Ockham's Razor, Truth, and Information

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

December 10, 2007

Abstract

In science, one faces the problem of selecting the true theory from a range of alternative theories. The typical response is to select the *simplest* theory compatible with available evidence, on the authority of "Ockham's Razor". But how can a fixed bias toward simplicity help one find possibly complex truths? A short survey of standard answers to this question reveals them to be either wishful, circular, or irrelevant. A new explanation is presented, based on minimizing the reversals of opinion prior to convergence to the truth. According to this alternative approach, Ockham's razor does not *inform* one which theory is true but is, nonetheless, the uniquely most efficient strategy for arriving at the true theory, where efficiency is a matter of minimizing reversals of opinion prior to finding the true theory.

1 Introduction

Suppose that several or even infinitely many theories are compatible with the information available. How ought one to choose among them, if at all? The traditional and intuitive answer is to choose the "simplest" and to cite *Ockham's razor* by way of justification. Simplicity, in turn, has something to do with minimization of entities, description length, causes, free parameters, independent principlies, or ad hoc hypotheses, or maximization of unity, uniformity, symmetry, testability, or explanatory power.

Insofar as Ockham's razor is widely regarded as a rule of scientific inference, it should help one to select the true theory from among the alternatives. The trouble is that it is far from clear how a fixed bias toward simplicity could do so (Morrison 2000). One wishes that simplicity could somehow indicate or inform one of the true theory, the way a compass needle indicates or informs one about direction. But since Ockham's razor always points toward simplicity, it is more like a compass needle that is frozen into a fixed position, which cannot be said to to indicate anything. Nor does it suffice to respond that a prior bias toward simplicity can be corrected, eventually, to allow for convergence to the truth, for alternative biases are also correctable in the limit.

This paper reviews some standard accounts of Ockham's razor and concludes that not one of them explains successfully how Ockham's razor helps one find the true theory any better than alternative empirical methods. Thereafter, a new explanation is presented, according to which Ockham's razor does not indicate or *inform* one of the truth like a compass but, nonetheless, keeps one on the straightest possible route to the true theory, which is the best that any inductive strategy could possibly guarantee. Indeed, no non-Ockham strategy can be said to guarantee so straight a path. Hence, a truth-seeker always has a good reason to stick with Ockham's razor even though simplicity does not indicate or inform one of the truth in the short run.

2 Standard Accounts

The point of the following review of standard explanations of Ockham's razor is just to underscore the fact that they do not connect simplicity with selecting the true theory. For the most part, the authors of the accounts fairly and explicitly specify motives other than finding the true theory—e.g., coherence, data-compression, or accurate estimation. But the official admonitions are all too easily forgotten in favor of a vague and hopeful impression that simplicity is a magical oracle that somehow extends or amplifies the information provided by the data. None of the following accounts warrants such a conclusion, even though several of them invoke the term "information" in one way or another.

2.1 Simple Virtues

Simple theories have attractive aesthetic and methodological virtues. Aesthetically, they are more unified, uniform and symmetrical and are less ad hoc or messy. Methodologically, they are more severely testable (Popper 1968, Glymour 1981, Friedman 1983, Mayo 1996), explain better (Kitcher 1981), predict better (Forster and Sober 1994), and provide a compact summary of the data (Li and Vitanyi 1997, Rissanen 1983[1]). However, if the truth happens not to be simple, then the truth does not possess the consequent virtues, either. To infer that the truth is simple because simple worlds and the theories that describe them have desirable properties is just wishful thinking, unless some further argument is given that connects these other properties with finding the true theory (van Fraassen 1981).

2.2 Bayesian Prior Probabilities

According to Bayesian methodology, one should update one's degree of belief $P(T)$ in theory T in light of evidence e according to the rule:

$$p(T|e) = \frac{p(T) \cdot p(e|T)}{p(e)}.$$

[1]Rissanen is admirably explicit that finding short explanations is an end-in-itself, rather than a means for finding the true theory.

Subjective Bayesians countenance any value whatever for the prior probability $p(T)$, so it is permissible to start with a prior probability distribution biased toward simple theories (Jeffreys 1985). But the mere adoption of such a bias hardly explains how finding the truth is facilitated better by that bias than by any other.

A more subtle Bayesian argument seems to avoid the preceding circle. Suppose that S is a simple theory that explains observation e, so that $p(e|S) \approx 1$ and that $C = \exists\theta C(\theta)$ is a competing theory that is deemed more complex due to its free parameter θ, which can be tuned to a small range of "miraculous" values over which $p(e|C(\theta)) \approx 1$. Strive, this time, to avoid any prior bias for or against simplicity. Ignorance between S and C implies that $p(S) \approx p(C)$. Hence, by the standard, Bayesian calculation:

$$\frac{p(S|e)}{p(C|e)} = \frac{p(S) \cdot p(e|S)}{p(C) \cdot p(e|C)} \approx \frac{p(e|S)}{p(e|C)} \approx \frac{1}{\int p(e|C(\theta)) \cdot p(C(\theta)|C)\ d\theta}.$$

Further ignorance about the true value of θ given that C is true implies that $p(C(\theta)|C)$ is flattish. Since $p(e|C(\theta))$ is high only over a very small range of possible values of θ and $p(C(\theta)|C)$ is flattish, the integral assumes a value near zero. So the posterior probability of the simple theory S is sharply greater than that of C (Rosenkrantz 1983). It seems, therefore, that simplicity is "truth conducive", starting from complete ignorance.

The magic evaporates when the focus shifts from theories to ways in which the alternative theories can be true. The S world carries prior probability $1/2$, whereas the prior probability of the range of worlds $C(\theta)$ in which θ is tuned to explain e is vanishingly small. That sharp, prior bias in favor of the S world is merely passed along through the Bayesian computation, accounting entirely for the sharp "confirmation" of S over C. More generally, Bayesian "ignorance" with respect to one partition of possibilities implies a strong prejudice with respect to another—e.g., "ignorance" between blue and non-blue together with ignorance between non-blue hues implies a strong bias against yellow—and that is all that is going on here. The point is not that science should be entirely free from biases. It is, rather, that direct appeal to one's bias hardly explains how that bias is better for finding the truth than alternative biases might be—every bias flatters itself.

2.3 Objective Prior Probabilities

One way to avoid the subjectivity of the preceding arguments is to select some particular prior probability distribution as special and to show that Ockham's razor follows. For example, R. Carnap (1950) viewed confirmation as a generalized notion of logical consequence in which $p(T|e)$ supposedly represents the degree to which premise e *partially entails* conclusion T. This putative degree of entailment is understood in terms of the total weight of possibilities satisfying $T\&e$ divided by the total weight of possibilities satisfying e. "Weight" is explicated in terms of probability, so there is the usual, Bayesian question of which prior probability measure to impose. Carnap imposed prior probabilities favoring uniform sequences of observable outcomes, with higher degrees of confirmation for predictions that resemble the past as a not-so-surprising result.

The trouble with Carnap's logical defense of Ockham's razor is that its prior bias toward uniformity is not preserved under linguistic translation and, hence, cannot be logical. On Carnap's proposal, a long run of green observations strongly confirms at stage n that the next observation will be green, rather than blue, because an invariantly green world is more uniform. N. Goodman (1955) responded that one can translate green/blue into grue/bleen, where grue means "green through n and blue thereafter" and bleen means "blue through n and green thereafter". A sequence of observations is uniform with respect to green/blue if and only if it is non-uniform with respect to grue/bleen, so uniformity and, hence, confirmation, is not preserved under translation. Against the objection that green/blue are "natural" predicates whereas grue/bleen involve a "magic time n", the predicates green/blue equally involve a magic time n in the grue/bleen language, so the situation is *logically* symmetrical. Therefore, Ockham's razor must be sought outside of logic.

Goodman, himself, proposed to rule out "grue-like" predicates by appealing to success in past inductions, which is a matter of history, rather than of logic. However, is hard to see how that can help if the "magic" time n still lies in the future, since then grue and green would have yielded identical success rates. A currently popular approach, called *algorithmic information theory* (Li and Vitanyi 1997), seeks uniformity not in pure logic, but in the presumably objective nature of computation. The algorithmic complexity of a string corresponds (roughly) to the length of the shortest computer program (in some fixed computer language) that generates the string. The intuitive idea is that a simple string has structure that a short program can exploit to reproduce it, whereas a complex or "random" string does not. This gives rise to the notion that good explanations are short theories that compress the data and that Ockham's razor is a matter of minimizing the sum of the lengths of the theory and of the compressed data. The proposal that one should infer the best explanation in this sense is called the *minimum description length* principle or MDL for short (Rissanen 1983). Algorithmic information theorists have developed the notion of a *universal* prior probability over bit strings with the property that more compressible strings tend to have higher prior probability. It can be shown that under certain conditions the MDL approach approximates Bayesian updating with the universal prior probability (Vitanyi and Li 2000).

