the disjunction is resolved.

to model selection problems like statistical curve fitting that require arbitrarily many retractions. The hope is that such an analysis will explain the intuitive bias toward simple laws in statistical curve fitting in terms of finding the true model.

(5) Finally, the preceding results do not begin to address the crucial question of trade-offs between ideal and computational efficiency when it is intractable or even impossible to compute the set of simplest answers at e. At least it provides an ideal framework that may serve as a guide toward the development of a more computationally realistic analysis (cf. Kelly 2004b for a discussion of Ockham's razor in the purely computational realm).

13 Acknowledgments

The author is indebted to (in alphabetical order) Seth Casana, Stephen Fancscali, Conor Mayo-Wilson, Joseph Ramsey, Richard Scheines, Cosma Shalizi, John Taylor, and Larry Wasserman for many patient, constructive, and critical discussions related to the material in this paper.

14 References

Akaike, H. (1973) "Information theory and an extension of the maximum likelihood principle," *Second International Symposium on Information Theory.* pp. 267-281.

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.

Carnap, R. (1950) *The Logical Foundations of Probability.* Chicago: University of Chicago Press.

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.

Friedman, M. (1983) *Foundations of Space-time Theories*, Princeton: Princeton University Press.

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.

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.

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

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.

15 Appendix: Proofs of Propositions

In the following results, (K, Π) is assumed to be an empirical problem satisfying 1, and 4. Adopt the abbreviation:

$$\omega^{(n)} = \underbrace{\omega, \ldots, \omega}_{n \text{ repetitions}} .$$

Let $\delta_e(\sigma, w)$ denote the number of errors committed by σ in w along $e < w$. Let

$$\delta_e(\sigma, 0) = \max_{w \in C_{e_-}(0)} \delta_e(\sigma, w).$$

Proof of proposition 3.1. Let the problem be nested and let σ' be a solution that is Ockham and stalwart from e onward. Since the problem is nested, σ' is also strongly Ockham from e onward. Let σ be a solution such that $\sigma \asymp_{e_-} \sigma'$. Set $q = \delta_{e_-}(\sigma, 0) = \delta_{e_-}(\sigma', 0)$. Let $r_1, \ldots, r_k$ be the times of the successive retractions performed by both σ and σ' along e_-. Suppose that $C_e(n+1) \neq \emptyset$.

Case: σ' does not retract at e. Then $\lambda_e(\sigma', 0) \leq \lambda_e(\sigma, 0)$ (by lemmas 10.1 and 12.4) and $\lambda_e(\sigma', n+1) \leq \lambda_e(\sigma, n+1)$ (by lemmas11.1 and 13.4).

Case: σ retracts at e. Then $\lambda_e(\sigma', 0) \leq \lambda_e(\sigma, 0)$ (by lemmas 10.2 and 12.3) and $\lambda_e(\sigma', n+1) \leq \lambda_e(\sigma, n+1)$ (by lemmas 11.2 and 13.3).

Case: σ' retracts at e and σ does not retract at e. Then since σ' is stalwart and strongly Ockham from e onward, $\sigma'(e_-) = \sigma(e_-) = \sigma(e)$ is not uniquely simplest at e. Since the problem is nested, $\sigma(e)$ is not simplest at e. Then $\lambda_e(\sigma', 0) \leq \lambda_e(\sigma, 0)$ (by lemmas 10.2 and 12.1) and $\lambda_e(\sigma', n+1) < \lambda_e(\sigma, n+1)$ (by lemmas 11.2 and 13.1). $\dashv$

Proof of proposition 3.2. Let σ' be a solution that is Ockham and stalwart at every stage. Let σ be a solution such that $\sigma' \asymp_{e_-} \sigma'$. Set $q = \delta_{e_-}(\sigma, 0) = \delta_{e_-}(\sigma', 0)$. Let $r_1, \ldots, r_k$ be the times of the successive retractions performed by both σ and σ' along e_-. Suppose that $C_e(n+1) \neq \emptyset$.

Case: σ' does not retract at e or σ retracts at e. Same as in preceding proof.