Algorithmic complexity may help to explicate some slippery but important methodological concepts, such as interest, beauty, or emergence (Adriaans 2007). The focus here, however, is on the putative connection, if any, between data-compression and finding the true theory. Some proponents of the approach (e.g., Rissanen, himself) deny that there is one and urge data-compression as an alternative aim. One reason for doubt is that program length depends heavily upon the particular programming language assumed in the definition of program length. In algorithmic complexity theory, a computer language is identified with a *universal machine*, which simulates an arbitrary program p, step by step, to produce the output of p. Suppose that, in a "natural" programming language L, the shortest program p that generates a random-looking string σ is almost as long as σ itself. But now one can specify a new programming language L' whose universal machine I' is just like the universal machine I for L except that,

when presented with a very short program p', I' simulates I on the long program p, generating σ. In other words, the complexity of p can be "buried" inside of I' so that it does not show up in the L' program p' that generates σ. This arbitrariness makes it hard to take program length seriously as an indicator of how simple the world really is unless a theory of "natural" programming languages is provided—but the theory of algorithmic complexity is stated in terms of an arbitrary, Turing-equivalent programming language.[2]

Quite aside from the relativity of program length to one's choice of computer language, there is a further question about the process by which observations are encoded or transduced into the bit-strings presupposed by algorithmic complexity theory. One transducer could encode green wavelengths as 0 and blue wavelengths as 1, whereas another, grue-like transducer could reverse these assignments at some random-looking times. Algorithmic complexity judges the same world to be either extremely complex or extremely simple depending upon which transducer is employed, but no bias that depends upon mere conventions about how the data are passed along to the scientist could plausibly be an indicator of truths lying behind the data-reporting process.

Finally, and most importantly, insofar as there is any theoretical connection between simplicity and truth in the MDL story, it amounts to the selection of a universal (i.e., simplicity-biased) prior probability measure, which adds nothing to the standard, circular, Bayesian account already discussed (cf. Mitchell 1997). Therefore, it is important not to be confused by talk of bits and nats into believing that simplicity somehow provides *information* about the true theory.

2.4 Over-fitting and Empirical Estimation

Classical statisticians have an alternative account of the connection between simplicity and truth based on the concept of "over-fitting" (cf. Wasserman 2003). Since this explanation does not invoke prior probabilities at all, it is free from the shadow of circularity characteristic of Bayesian explanations. However, the underlying aim is not to choose the true theory, but to find a false theory that yields accurate empirical estimates at small sample sizes. One might expect that no theory predicts more accurately than the true theory, but that is emphatically not how "accuracy" is understood in the over-fitting literature. Hence, the over-fitting explanation of Ockham's razor avoids circular appeal to a prior simplicity bias only by adopting a skeptical or instrumentalistic stance toward theories (Forster and Sober 1994).

[2] Algorithmic complexity theorists respond to the preceding concern as follows. The first universal machine I has a program $p_{I'}$ that simulates universal machine I'. Let p' be the shortest program producing some string σ according to I'. Then the result p of chaining together the programs $p_{I'}$ and p' generates σ in L. Chaining $p_{I'}$ onto p' adds only constant length to p', so there exists a constant k that bounds the difference in length of the shortest program in L from the length the shortest program in L' that generates an arbitrary string σ. But that is scant comfort when one applies Ockham's razor in a particular instance, for it is still the case that an arbitrarily complex theory in the first universal machine could be the simplest possible theory for a second. The constants connecting systems can be arbitrarily large, so no matter how many reversals of simplicity ranking one wishes to effect, one could fish for an alternative universal machine that effects them.

To see how false theories can predict more "accurately" than true ones, imagine a marksman firing a rifle at a target from a tripod that can be locked in both the vertical and the horizontal dimensions. When both locks are off, the best marksman produces a cloud of shots centered on the bull's eye. Suppose that the inaccuracy of a marksman is measured in terms of the expected distance from the bull's eye of a single shot. Call this the marksman's "risk" (of missing the bull's eye). A good marksman's risk is due entirely to the spread or *variance* of his shots around the bull's eye. Now consider a lazy marksman, who locks the tripod in both dimensions, so every shot hits at the same point at a distance b from the bull's eye. The lazy marksman has no variance, but has *bias* b, because his average shot hits at distance b from the bull's eye. There is a critical bias $b > 0$ below which the lazy marksman is more "accurate" than the good marksman as measured by risk. Think of the bull's eye as the true value of an empirical parameter and of a shot as an empirical estimate of the parameter based on a random sample. Free aim corresponds to an empirical estimate using a complex theory. The locked tripod corresponds to a fixed empirical estimate based on a simple theory with no free parameters. The bias of the simple theory implies its falsehood (it rules out the true sampling distribution). So even if the true theory is very complex and is known in advance, risk minimization argues for using a false, over-simplified theory for estimation purposes. Hence, over-fitting hardly explains how Ockham's razor helps one find the true theory. That conclusion may sound odd in light of popular glosses of over-fitting such as the following:

> It is overwhelmingly probable that any curve that fits the data perfectly is false. Of course, this negative remark does not provide a recipe for disentangling signal from noise. We know that any curve with perfect fit is probably false, but this does not tell us which curve we should regard as true. What we would like is a method for separating the *trends* in the data from the random deviations from those trends generated by error. A solution to the curve fitting problem will provide a method of this sort (Forster and Sober 1994).

One might naturally conclude that the *trend* in the data is the *true signal* and that the aim is to strike the *true* balance between signal and noise, which only the true theory can do. However, as the authors of the passage later explain with care, over-fitting and under-fitting are defined in terms of estimation risk at a given sample size, rather than in terms of the true curve: "under-fitting" occurs when sub-optimal risk is due to bias and "over-fitting" occurs when sub-optimal risk is due to variance. Thus, as discussed above, if the sample size is small and the truth is not as simple as possible, risk minimization recommends selection of an over-simplified theory that falsely explains true signal as noise.

In the scientific case, one does not know the true sampling distribution a priori, so one does not know the bias and, hence, the risk, of using a given theory for estimation purposes. One can estimate the risk from the sample by calculating the average squared distance of data points from predictions by the theory. But the estimated risk of a complex theory is biased toward optimism because risk is estimated as fit to the

data and a sufficiently complex theory can fit the data exactly, even if the true risk of estimation is considerable due to noise. To assuage this systematic estimation bias, the risk estimate must incorporate a tax on free parameters. Then one can choose, for estimation purposes, a theory whose corrected estimated risk is minimal. This is the basic logic between such standard, classical estimation procedures as *Akaike's information criterion (AIC))* (1973), cross-validation, and *Mallow's statistic* (cf. Wasserman 2003).

Structural risk minimization (SRM) is an interesting generalization and extension of the over-fitting perspective (Vapnik 1998). In the SRM approach, one does not merely construct an (approximately) unbiased estimate of risk; one solves for objective, worst-case bounds on the chance that estimated risk differs by a given amount from actual risk. A crucial term in these bounds is called the *Vapnik Chervonenkis* dimension or VC dimension for short. The VC dimension is a measure of the range of possible samples the theory in question has the "capacity" to accommodate, which suggests a connection to simplicity and Ockham's razor. As in the over-fitting account, one can seek the "sweet spot" between simplicity (low VC-dimension) and fit (estimated risk) that minimizes the worst-case bound on the error of the risk estimate. Then one can choose the parameter setting that minimizes estimated risk within that theory.

Again, the aim is not to find the true theory. And yet, the SRM approach can explain other approaches (e.g., MDL and Bayesianism) as respectable ways to control worst-case estimation risk, eliminating the circular appeals to prior simplicity biases (Vapnik 1998). The moral is skeptical. If risk minimization is the last word on Ockham's razor, then the apparent rhetorical force of simplicity is founded upon a fundamental confusion between theories as true propositions and theories as useful instruments for controlling variability in empirical estimates.

It is tempting, at this point, to ask whether theoretical truth really matters—accurate predictions should suffice for all practical purposes. That is true so far as passive prediction is concerned. But beliefs are for guiding action and actions can alter the world so that the sampling distribution we drew our conclusions from is altered as well—perhaps dramatically. Negligible relativistic effects are amplified explosively when a sufficient quantity of uranium ore is processed. A crusade to eliminate ash trays breaks the previously observed, strong correlation between ash trays and cancer, undermining the original motivation for the policy. Theories that guide action are supposed to provide accurate *counterfactual* estimates about what would happen if the world (and, hence, the sampling distribution) were altered in various ways (Spirtes et al. 2000). An accurate estimate of the true sampling distribution is not enough in such cases, because distributions corresponding to complex theories can be arbitrarily similar to distributions corresponding to simple theories, having very different counterfactual import. This point will be sharpened below, when the details of the contemporary literature on causal discovery are discussed.

Finally, it is clear that the over-fitting story depends, essentially, upon noise in the data and, hence, in the shots at the truth taken by the estimator, since non-noisy estimates involve no variance and, hence, no bias-variance balance. However, Ockham's razor seems no less compelling in deterministic settings. One would prefer

that the connection between simplicity and theoretical truth not depend essentially upon randomness.

2.5 Convergence

The preceding explanations promise something in the short run, but theoretical truth cannot be guaranteed in the short run, even with high chance, because complex effects in nature may be two small or subtle to notice right away. Bayesians address this difficulty by circular appeal to the very bias to be explained. Risk minimization responds by shifting the focus from theoretical truth to predictive risk. A third option is to relax the demand for immediate success. Arbitrarily small, complex effects requiring free parameters to explain them can be detected eventually, as more data are collected, as more regions of the universe are explored, and as observational technology improves, so if it is assumed in advance (as in polynomial curve fitting) that there are at most finitely many such effects to be found, then at some point all the effects are noticed and Ockham's razor converges to the true theory. For example, it can be shown that, in a wide range of cases, Bayesian updating armed with a simplicity-biased prior probability does converge to the true theory in the limit. However, if indication or pointing to the true theory is too stringent to be feasible, mere convergence to the true theory is too weak to single out Ockham's razor as the best truth-finding policy in the short run. Convergence requires merely that a prior simplicity bias "wash out", eventually, in complex worlds. But the question is not how to overcome a prior simplicity bias; it is, rather, how such a bias helps one find the truth better than alternative biases. Convergence, alone, cannot answer that question, since if a method converges to the truth, so does every finite variant of that method (Salmon 1967). Hence, mere convergence says nothing about how the interests of truth-finding are *particularly* furthered by choosing the simplest theory *now*. But that is what the puzzle of simplicity is about.

3 Diagnosis

To recapitulate, the two standard notions of finding truth are (1) *indication* or *informing* of the truth in the short run and (2) *convergence* in the long run. The former aim is too strong to support an a priori explanation of Ockham's razor, since an arbitrarily complex world can appear arbitrarily simple in the short run, before the various dimensions of complexity have been detected. The latter aim is too weak to support an a priori explanation of Ockham's razor, since a prior bias toward complexity can also be washed out by further information. Therefore, if the apparent connection between simplicity and theoretical truth has an explanation, it should be sought somewhere between these two extremes: Ockham's razor should somehow help one converge to the true theory better or more efficiently than alternative strategies. Just such an account will now be presented. The basic idea is that a bias toward simplicity neither points at the truth nor merely converges to it, but converges to it in the most efficient or direct manner possible, where efficiency is measured in terms of errors, reversals of opinion,

and the time delay to such reversals.[3]

4 Traveler's Aid

To state the simplicity puzzle in its most basic terms, how could fixed, one-size-fits-all advice be guaranteed to help one find something that might be anywhere—in a sense stronger than merely guaranteeing that one will find it by exhaustive search? It happens every day. Suppose that a city dweller is lost in a small town on a long automobile journey6. He asks a local resident for directions. The resident directs him to the freeway entrance ramp. The traveler follows the advice and travels as directly as possible to the freeway, which is by far the most direct route home—in spite of a few, unavoidable curves around major geographical features.

Now suppose that the traveler stubbornly ignores the resident's advice. Indeed, suppose that, in so doing, the traveler follows a road on the true compass heading to his destination, whereas getting on the freeway requires a short jog in the opposite direction. The chosen route narrows and begins to meander through the mountains. The traveler finally concedes that it wasn't a good idea and retraces his route back to the resident. He then follows her directions to the freeway and proceeds home via the best possible route. The traveler's reward for ignoring the resident's advice is a humiliating U-turn right back to where he started, followed by all the unavoidable twists and turns encountered on the freeway over the mountains. Had he heeded the advice, he would have encountered only the unavoidable curves along the freeway. So he should have heeded it.

In connection with the simplicity puzzle, this unremarkable tale has some remarkable features.

1. The resident's advice is the *best possible* advice in the sense that it puts one on the most direct route to the goal, for violating it incurs at least one extra, initial U-turn.

2. The advice is the best possible even if it aims the traveler in the wrong direction initially.

3. The resident can give precisely the same, *fixed* advice to every stranger who asks, even though she does not know where they are headed—no Ouija board or other occult channel of information is required.

[3]The basic idea of counting mind-changes is originally due to H. Putnam (1965). It has been studied extensively in the computational learning literature— for a review cf. (Jain et al. 1999). But in that literature, the focus is on categorizing the complexities of problems rather than on singling out Ockham's razor as an optimal strategy. I viewed the matter the same way in (Kelly 1996). Schulte (1999a, 1999b) derives short-run constraints on strategies from retraction minimization. (Kelly 2002) extends the idea, based on a variant of the ordinal mind-change account due to (Freivalds and Smith 1993), but that approach does not apply to cases like curve fitting, in which theory complexity is unbounded. Subsequent steps to the present approach may be found in (Kelly 2004, 2006) and in (Kelly and Glymour 2004).

So directions to the nearest freeway entrance ramp satisfy all the apparently arcane and paradoxical demands that a successful explanation of Ockham's razor must satisfy. It remains to explain what the freeway to the truth is and how Ockham's razor keeps one on it.

5 Some Examples

For some guidance in the general developments that follow, consider some familiar examples.

Polynomial structures. Let S be a finite set of natural numbers and suppose that the truth is some unknown polynomial law:

$$y = f(x) = \sum_{i \in S} a_i x^i,$$

where for each $i \in S$, $a_i \neq 0$. Say that S is the *structure* of the law, as it determines the form of the law as it would be written in a textbook. Suppose that the problem is to infer the true structure S of the law. It is implausible to suppose that for a given value of the independent variable x one could observe the exact value of the dependent variable y, so suppose that for each queried value of x at stage k of inquiry, the scientist receives an arbitrarily small, open interval around the corresponding value of y and that repeated queries of x result in an infinite sequence of open intervals converging to $\{y\}$.

It is impossible to be sure that one has selected S correctly by any finite time, since there may be some $i \in S$ such that a_1 is set to a very small value in f, making it appear that the monomial $a_i x^i$ is missing from f. Ockham's razor urges the conclusion that $i \notin S$ until the corresponding monomial is noticed in the data.

There is a connection between the complexity of the true polynomial structure and what scientists and engineers call *effects*. Suppose that $S_0 = \{0\}$, so for some $a_i > 0$, $f_0(x) = a_i$. Let experience e_0 present a finite sequence of interval observations of the sort just described for f_0. Then there is a bit of wiggle room in each such interval, so that for some suitably small $a_1 > 0$, the curve $f_1(x) = a_1 x + a_0$ of form $S_1 = \{0, 1\}$ is compatible with e_0. Eventually, some open interval around $y = a_0$ is presented that excludes f_0. Call such information a first-order *effect*. If e_1 extends that information and presents an arbitrary, finite number of shrinking, open intervals around f_1 then, again, there exists suitably small $a_2 > 0$ such that $f_2(x) = a_2 x^2 + a_1 x + a_0$ of form $S_2 = \{0, 1, 2\}$ passes through each of the intervals presented in e_1. Eventually, the intervals tighten so that no linear curve passes between them. Call such information a second-order effect, and so forth. The number of effects presented by a world corresponds to the cardinality of S, so there is a correspondence between empirical effects and empirical complexity.