Case: σ' retracts at e and σ does not. Since σ' is always stalwart and strongly Ockham and σ' retracts at e, it follows that $\sigma'(e_-) = \sigma(e_-) = \sigma(e)$ is not uniquely simplest at e. Then $\sigma(e_-) \neq$ '?' and, since $\sigma \asymp_{e_-} \sigma'$, e is the first time at which σ returns an answer that is not uniquely simplest. Then $\lambda_e(\sigma', 0) < \lambda_e(\sigma, 0)$ (by lemmas 10.2 and 12.2) and $\lambda_e(\sigma', n+1) \leq \lambda_e(\sigma, n+1)$ (by lemmas 11.2 and 13.2). $\dashv$

Proof of proposition 4 Let σ be a solution that violates either Ockham's razor or stalwartness at e of length j. There exists $\sigma' \asymp_{e_-} \sigma$ such that σ' is strongly Ockham and stalwart from e onward (by lemma 1). Set $q = \delta_{e_-}(\sigma, 0) = \delta_{e_-}(\sigma', 0)$. Let $r_1, \ldots, r_k$ be the times of the successive retractions performed by both σ and σ' along e_-. Suppose that $C_e(n+1) \neq \emptyset$.

Case: σ violates Ockham's razor at e. Hence, $\sigma(e)$ is not simplest at e. Then $\lambda_e(\sigma', 0) < \lambda_e(\sigma, 0)$ (by lemmas 10.2 and 12.1) and $\lambda_e(\sigma', n+1) < \lambda_e(\sigma, n+1)$ (by lemmas11.2 and 13.1).

Case: σ violates stalwartness at e. So there exists a uniquely simplest answer T at e such that $\sigma(e_-) = T$ but $\sigma(e) \neq T$. Then σ retracts at e but σ' does not. So $\lambda_e(\sigma', 0) < \lambda_e(\sigma, 0)$ (by lemmas 10.1 and 12.3) and $\lambda_e(\sigma', n+1) < \lambda_e(\sigma, n+1)$ (by lemmas 11.1 and 13.3). $\dashv$

Proof of proposition 5. The stalwartness case is immediate from proposition 4. Now suppose that σ is a solution that violates the strong Ockham principle (somewhere). Then there exists finite input sequence e compatible with K such that σ violates the strong Ockham principle at e, but not at any proper sub-sequence of e. Then $\sigma(e) = T$, where T is not the uniquely simplest answer compatible with e. Let j denote the length of e. There exists solution $\sigma' \asymp_{e_-} \sigma$ that is stalwart and strongly Ockham from e onward (by lemma 1). Set $q = \delta_{e_-}(\sigma, 0) = \delta_{e_-}(\sigma', 0)$. Let $r_1, \ldots, r_k$ be the times of the successive retractions performed by both σ and σ' along e_-. Then $\lambda_e(\sigma', 0) < \lambda_e(\sigma, 0)$ (by lemmas 10.2 and 12.2). Suppose that $C_e(n+1) \neq \emptyset$.

Case: σ does not retract at e. Then $\lambda_e(\sigma', n+1) < \lambda_e(\sigma, n+1)$ (by lemmas 11.2 and 13.2).

Case: σ retracts at e. Then $\sigma(e_-) \neq$ '?', so $\lambda_e(\sigma', n+1) \leq \lambda_e(\sigma, n+1)$ (by lemmas 11.2 and 13.3). $\dashv$

Proof of corollary 2. (1) implies (2) implies (3) by definition. (3) implies (4) by proposition 4. (4) implies the strong Ockham principle and stalwartness since the problem is nested. The strong Ockham principle and stalwartness imply (1) by proposition 3.1. $\dashv$

Proof of corollary 3. (1) implies (2) by definition. (2) implies (3) by propositions 5. (3) implies (1) by proposition 3.2. ⊣

16 Appendix: Lemmas

Lemma 1 (solution variants) *For each solution σ, $e \in K_{fin}$, there exists solution $\sigma' \asymp_{e\_} \sigma$ such that σ' is strongly Ockham and stalwart from e onward.*

Proof. Let σ' agree with σ along e and then produce the uniquely simplest answer compatible with e' if it exists and '?' otherwise, for each e' properly extending e. Then by construction, σ' is stalwart and Ockham from e onward and $\sigma' \asymp_{e\_} \sigma$. Since (K, Π) is strongly Ockham solvable (axiom 4), σ' solves (K, Π), because σ' converges, in each world, to whatever the assumed, strongly Ockham solution converges to in that world. ⊣

Lemma 2 (forcibility is asymptotic) *Let pattern $T * b$ be forcible given e. Then there exists a world $w \in K_e \cap T$ extending e such that for each finite e' such that $e \leq e' < w$, $T * b$ is forcible given e'.*