Linear dependence. Suppose that the truth is a multivariate linear law

$$y = f(x) = \sum_{i \in S} a_i x_i,$$

where for each $i \in S$, $a_i \neq 0$. Again, the problem is to infer the structure S of f. Let the data be presented as in the preceding example. As before, it seems that complexity corresponds with the cardinality of S which is connected, in turn, to the number of effects presented by nature if f is true.

Conservation laws. Consider an idealized version of explaining reactions with conservation laws, as in the theory of elementary particles (Schulte 2001, Valdez-Perez 1996). Suppose that there are n observable types of particles, and it is assumed that they interact so as to conserve n distinct quantities. In other words, each particle of type p_i carries a specific amount of each of the conserved quantities and for each of the conserved quantities, the total amount of that quantity going into an arbitrary reaction must be the total amount that emerges. Usually, one thinks of a reaction in terms of inputs and outputs; e.g.,

$$r = (p_1, p_1, p_1, p_2, p_2, p_3 \rightarrow p_1, p_1, p_2, p_3, p_3).$$

One can represent the inputs by a vector in which entry i is the number of particles of type p_i in r, and similarly for the output:

$$\begin{aligned} \mathbf{a} &= (3, 2, 1); \\ \mathbf{b} &= (2, 1, 2); \\ r &= (\mathbf{a} \rightarrow \mathbf{b}). \end{aligned}$$

A *quantity* $\mathbf{q}$ (e.g., mass or spin) is an assignment of real numbers to particle types, as in $\mathbf{q} = (1, 0, 1)$, which says that particles a_1, a_3 both carry a unit of $\mathbf{q}$ and a_2 carries none. Quantity $\mathbf{q}$ is *conserved* in r just in case the total $\mathbf{q}$ in is the total $\mathbf{q}$ out. That condition is just:

$$\sum_{i=1}^{3} q_i a_i = \sum_{i=1}^{3} q_i b_i,$$

or, in vector notation,

$$\mathbf{q} \cdot \mathbf{a} = \mathbf{q} \cdot \mathbf{b},$$

which is equivalent to:

$$\mathbf{q} \cdot (\mathbf{a} - \mathbf{b}) = 0.$$

Since reaction r enters the condition for conservation solely as the vector difference $\mathbf{a} - \mathbf{b}$, there is no harm, so far as conservation is concerned, in identifying reaction r with the difference vector:

$$\mathbf{r} = \mathbf{a} - \mathbf{b} = (1, 1, -1).$$

Then the condition for $\mathbf{r}$ conserving $\mathbf{q}$ can be rewritten succinctly as:

$$\mathbf{q} \cdot \mathbf{r} = 0,$$

which is the familiar condition for geometrical orthogonality of $\mathbf{q}$ with $\mathbf{r}$. Thus, the reactions that preserve quantity $\mathbf{q}$ are precisely the integer-valued vectors orthogonal to $\mathbf{q}$. In this example, $\mathbf{r}$ does conserve $\mathbf{q}$, for:

$$(1, 0, 1) \cdot (1, 1, -1) = 1 + 0 - 1 = 0.$$

But so do reactions $\mathbf{u} = (1, 0, -1)$ and $\mathbf{v} = (0, 1, 0)$, which are linearly independent. Since the subspace of vectors orthogonal to $\mathbf{q}$ is two-dimensional, every reaction that conserves $\mathbf{q}$ is a linear combination of $\mathbf{u}$ and $\mathbf{v}$ (e.g., $\mathbf{r} = \mathbf{u} + \mathbf{v}$). If the only conserved quantity were $\mathbf{q}$, then it would be strange to observe only scalar multiples of $\mathbf{r}$. In that case, one would expect that the possible reactions are constrained by some other conserved quantity linearly independent of $\mathbf{q}$, say $\mathbf{q}' = (0, 1, 1)$. Now the possible reactions lie along the intersection of the planes respectively orthogonal to $\mathbf{q}$ and $\mathbf{q}'$, which are precisely the scalar multiples of $\mathbf{r}$. Notice that any two linearly independent quantities orthogonal to $\mathbf{r}$ would suffice—the quantities, themselves, are not uniquely determined.

Now suppose that the problem is to determine how many quantities are conserved, assuming that some conservation theory is true and that every possible reaction is observed, eventually. Let an "effect" be the observation of a reaction linearly independent of the reactions seen so far. As in the preceding applications, effects may appear at any time but cannot be taken back after they occur and the correct answer is uniquely determined by the (finite) number of effects that occur.

In this example, favoring the answer that corresponds to the fewest effects corresponds to positing the greatest possible number of conserved quantities, which corresponds to physical practice (cf. Ford 1963). In this case, simplicity intuitions are consonant with testability and explanation, but run counter to minimization of free parameters (posited conserved quantities).

Discovering causal structure. If one does not have access to experimental data, due to cost, feasibility, or ethical considerations, one must base one's policy recommendations on purely observational data. In spite of the usual advice that correlation does not imply causation, sometimes it does. Suppose that there are n observable, jointly normal random variables and and that one wishes to determine as much as possible about causal relations among them.

The following setup is based upon (Spirtes et al. 2000). A *causal structure* associates with each unordered pair of variables $\{X, Y\}$ one of the following statements:

$$X \to Y; \quad X \leftarrow Y; \quad X \| Y;$$

interpreted, respectively, as X is a direct cause of Y, Y is a direct cause of X, and X, Y have no direct causal connection. The first two cases are *direct causal connections* and the fourth case denies such a connection. A causal structure can, therefore, be presented as a directed, acyclic *graph* (DAG) in which variables are *vertices* and arrows are direct causal connections. The notation $X - Y$ means that there is a direct connection in either direction between X and Y without specifying which. A partially oriented graph with such ambiguous edges is understood, for present purposes, to represent the disjunction of the structures that result from specifying them in each possible way.

At the core of the approach is a pair of rules for associating causal structures with probability distributions. Let p be a joint probability distribution on a set of random variables V. If S is a subset of V, let $(X \amalg Y)|S$ abbreviate that X is statistically independent of Y conditional on S in p. A sequence of variables is a *path* if each successive pair is immediately causally connected. A *collision* on a path is a variable

with arrows coming in from adjacent variables on the path (e.g., variable Y in path $X \rightarrow Y \leftarrow Z$). A path is *activated* by variable set S just in case the only variables in S that occur on the path are collisions and every collision on the path has a descendent in S. Then the key assumption relating probabilities to causal structures is simply:

$$(X \amalg Y)|S \text{ if and only if } X \text{ has no activated path to } Y.$$

Let T_p denote the set of all causal structures satisfying this relation to probability measure p.

To see why it is intuitive to associate T_p with p, suppose that $X \rightarrow Y \rightarrow Z$ and that none of these variables are in conditioning set S. Then knowing something about Z tells one something about X and knowing something about the value of X tells one something about Z. But the ultimate cause X yields no further information about Z when the intermediate cause Y is known (unless there is some other activated path between X and Y). On the other hand, suppose that the path is $X \rightarrow Y \leftarrow Z$ with collision Y. If there is no further path connecting X with Z, knowing about X says nothing about Z (they are independent causes of Y), but since X and Z may cooperate or compete in a systematic way to produce Y, knowing the value of Y together with the value of X yields some information about the corresponding setting of Z. The dependency among causes given the state of the common effect turns out to be an important clue to causal orientation.

It follows from the preceding assumptions that there is a direct connection $X - Y$ just in case X and Y are dependent conditional on each set of variables not including X, Y. There is a collision $(X \rightarrow Y \leftarrow Z)$ if $(X - Y - Z)$ holds (by the preceding rule) and $(X - Z)$ does not hold (by the preceding rule) and, furthermore, X, Z are dependent given every set of variables including Y but not X, Z (Spirtes et al. 2000, theorem 3.4). Further causal orientations may be entailed in light of background assumptions. These rules (actually, more computationally efficient heuristic versions thereof) have been implemented in "data-mining" software packages that search for causal structures governing large sets of observational variables. The key points to remember are that (1) a direct causal connection is implied by the appearance of some set of effects and (2) edge orientations depend both on the appearance of some effects and on the non-appearance in the future of further effects.

The above considerations are taken to be general. However, much of the literature on causal discovery focuses on two special cases. In the *discrete multinomial* case, say that $G \in D_g$ if and only if $G \in T_p$ and p is a discrete, joint distribution over a finite range of possible values for each variable in G. In the *linear Gaussian* case, say that $G \in L_p$ if and only if $G \in T_p$ and p is generated from G as follows: each variable in G is assumed to be a linear function of its parents, together with an extra, normally distributed, unobserved variable called an *error term* and the error terms are assumed to be uncorrelated. For brevity, say that G is *standard* for p if and only if $G \in D_p$ or $G \in L_p$. The following discussion is restricted to the standard cases because that is where matters are best understood at present.

In practice, not all variables are observable, but assume, optimistically, that all causally relevant variables are observable. Even then, in the standard cases, the DAGs

in T_p cannot possibly be distinguished from one another from samples drawn from p, so one may as well require only convergence to T_p in each p compatible with background assumptions.[4]

Moreover, statistical dependencies among variables must be inferred from finite samples, which can result in spurious causal conclusions because finite samples cannot reliably distinguish statistical independence from weak statistical dependence. Idealizing, as in the preceding examples, suppose that one receives the outputs of a data-processing laboratory that merely informs one of the dependencies.[5] that have been verified so far (at the current, growing sample size) by a standard statistical dependency test, where the null hypothesis is independence.[6] Think of an *effect* as data verifying that a partial correlation is non-zero. Absence of an effect is compatible with noticing it later (the correlation could be arbitrarily small). If it is required only that one infer the true indistinguishability class $T(p)$ for arbitrary p representable by a DAG, then effects determine the right answer.

What does Ockham say in the standard cases? Presumably, something like: assume no more dependencies than one has seen so far, unless background knowledge and other dependencies entail them. It follows, straightforwardly, that direct causal connections add complexity, and that seems intuitively right. Causal orientation of causal connections is more interesting. It may seem that causal orientation does affect complexity, because, with binary variables, a common effect depends in some manner that must be specified upon four states of the joint causes whereas a common cause affects each effect with just two states. Usually, free parameters agree with complexity, as in the curve-fitting example above. But given the overall assumptions of causal discovery, a result due to Chickering (2003) implies that these extra parameters do not correspond to potential empirical effects and, hence, do not really contribute to empirical complexity. In other words, given that no further edges are coming, one can afford to wait for data that *decide* all the discernable facts about orientation (Schulte 2007). Standard MDL procedures that tax free parameters can favor non-collisions over collisions before the data resolve the issue, risking extra surprises.[7]

[4]It is known that in the linear, non-Gaussian case, causal structure can be recovered uniquely if there are no unobserved variables (Shimizu et al. 2006). The same may be true in the non-linear Gaussian case.

[5]In the standard cases, it is known that all of the over-identifying constraints follow from conditional independence constraints (Richardson and Spirtes 2002). That is known to be false in the linear, non-Gaussian case (Shimizu et al. 2006), so in that case simplicity must be relativized to a wider range of potential effects. Indeed, in the linear, non-Gaussian case, the set of possible empirical effects is so rich that there are no proper inclusion relations among the sets of effects corresponding to alternative causal models, so the simplicity ranking is flat.

[6]Also, the significance level is tuned down at a sufficiently slow rate to ensure that the test converges in probability to the right answer. At the end of the paper, some of the issues that arise in a serious application to statistical model selection are raised.

[7]A similar issue arises in the inference of regular sets from positive examples. The most liberal automaton is a one-state universal acceptor with a loop for each input character. But assuming that the language is learned from positive examples only, that is the most complex hypothesis in terms of empirical effects. In typical scientific applications, such as curve fitting, extra parameters imply extra effects. But not always, and then it is the effects, rather than the parameters, that determine retraction

For example, when there are three variables X, Y, Z and $(X - Y - Z)$ is known, then, excluding unobserved causes, there are two equivalence classes of graphs, the collision orientation $(X \rightarrow Y \leftarrow Z)$ in one class C and all the other orientations in the complementary class $\neg C$. Looking at the total set of implied dependencies for C, C', it turns out that the only differences are that C entails $\neg((X \amalg Z)|Y)$ but not $\neg(X \amalg Z)$, whereas $\neg C$ entails $\neg(X \amalg Z)$ but not $\neg((X \amalg Z)|Y)$, so there is no inclusion relationship between the dependencies characterizing C and the dependencies characterizing $\neg C$. No such inclusion relationship obtains in the example under consideration: both hypotheses are among the simplest compatible with the data, so Ockham's razor does not choose among them. Moreover, given that $(X - Y - Z)$ has been verified, nature must present either $\neg(X \amalg Z)$ or $\neg((X \amalg Z)|Y)$ eventually (given that the causal truth can be represented by some graph over the observable variables) so it seems that science can and should wait for nature to resolve the matter instead of racing ahead—and that is just how Ockham's razor is interpreted in the following discussion. Regardless of which effect nature elects to present, it remains possible, thereafter, to present the other one as well, in which case each variable is connected immediately to every other and one can say nothing from the orientation rules about causal directionality. This situation involves more effects than either of the two preceding cases, but another direct causal connection is also added as a syntactic indication of the increase in complexity.

In the standard cases, the preceding evolution can result in spectacular reversals of causal conclusions as experience increases, not just in terms of truth, but in terms of practical consequences as well. Suppose that it is known that $(X \rightarrow Y - Z)$ and none of these variables has yet exhibited any dependence with W. Then discovery of $\neg((X \amalg Z)|Y)$, background knowledge, and Ockham's razor unambiguously imply $(X \rightarrow Y \leftarrow Z)$, a golden invitation to exploit Z to control Y. Indeed, the connections may be obvious and strong, inviting one to invest serious resources to exploit Z. But the conclusion rests entirely on Ockham's razor, for the further discovery of $\neg(X \amalg Z)$ is incompatible with $(X \rightarrow Y \leftarrow Z)$ and the new Ockham answer is $(X \rightarrow Y - Z)$ with edge $(X - Z)$ added. Further discovery that $\neg((Z \amalg W)|X, Y)$ and that $\neg((Y \amalg W)|Z)$ results in the conclusion $Y \rightarrow Z \leftarrow W$, reversing the original conclusion that Y can be controlled by Z.[8] The orientation of the direct causal connection $Y - Z$ can be flipped n times in sequence by assuming causes $X_0, \ldots X_n$ of Y in the role of X and potential collisions $W_0, \ldots, W_n$ in the role of W. There is no way that a convergent strategy can avoid such discrete flips of $Y - Z$; they are an ineluctable feature of the problem of determining the efficacy of Z on Y from non-experimental data, no matter how strong the estimate of the strength of the cause $Y \rightarrow Z$ is prior to the reversal. Indeed, standard causal discovery algorithms exhibit the diachronic retractions just discussed in computer simulations. The practical consequences of getting the edge orientation wrong are hardly negligible, for if Z does not cause Y, the policy of manipulating Z to achieve results for Y will have no benefits at all to justify its cost. Indeed, in the case just described, sample size imposes no non-trivial bound on arbitrarily large mis-

efficiency.

[8] I am indebted to Richard Scheines for suggesting this example.

estimates of the effectiveness of Y in controlling Z (cf. Robins et al. 2003, Zhang and Spirtes 2003). A skeptical stance toward causal inference is tempting in the standard cases:

> We could try to learn the correct causal graph from data but this is dangerous. In fact it is impossible with two variables. With more than two variables there are methods that can find the causal graph under certain assumptions but they are large sample methods and, furthermore, there is no way to ever know if the sample size you have is large enough to make the methods reliable (Wasserman 2003, p. 275).

This skepticism is one more symptom of the unrealizable demand that simplicity should reliably point toward or inform one of the true theoretical structure, a popular—if infeasible—view both in statistics and philosophy (Goldman 1986, Mayo 1996, Dretske 1981). The approach developed below is quite different: insofar as finding the truth makes reversals of opinion unavoidable, they are not only justified but laudable—whereas, insofar as they are avoidable, they should be avoided. So the best possible strategies are those that converge to the truth with as few course-reversals as possible. That is what standard causal inference algorithms tend to do, and it is the best they could possibly do in the standard cases.