Proof. Suppose $T * b$ is forcible given e. Suppose for reductio that the consequent of the lemma is false. Then for each $w \in T \cap K_e$ there exists e' extending e and extended by w such that $T * b$ is not forcible given e'. For each $w \in T \cap K_e$, let e_w be the shortest such e'. Let $\Pi_{e_w} = \{T \cap K_{e_w} : T \in \Pi\}$. For each e_w, $T * b$ is not forcible at e_w, so there exists a solution σ_w for (K_{e_w}, Π) that never produces $T * b$ after e_w. Let σ solve (K, Π) and let σ^* be just like σ except that control is shifted permanently to σ_w when e_w is encountered. So σ^* is a solution that never produces $T * b$ after seeing some e_w. Let $\sigma^\dagger$ be like σ^* except that $\sigma^\dagger$ produces '?' along each e_w and at each e not extended by some e_w such that σ returns T at e. Then $\sigma^\dagger$ is still a solution, since σ^* converges to the truth over $K_e \cap T$ (the question marks eventually end in each $w \in K_e \cap T$) and over $K_e - T$ (σ does not converge to T in any such world, so again, the question marks end eventually in each $w \in K_e - T$). But $\sigma^\dagger$ doesn't produce $T * b$ after e along any e' extending e. So $T * b$ is not forcible given e. Contradiction. ⊣

Lemma 3 (forcible pattern existence) *Suppose that $C_e(n)$ is non-empty. Then there exists a finite pattern $b \in \Delta_e$ of length $\geq n + 1$.*

Proof. Let $w \in C_e(0)$. In the base case, nature can force the answer T true in w from an arbitrary solution since a solution must converge to T along w. So (by axiom 1) there exists some pattern $b \in \Delta_e$ of length ≥ 1. For induction, suppose that $w \in C_e(n+1)$. Let e' be the first effect along w after e. So there are n effects occurring along w after e'. By the induction hypothesis, there exists pattern $a \in \Delta_{e'}$ of length at least $n + 1$. Since e' is an effect, there exists pattern $T * b$ such that $T * b * a$ is a pattern in $\Delta_{e'\_}$. Hence, $T * b * a$ has length at least $n + 2$. Since $T * b * a$ is forcible at $e'\_$, $T * b * a$ is forcible at e as well. So there exists some pattern $d \in \Delta_e$ of which $T * b * a$ is a sub-pattern (by axiom 1), so d has length at least $n + 2$. ⊣

Lemma 4 (nature's starting point) *Let $T*b \in \Delta_e$. Then there exists $w \in C_e(0) \cap T$ such that for each finite e' such that $e \leq e' < w$, $T * b \in \Delta_{e'}$.*

Proof. Let $T * b \in \Delta_e$. So $T * b$ is forcible given e. There exists $w \in K_e \cap T$ such that $T * b$ is forcible at each finite e' such that $e \leq e' < w$ (by lemma 2). Suppose that $T * b$ is a subpattern of d and $T * b \not\leq d$ and d is forcible at e'. Then d is forcible at e, so $T * b \notin \Delta_e$, which is a contradiction. Hence, $T * b \in \Delta_{e'}$.

Now suppose for reductio that there exists effect e' such that $e < e' < w$. Then there exists $T' * c$ such that $T' * c * \Delta_{e'} \subseteq \Delta_{e'_-}$ and no member of $\Delta_{e'_-}$ begins with T'. Recall that $T * b \in \Delta_{e'}$, so $T' \neq T$ and $T' * c * T * b \in \Delta_{e_-}$. So since $e \leq e'_- < e'$, $T' * c * T * b$ is forcible at e. Since $T' \neq T$, it follows that $T * b \not\leq T' * c * T * b$, so $T * b \notin \Delta_e$. Contradiction. So no effect occurs in w after e, so $w \in C_e(0)$. $\dashv$

Lemma 5 (simple world existence) *Let $K_e \neq \emptyset$. Then there exists $w \in C_e(0)$.*

Proof. Suppose there exists $w \in K_e$. If $c(w, e) = 0$, we are done. So suppose $c(w, e) = k > 0$. Then (by lemma 3) there exists $T * b \in \Delta_e$ of length $k + 1$. So by lemma 4, there exists $w' \in C_e(0) \cap T$. $\dashv$

Lemma 6 (simplest answers forcible first) *Let $T * b \in \Delta_e$. Then T is a simplest answer.*

Proof. Suppose that $T * b \in \Delta_e$. Then there exists $w \in T \cap K_e$ such that for each e' for which $e \leq e' < w$, $T * b \in \Delta_{e'}$ (by lemma 4). Suppose for reductio that T is not a simplest answer at e. Then for each $w' \in T \cap K_e$, $c(w', e) > 0$, so $c(w, e) > 0$. Hence, there exists effect e' such that $e < e' < w$. Hence, there exists $T' * c$ such that $T' * c * \Delta_{e'} \subseteq \Delta_{e'_-}$. Since $T * b \in \Delta_{e'}$, $T' * c * T * b \in \Delta_{e'_-}$. Hence, $T' * c * T * b$ is forcible at e. So $T * b \not\leq T' * c * T * b$ (by axiom 2). Hence, $T * b$ is not backwards-maximally forcible at e, so $T * b \notin \Delta_e$. Contradiction. $\dashv$

Lemma 7 (uniquely simplest answer and forcibility) *Let answer $T \in \Pi$ be uniquely simplest at e. Then each pattern in Δ_e begins with T.*

Proof. Suppose that for some answer $T' \neq T$, pattern $T' * a \in \Delta_e$. So (by lemma 6), T' is simplest at e. So T is not uniquely simplest. $\dashv$

Lemma 8 (simplest answer defeated only by effects) *Let $K_e \neq \emptyset$, let e be non-empty, and let answer $T \in \Pi$ be uniquely simplest at e_- but not at e. Then e is an effect.*

Proof. Let K_e, e be non-empty. Then K_{e_-} is non-empty, so (by lemma 5) $C_{e_-}(0), C_e(0)$ are non-empty. So since T is uniquely simplest at e_- but not at e, we have $C_{e_-}(0) \subseteq T$ but $C_e(0) \not\subseteq T$. So there exists $w \in C_e(0) - C_{e_-}(0)$. Hence, $c(w, e_-) > 0$ and $c(w, e) = 0$, so e is an effect. $\dashv$

Lemma 9 (sequential forcing) *Let e be a finite input sequence and let pattern $a = (T_1, \ldots, T_k) \in \Delta_e$ and let i be a natural number and let $1 \leq m \leq k$. Then there exists $w \in C_e(m-1) \cap T_m$ and e' such that $e < e' < w$ and:*

1. *for each $m' \leq m$, σ produces $T_{m'}$ at least i times in immediate succession after e along e' and*

2. *exactly $m-1$ effects occur along e' after e and*

3. *for each e'' such that $e' \leq e'' < w$, $(T_m, \ldots, T_k) \in \Delta_{e''}$.*

Proof. In the base case, let $m = 1$. Let i be arbitrary. There exists $w \in C_e(0) \cap T_1$ such that for all e' such that $e \leq e' < w$, $(T_1, \ldots, T_k) \in \Delta_{e'}$ (by lemma 4). Since $w \in T_1$ and σ is a solution, σ converges to T_1 in w and, hence, produces T_1 at least i times after e, say by e' such that $e < e' < w$. Since $w \in C_e(0)$ and $e' < w$, no effects occur along e' after e.

In the inductive case, let $1 < m \leq k$. By the induction hypothesis, there exists $w \in C_e(m-1) \cap T_{m-1}$ and e_0 such that $e < e_0 < w$ and:

1. for each $m' \leq m-1$, σ produces $T_{m'}$ at least i times in immediate succession after e along e_0 and

2. exactly $m-2$ effects occur along e_0 after e and

3. for each e' such that $e_0 \leq e' < w$, $(T_{m-1}, \ldots, T_k) \in \Delta_{e'}$.

There exists $e_1 > e_0$ such that exactly one effect occurs along e_1 after e_0 and $(T_m, \ldots, T_k) \in \Delta_{e_1}$ (by axiom 3). There exists $w' \in C_{e_1}(0) \cap T_m$ such that for each e' such that $e_1 \leq e' < w'$, $(T_m, \ldots, T_k) \in \Delta_{e'}$ (by lemma 4). Since $w' \in T_m$, there exists e_2 such that $e_1 < e_2 < w'$ and solution σ produces T_m at least i times along e_2 after e. Then $m-1$ effects occur after e along e_0, one effect occurs along e_1 after e_0 and no effects occur along w' after e_1 since $w' \in C_{e_1}(0)$. Hence, $w' \in C_e(m-1)$. So w', e_2 have the required properties at level m. $\dashv$