To summarize, an adequate explanation of Ockham's razor should isolate what is common to the simplicity intuitions in standard examples like the preceding ones and should also explain how favoring the simplest theory compatible with experience helps one find the truth more directly or efficiently than competing strategies when infallibility or even probable infallibility is hopeless. Such an explanation, along the lines of the freeway metaphor, will now be presented. First, simplicity and efficient convergence to the truth must be defined with mathematical rigor and then a proper proof must be provided that Ockham's razor is the most efficient possible strategy for converging to the truth.

6 Inference of Theoretical Structure

In light of the preceding examples, say that an empirical *effect* is experience that (1) may take arbitrarily long to appear due to its subtlety or difficulty to produce and that (2) never disappears once it has been seen. Furthermore, (3) at most finitely many effects appear for eternity and (4) the correct theoretical structure is uniquely determined by the (finite) set of effects one encounters for eternity. In light of (4), one may as well understand the problem of finding the true theory as a matter of inferring which finite set of effects (corresponding to some structure or other) one will encounter for eternity.

Accordingly, let E be a countable set of potential effects satisfying (1-4) which, for the time being, will not be analyzed further (a deeper analysis, explaining what effects are, is provided below). Let Ω denote the set of all finite subsets of E. It may happen that one knows *a priori* that some theoretical structures are impossible (e.g.,

not every finite set of statistical dependencies corresponds to a causal graph). Let $\Gamma \subseteq \Omega$ be the set of possible sets of effects compatible with background knowledge. An empirical *world* w is an infinite sequence of mutually disjoint, finite subsets of E that converges to $\emptyset$, where the finite set $w(i)$ corresponds to the set of as-yet unobserved effects encountered for the first time at stage i of inquiry. Let W denote the set of all such worlds. If no new effects are encountered at i, then $w(i)$ is empty. Let $w|k = (w_0, \ldots, w_{k-1})$, the finite initial segment of w of length k. The finite set of all effects presented by w (or by finite sequence $e = w|k$) is given by:

$$S_w = \bigcup_{i=0}^{\infty} w(i); \quad S_e = \bigcup_{i=0}^{k-1} w(i).$$

For each $w \in W$ define the *modulus* of w to be the first moment from which no more new effects appear:

$$\mu(w) = \text{ the least } k \text{ such that } S_w = S_{w|k}.$$

The background restriction $\Gamma \subseteq \Omega$ on sets of effects can be viewed as a material restriction on empirical worlds as follows:

$$K_\Gamma = \{w \in W : S_w \in \Gamma\}.$$

Recall that each theoretical structure T corresponds uniquely to some finite set S of effects. Let theoretical structure T_S corresponding to finite set $S \subseteq E$ be identified with the set of all worlds in which T_S is correct—namely, the set of all worlds that present exactly S:

$$T_S = \{w \in W : S_w = S\}.$$

The set:

$$\Pi_\Gamma = \{T_S : S \in \Gamma\}$$

partitions W into mutually exclusive and exhaustive alternative propositions called *potential answers* and will be referred to as the *question* posed by the problem of inferring theoretical structures. Then T_{S_w} is the unique answer in Π_Γ that contains (is true of) w. Finally, the *theoretical structure inference problem* with *possible structures* Γ is represented by the ordered pair:

$$\mathcal{P}_\Gamma = (K_\Gamma, \Pi_\Gamma),$$

where Π_Γ is the empirical *question* and K_Γ is the empirical *background presupposition.*

Every concept and proposition that follows is relative to Γ so, to eliminate some symbolic clutter, think of Γ as a "global variable" held fixed in the background, to be referred to as clarity demands.

7 Empirical Strategies and Convergent Solutions

What makes science unavoidably fallible is that one does not get to see the entire empirical world w all at once; rather, one sees incrementally longer, finite, initial segments of w as time passes. The set of all possible finite sequences the scientist might see as time passes is given by:

$$F_\Gamma = \{w|i : w \in K_\Gamma \text{ and } i \in N\}.$$

When e is a finite, initial segment of e' (i.e., there exists i such that $e = e'|i$), say that $e \leq e'$. When e is a sub-sequence but not necessarily an initial segment of e', then abuse notation by writing $e \subseteq e'$. Let $e * e'$ denote sequence concatenation, where it is always understood that e is finite and that e' may be finite or infinite. Finally (only in the proofs in the Appendix), if x is some generic set-theoretic object, let x^∞ denote the infinite sequence in which only x occurs.

An empirical *strategy* M for problem $\mathcal{P}_\Gamma$ is a mapping of type:

$$M : F_\Gamma \to \Pi \cup \{\text{`?'}\}.^9$$

In other words, M maps each finite sequence $e \in F_\Gamma$ either to an answer $T_S \in \Pi$ or to '?', indicating refusal to choose an answer. Then in world $w \in K_\Gamma$, M produces the unending sequence of outputs:

$$M[w] = (M(w|0), M(w|1), M(w|2), \ldots),$$

where the square brackets are a reminder that M does not get to see w "all at once".

After seeing finite input sequence e, background presupposition K_Γ entails that one must live in a world $w \in K_\Gamma$ that extends e, so let:

$$K_\Gamma|e = \{w \in K_\Gamma : w \geq e\}$$

denote the set of all such extensions. Then one may resrict Π_Γ to the answers compatible with e as follows:

$$\Pi_\Gamma|e = \{T \in \Pi_\Gamma : T \cap K_\Gamma|e \neq \emptyset\}.$$

Say that M *solves* $\mathcal{P}_\Gamma$ *in the limit* given e if and only if for each $w \in K_\Gamma|e$,

$$\lim_{i \to \infty} M(w|i) = T_{S_w},$$

in which case, say that M is a *convergent solution* to $\mathcal{P}_\Gamma$ given e. A *convergent solution* to $\mathcal{P}_\Gamma$ is just a convergent solution given the empty sequence ().

One obvious, convergent solution to $\mathcal{P}_\Omega$ (i.e., no finite set of effects is ruled out *a priori*) is just:

$$M(e) = T_{S_e},$$

[9] In a more realistic setup, M could output disjunctions of answers in Π_Γ or degrees of belief distributed over Π_Γ. The ideas that follow extend to both situations.

for if $w \in K_\Gamma$, new effects stop appearing, eventually—say by stage n—so for all $m \geq n$, $M(w|m) = T_{S_{w|m}} = T_w$. But there are infinitely many alternative, convergent solutions as well—each finite variant of the obvious, convergent solution is a convergent solution—and it is less trivial to say how and in what sense the obvious strategy helps one to find the truth better than these do. That is the question answered by the following argument.

8 Empirical Complexity Defined in Terms of Effects

If $\Gamma = \Omega$, as in the polynomial structure problem, then an obvious definition of the *empirical complexity* of world w given e is

$$c(w, e) = |S_w| - |S_e|,$$

the number of new effects presented by w after the end of e (cf. Kelly 2007). When $\Gamma \subset \Omega$, as in the causal inference problem (some finite sets of partial correlations correspond to no causal graph), a slightly more general approach is required.[10] The basic idea is that effects, relative to a problem, correspond to successive opportunities to force the scientist to switch from one answer to another. Restrict Γ to those sets of effects compatible with e:

$$\Gamma|e = \{S \in \Gamma : S_e \subseteq S\}.$$

This set includes all the possible theoretical structures that might serve as potential interpretations of what has been presented by e. Say that a *path* in $\Gamma|e$ is a finite, non-repetitive, ascending sequence of elements of $\Gamma|e$. If $S, S' \in \Gamma|e$, let $\pi_e(S, S')$ denote the set of all paths in π_e that start with S and terminate with S'. Then $\pi_e(*, S')$ denotes all paths in $\Gamma|e$ that terminate with S' and $\pi_e(S, *)$ denotes all paths in $\Gamma|e$ that start with S. So $\pi_e(*, S)$ represents all the possible paths nature might have taken to S from some arbitrary starting point in $\Gamma|e$. Then for $e \in F_\Gamma$, $w \in K_\Gamma|e$, and $P \subseteq K_\Gamma$, define *empirical complexity* as follows:

$$\begin{aligned} c(w, e) &= \max\{\text{length}(p) : p \in \pi_e(*, S_w)\} - 1; \\ c(P, e) &= \min\{c(w, e) : w \in P \cap K_\Gamma|e\}. \end{aligned}$$

Then since $(S) \in \pi_e(*, S)$ if $S \in \Gamma|e$ and lengths are discrete, it is immediate that:

Proposition 1 (empirical complexity is non-negative) *If $w \in K_\Gamma|e, P \in \Pi|e$, then $c(w, e), c(P, e)$ assume values in the natural numbers.*

Hence, answers with complexity zero are simplest. Define:

$$(\Gamma|e)_{\min} = \{S \in \Gamma|e : \text{for all } S' \in \Gamma|e, S' \not\subset S\},$$

and say that S is *minimally compatible* with e if and only if $S \in (\Gamma|e)_{\min}$.