Lemma 10 (upper bound over $C_e(0)$) *Let $K_e \neq \emptyset$, let e have length j, and let solution σ be stalwart and strongly Ockham from e onward. Let $(r_1, \ldots, r_k)$ be the timed retractions of σ along e_-. Set $q = \delta_{e_-}(\sigma, 0)$. Then*

$$\lambda_e(\sigma, 0) \leq \begin{cases} 1.\ (q, (r_1, \ldots, r_k)) & \text{if } \sigma \text{ does not retract at } e; \\ 2.\ (q, (r_1, \ldots, r_k, j)) & \text{in general.} \end{cases}$$

Proof. Suppose that $w \in C_e(0)$, where e has length j. Let $(r_1, \ldots, r_k)$ be the retraction times for both σ and σ' along e_-. Also, σ' produces only the uniquely simplest answer compatible with experience or '?' and never retracts it along w from e onward (lemma 8), so σ' produces no false answers along w from the end of e onward. So if σ does not retract at e, then:

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

In the general case, σ may retract at e:

$$\lambda(\sigma, w) \leq (q, (r_1, \ldots, r_k, j)).$$

Since w is an arbitrary element of $C_e(0)$,

$$\lambda_e(\sigma, 0) \leq \lambda(\sigma, w). \dashv$$

Lemma 11 (upper bound over $C_e(n+1)$) *Let e have length j, and let solution σ be stalwart and strongly Ockham from e onward. Let $(r_1, \ldots, r_k)$ be the timed retractions of σ along e_-. Then*

$$\lambda_e(\sigma, n+1) \leq \begin{cases} 1.\ (\omega, (r_1, \ldots, r_k, \omega^{(n+1)})) & \text{if } \sigma \text{ does not retract at } e; \\ 2.\ (\omega, (r_1, \ldots, r_k, k, \omega^{(n+1)})) & \text{in general.} \end{cases}$$

Proof. Since σ retracts at most once at each effect from e (by lemma 8), it follows that even if σ retracts at e:

$$\lambda_e(\sigma, n+1) \leq (\omega, (r_1, \ldots, r_k, j, \omega^{(n+1)})).$$

If σ does not retract at e:

$$\lambda_e(\sigma, n+1) \leq (\omega, (r_1, \ldots, r_k, \omega^{(n+1)})). \dashv$$

Lemma 12 (lower bound over $C_e(0)$) *Let $K_e \neq \emptyset$ and let e have length j. Let $(r_1, \ldots, r_k)$ be the timed retractions of solution σ along e_-. Set $q = \delta_{e_-}(\sigma, 0)$. Then*

$$\lambda_e(\sigma, 0) \geq \begin{cases} 1.\ (q+1, (r_1, \ldots, r_k, j+1)) & \text{if } \sigma(e) \text{ is not simplest at } e; \\ 2.\ (q, (r_1, \ldots, r_k, j+1)) & \text{if } \sigma(e) \text{ is not uniquely simplest at } e; \\ 3.\ (q, (r_1, \ldots, r_k, j)) & \text{if } \sigma \text{ retracts at } e; \\ 4.\ (q, (r_1, \ldots, r_k)) & \text{in general.} \end{cases}$$

Proof. (1) Suppose that $T = \sigma(e)$ is not simplest at e. Hence, $C_e(0) \cap T = \emptyset$. Let $w \in C_0(e)$ be such that $\delta_e(\sigma, w) = \delta_e(\sigma, 0)$. Then $w \notin T$, so σ commits at least $\delta_e(\sigma, w) + 1$ errors in w. Since σ is a solution, σ converges to $T' \in w$, so there exists some e' such that $e < e' < w$ and $\sigma(e') \neq T$. Hence,

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

(2) Suppose that e is least such that answer $\sigma(e)$ is not uniquely simplest at e. Let $D = \sigma(e)$. So there exists $w \in C_e(0) - D$. Since σ is a solution, there exists e' such that $e < e' < w$ and $\sigma(e') = T_w$, so σ retracts along w no sooner than stage $j+1$. Also, since $\delta_e(\sigma, 0) = d$ there exists $w' \in C_e(0)$ in which σ commits d errors after e along w'. Hence,

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

(3) Suppose that σ retracts at e. There exists $w \in C_e$ (by lemma 5), so choose $w \in C_e(0)$ so that σ commits $q = \delta_e(\sigma, w)$ errors along e_- in w. The retractions by σ along e_- get counted along with the retraction at e. Hence:

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

(4) In general, simply drop the retraction at j from the argument in case (3). $\dashv$

Lemma 13 (lower bound over $C_e(n+1)$) *Let $C_e(n+1)$ be non-empty, let e have length j. Let $(r_1, \ldots, r_k)$ be the timed retractions of solution σ along e_-. Then*

$$\lambda_e(\sigma, n+1) \geq \begin{cases} 1.\ (\omega, (r_1, \ldots, r_k, j+1, \omega^{(n+1)})) & \text{if } \sigma(e) \text{ is not simplest at } e \\ 2.\ (\omega, (r_1, \ldots, r_k, j, \omega^{(n+1)})) & \text{if } \sigma(e_-) \neq \text{`?' and} \\ & e \text{ is least such that} \\ & \sigma(e) \text{ is not uniquely simplest at } e; \\ 3.\ (\omega, (r_1, \ldots, r_k, j, \omega^{(n+1)})) & \text{if } \sigma \text{ retracts at } e; \\ 4.\ (\omega, (r_1, \ldots, r_k, \omega^{(n+1)})) & \text{in general.} \end{cases}$$

Proof. Suppose that $C_e(n+1) \neq \emptyset$. There exists pattern $T * T' * d \in \Delta_e$ of length $n+2$ (by lemma 3). Let arbitrary natural number i be given. There exists $w_0 \in C_e(0)$ such that for each e' such that $e \leq e' < w$, $T * T' * d \in \Delta_{e'}$ (by lemma 4).

(1) $\sigma(e)$ is not a simplest answer at e. Hence, $w_0 \notin \sigma(e)$. So since σ is a solution, σ eventually converges to the true answer T_0 in w_0, so there exists e_0 such that $e < e_0 < w$ and σ produces T_0 successively at least i times along e_0 after the end of e and σ retracts $\sigma(e)$ back to T_0 no sooner than stage $j+1$. Pattern $T * T' * d \in \Delta_{e_0}$. So (by lemma 9) there exists $w_{n+1} \in C_{e_0}(n+1)$ along which each answer occurring in $T' * d$ is repeated by σ at least i times after the end of e_0. Since $T \neq T'$, σ commits at least i errors along w_{n+1} after the end of e_0. Since $e_0 < w_0 \in C_e(0)$, no effects occur in e_0 after the end of e. So since $w_{n+1} \in C_{e_0}(n+1)$, $w_{n+1} \in C_e(n+1)$. Hence,

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

Since i is arbitrary,

$$\lambda_e(\sigma, n+1) \geq \lambda(\sigma, w_{n+1}) \geq (\omega, (r_1, \ldots, r_k, j+1, \omega^{(n+1)})).$$

(2) Suppose that $\sigma(e_-) \neq$ '?' and e is least such that $\sigma(e)$ is not uniquely simplest at e. Let $D = \sigma(e)$. So there exists $w \in C_e(0) - D$. Case: $\sigma(e_-) = D$. Then D is uniquely simplest at e_-. Hence, $w \notin C_{e_-}(0)$. So e is an effect in w. So there exists pattern $G * g$ such that $G * g * \Delta_e \subseteq \Delta_{e_-}$. Since the uniquely simplest hypothesis D at e_- begins each pattern in Δ_{e_-} (by lemma 7), $G = D$, so $D * g * \Delta_e \subseteq \Delta_{e_-}$. So (by axiom 2) no pattern in Δ_e begins with D. Hence, D is not even simplest given e. Revert to case (1). Case: $\sigma(e_-) \neq D$. Then σ retracts at e. Revert to case (3).

(3) σ retracts at e. Since $T * T' * b \in \Delta_e$ and the length of $T * T' * b$ is at least $n+2$, it follows (by lemma 9) that there exists $w_{n+1} \in C_e(n+1)$ along which each answer occurring in $T * T' * b$ is repeated by σ at least i times after the end of e. Since $T \neq T'$, σ commits at least i errors along w_{n+1} after the end of e. Hence,

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

Since $i > j+1$ is arbitrary,

$$\lambda_e(\sigma, n+1) \geq (\omega, (r_1, \ldots, r_k, j, \omega^{(n+1)})).$$

(4) general case. Follow case 3, ignoring the retraction at j. $\dashv$