[10] E.g., suppose that $\Gamma = \{\emptyset, \{a, b\}\}$. Then seeing a implies that one will see b, so a and b are not independent effects. They are more like correlated aspects of one effect, so they should not be counted separately.

Proposition 2 (characterization of zero complexity) *Let $w \in K_\Gamma|e$ and $e \in F_\Gamma$ and $T_S \in \Pi_\Gamma|e$. Then:*

1. *$c(w, e) = 0$ if and only if $S_w \in (\Gamma|e)_{min}$;*
2. *$c(T_S, e) = 0$ if and only if $S \in (\Gamma|e)_{min}$.*

Maximum simplicity is minimum complexity. Borrowing a standard re-scaling trick from information theory, one can convert complexity degrees to *simplicity degrees* in the unit interval as follows:

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

Unconditional complexity and simplicity are definable as:

$$\begin{aligned} c(P) &= c(P, ()); \\ s(P) &= s(P, ()). \end{aligned}$$

9 Ockham's Razor

The *Ockham* answer given e, if it exists, is the unique answer $T \in \Pi_\Gamma|e$ such that $c(T, e)$ is minimal over all alternative theories $T' \in \Pi_\Gamma|e$. In light of proposition 1, the Ockham answer is the unique answer in $T \in \Pi_\Gamma|e$ such that $c(T, e) = 0$. Empirical strategy M *satifies Ockham's razor* (or is *Ockham*, for short) at e iff

$$M(e) \text{ is Ockham given } e \text{ or } M(e) = \text{`?'}.^{11}$$

Furthermore, M is *Ockham* from e onward iff M is Ockham at each e' extending e; and M is *Ockham* if M is Ockham at each $e \in F_\Gamma$.

When S is in $\Gamma|e$ and S is a subset of each $R \in \Gamma|e$, say that S is the *minimum* in $\Gamma|e$. The Ockham answer, if it exists, can be characterized both in terms of uniquely minimal compatibility and in terms of minimality.

Proposition 3 (Ockham answer characterization) *Let $e \in F_\Gamma$ and $T_S \in \Pi_\Gamma|e$. Then the following statements are equivalent:*

1. *T_S is Ockham given e;*
2. *$(\Gamma|e)_{min} = \{S\}$;*
3. *S is the minimum in $\Gamma|e$.*

[11] If M is allowed to output disjunctions of answers in Π_Γ, then Ockham's razor requires that $\bigcup\{T_S : S \in (\Gamma|e)_{\min}\} \subseteq M(e)$.

10 Stalwartness and Eventual Inormativeness

Ockham's razor does not constrain suspension of judgment in any way, but it would be odd to adopt the Ockham answer T at e and then to drop T later, even though T is still the Ockham answer—further effect-free experience would only seem to "confirm" the truth of T. Accordingly, let $e \in F_\Gamma$ and let $e * S$ denote the extended, finite input sequence along which finite $S \subseteq E$ is reported right after the end of e. Say that strategy M is *stalwart* at $e * S$ if and only if for each answer $T \in \Pi_\Gamma$, if $M(e) = T$ and $M(e*S) \neq T$ then T is not the Ockham answer at $e*S$ (i.e., an answer is dropped only if it is not the Ockham answer when it is dropped). As with the Ockham property, itself, one may speak of M being stalwart from e onward or as just being stalwart, which means that M is stalwart at each e.

Similarly, it would be too skeptical never to conclude that no more effects are forthcoming, no matter how much effect-free experience has been collected. Accordingly, say that a strategy is *eventually informative* from e onward if there is no world $w \in K_\Gamma|e$ on which M converges to '?'. Then M is *eventually informative* if M is eventually informative from the empty input sequence onward.

Finally, a *normal* Ockham strategy from e onward is an eventually informative, stalwart, Ockham strategy from e onward and a normal Ockham strategy is normally Ockham from the empty sequence onward. The normal Ockham strategies are intuitively quite plausible. Such a strategy M may wait for a while but eventually chooses the Ockham answer and retains it until it is no longer Ockham. Furthermore, after each new effect is encountered, there is some finite amount of effect-free experience that lulls M to plump for the simplest theory once again. That is pretty much what people and animals do, and also describes, approximately, the behavior of a simplicity-biased Bayesian agent who selects only the theory whose posterior probability is above some high threshold. But plausibility and rhetoric are not the points at issue—finding the true theory is—so it is more pertinent to observe that normally Ockham strategies are, at least, guaranteed to converge to the truth.

Proposition 4 (normal Ockham Convergence) *If M is normally Ockham for $\mathcal{P}_\Gamma$ from e onward, then M is a solution to P_Γ from e onward.*

Furthermore, eventual informativeness is a necessary condition for being a solution, for a strategy that is not eventually informative evidently fails to converge to any theory in some world w:

Proposition 5 (convergence implies eventual informativeness) *If M solves $\mathcal{P}_\Gamma$ from e onward, then M is eventually informative from e onward.*

So there is always a motive to be eventually informative, if one wishes to find the truth at all. The same is not clear, yet, for Ockham's razor and stalwartness, since there are infinitely many eventually informative, non-Ockham solutions. For example, an alternative solution favors some set S of size fifty until the anticipated fifty effects fail to appear for ten thousand stages, after which it concedes defeat and reverts back to Ockham's razor. So it remains to determine how, if at all, Ockham strategies are better at finding the true theory than these variants are.

11 Epistemic Costs of Convergence

As in the parable of the traveler, the aim is to show that normal Ockham strategies are the properly most *efficient* strategies for finding the truth, where efficiency is a matter of minimizing epistemic costs en route to convergence to the truth.

(1) Since the aim is to find the truth, an evident cost of inquiry is the number of times one selects a false answer prior to convergence.

(2) Nobody likes it when science changes its tune, but the intrinsic fallibility of theory choice makes some reversals of course unavoidable. Therefore, the best one can demand of an optimally truth-conducive strategy for theory choice is that it not reverse course more than necessary. A method *retracts* its previous answer whenever its current answer fails to entail its previous answer.[12] In the narrow context of methods that produce answers in $\Pi \cup \{\text{`?'}\}$ (where '?' is interpreted as the most uninformative answer $W$), strategy $M$ *retracts* at $e * S$ if and only if $M(e) \neq \text{`?'}$ and $M(e * S) \neq M(e)$.

Retractions have been studied as an objective feature of the complexity of problems, both computational and empirical. H. Putnam (1965) noticed that the concept of computability can be extended by allowing Turing machines to "take back" their answers some fixed number of times and called properties having such generalized decision procedures *n-trial predicates*. In a similar spirit, computational learning theorists speak of *mind-changes* and have studied bounds on the number of mind-changes required to find the truth in various empirical questions (Jain et al. 1999). The body of results obtained makes it clear that mind-changes are an invariant feature both of empirical and of purely formal inquiry. The idea here is to shift the focus from problems back to methods.

(3) A third cost of inquiry is elapsed time to each retraction. Theories are used to derive further conclusions and these conclusions tend to accumulate through time. When the theory is retracted, all of these subsidiary conclusions are called into question with it. The accompanying angst is not merely practical but cognitive and theoretical, and it should be minimized by getting retractions over with as soon as possible. Also, aside from such subsidiary conclusions, there is a tragic aspect of insouciant hubris or of unwittingly "living a lie" when one is destined to retract in the future, even if the retracted theory happens to be true. The insouciance is all the worse if one is destined to retract many times. It would be better to relieve the hubris as soon as possible.[13]

Taken together, errors, retractions, and retraction times paint a fairly representative picture of what might be termed the quality or directness of a strategy's connection with or route to the truth. If e is an input stream, let the *cumulative cost* or *loss* of strategy M on $e \in K_\Gamma$ be given by the pair $\lambda(M, w) = (b, \tau)$, where b is the total

[12] In belief revision theory, a belief change that adds content is an *expansion*, a belief change that removes content is a *contraction* and a belief change that does any of the above is a *revision* (Gardenfors 1988). In that terminology, a retraction is any revision in which content is lost and, hence, may be expressed as a non-trivial contraction followed by an expansion. In spite of this connection, belief revision theorists have not begun to examine the normative consequences of minimizing contractions (or of finding the truth).

[13] Elimination of hubris as soon as possible is a Platonic theme, arising, for example, in the *Meno*.

number of false answers produced by M along e and τ is the sequence of times at which the successive retractions performed by M along e occur. The length of τ (which is finite for convergent strategies) is, then, the total number of retractions performed.

It would be a shame if Ockham's razor were to rest upon some idiosyncratic, subjective weighting of errors, retractions, and retraction times but, happily, the proposed argument for Ockham's razor rests only on comparisons that agree in all dimensions (i.e. on *Pareto* comparisons). First, consider retractions and retraction times. If σ, τ are finite, ascending sequences of natural numbers, define:[14]

$$\sigma \leq \tau \quad \text{iff} \quad \begin{array}{l}\text{there exists a subsequence } \gamma \text{ of } \tau \text{ such that} \\ \text{for each } i \leq \text{length}(\sigma), \sigma(i) \leq \gamma(i).\end{array}$$

Hence, $(1, 3, 7) \leq (2, 3, 4, 8)$ in virtue of sub-sequence $(2, 3, 8)$. Then if (b, σ) and (c, τ) are both cumulative costs, define:

$$\begin{array}{lll}(b, \sigma) \leq (c, \tau) & \text{iff} & b \leq c \text{ and } \sigma \leq \tau; \\ (b, \sigma) \equiv (c, \tau) & \text{iff} & (b, \sigma) \leq (c, \tau) \text{ and } (c, \tau) \leq (b, \sigma); \\ (b, \sigma) < (c, \tau) & \text{iff} & (b, \sigma) \leq (c, \tau) \text{ and } (c, \tau) \not\leq (b, \sigma).\end{array}$$

12 Worst-case Cost Bounds

Non-Ockham strategies do not necessarily incur greater costs prior to convergence: Nature could be so kind as to present the extra effects posited by a non-Ockham strategy immediately, in which case it would beat all Ockham competitors in the race to the truth. The same point is familiar in the theory of computational complexity: if an inefficient algorithm is optimized for speed on a single input, even the best algorithms will fail to dominate it in terms of computational resources expended before the answer is found. For that reason, algorithmic efficiency is ordinarily understood to be a matter of optimizing worst-case cost (Garey and Johnson 1979). Adoption of a similar approach to empirical strategies and to Ockham's razor requires some careful attention to worst-case bounds on total costs of inquiry. Let ω denote the first infinite ordinal number. A *potential cost bound* is a pair (b, σ), where $b \leq \omega$ and σ is a finite or infinite, non-descending sequence of entries $\leq \omega$ in which no finite entry occurs more than once. If (b, σ) is a cost vector and (c, τ) is a cost bound, then $(b, \sigma) \leq (c, \tau)$ can be defined just as for cost vectors, themselves. Cost bounds $(c, \tau), (d, \gamma)$ may now be compared as follows:

$$\begin{array}{lll}(c, \tau) \leq (d, \gamma) & \text{iff} & \text{for each cost vector } (b, \sigma), \text{ if } (b, \sigma) \leq (c, \tau) \text{ then } (b, \sigma) \leq (d, \gamma); \\ (c, \tau) \equiv (d, \gamma) & \text{iff} & (c, \tau) \leq (d, \gamma) \text{ and } (d, \gamma) \leq (c, \tau); \\ (c, \tau) < (d, \gamma) & \text{iff} & (c, \tau) \leq (d, \gamma) \text{ and } (d, \gamma) \not\leq (c, \tau).\end{array}$$

Thus, for example, $(4, (2)) < (\omega, (2, \omega)) < (\omega, (0, 1, 2, \ldots)) \equiv (\omega, (\omega, \omega, \omega, \ldots))$. Now, each set C of cost vectors has a unique (up to equivalence) least upper bound $\sup(C)$

[14] Context will distinguish whether $\leq$ denotes this relation or the initial segment relation.

among the potential upper bounds (Kelly 2007). Suppose that finite input sequence e has already been seen. Then one knows that possibilities incompatible with e cannot happen, so define the *worst-case cost* of M at e as:

$$\lambda_e(M) = \sup_{w \in K_\Gamma | e} \lambda(M, w).$$

13 Relative Efficiency

The final hurdle in arguing for the efficiency of Ockham's razor is the triviality of worst-case cost bounds: for each e and for each convergent solution M to $\mathcal{P}_\Omega$, the worst-case cost bound achieved by M at e is just:

$$\lambda_e(M) = (\omega, (\omega, \omega, \omega, \ldots)).$$

For let m be an arbitrary, natural number and let $\{a_0, \ldots a_n, \ldots\}$ be an arbitrary enumeration of the set E of possible effects. Nature can present $\emptyset$ until, on pain of not converging to the truth, M produces answer $T_\emptyset$ at least m times consecutively. Then Nature can present $\{a_0\}$ followed by repetitions of $\emptyset$ until, on pain of not converging to the truth, M produces $T_{\{e_0\}}$ at least m times, consecutively, etc. Hence, normal Ockham strategies are not distinguished from alternative, convergent solutions in terms of worst-case efficiency.

Again, a similar difficulty is familiar in the assessment of computer algorithms: typically, the number of steps required by an algorithm is not finitely bounded across all possible inputs since larger inputs require more steps of computation. The problem disappears if worst-case bounds are taken over problem instances (inputs) of a given size, rather than over all possible problem instances, for there are at most finitely many such inputs, so the worst-case performance of an algorithm over inputs of a given size is guaranteed to exist (Garey and Johnson 1979). Then one compares the worst-case cost bounds over each instance size as instance size increases. In $\mathcal{P}_\Omega$, every problem instance (input stream) is of infinite length, so length is no longer a useful notion of instance size. But *empirical complexity* $c(w, e)$, defined above, is such a notion. Furthermore, each normal Ockham strategy is a convergent solution that retracts at most n times over instances of empirical complexity n, so non-trivial cost bounds are achievable. Accordingly, define the nth *empirical complexity class* $C_e(n)$ of worlds in $K_\Gamma|e$ as:

$$C_e(n) = \{w \in K_\Gamma | e : c(w, e) = n\}.$$

Then one may define the *worst-case cost* of strategy M given e over $C_e(n)$ as follows:

$$\lambda_e(M, n) = \sup_{w \in C_e(n)} \lambda(M, w).$$

Now it is possible to compare strategies in terms of their worst-case costs over problem instances of various sizes.

$$\begin{array}{lll} M \leq_e M' & \text{iff} & (\forall n)\ \lambda_e(M, n) \leq \lambda_e(M', n); \\ M <_e M' & \text{iff} & M \leq_e M' \text{ and } M' \not\leq_e M; \\ M \prec_e M' & \text{iff} & (\forall n) \text{ if } C_e(n) \neq \emptyset \text{ then } \lambda_e(M, n) < \lambda_e(M', n). \end{array}$$

When $M \leq_e M'$, say that M is *as efficient as* M' given e. If $M <_e M'$ say that M is (weakly) *more efficient than* M' given e. Finally, when $M \prec_e M'$, say that M is *strongly* more efficient than M'.

The concept "more efficient than" is a hybrid, lying between dominance (doing as well in each world and better in some world) and worst-case (minimax) reasoning (doing better in the worst case overall). The hybrid character of "more efficient than" is just what is required to expose the superiority of normal Ockham strategies: dominance is too strict because non-Ockham strategies can get lucky and the worst-case overall is too loose because even normal Ockham strategies guarantee no nontrivial, worst-case bound.

14 Optimality

Suppose that a scientist facing problem $\mathcal{P}_\Gamma$ has been using strategy M for a while and that the final datum in finite input sequence $e = (x_1, \ldots, x_n)$ has just been presented. Let the data e_- observed just prior to the end of e be defined by:

$$\begin{aligned} e_-(()) &= (); \\ e_-((S_0, \ldots, S_n, S_{n+1})) &= (S_0, \ldots, S_n). \end{aligned}$$

At e, past actions along e_- directed by the scientist's strategy M can no longer be "taken back" at e. Hence, an alternative strategy M' cannot possibly be adopted and implemented at e unless its outputs agree with those of M all along e_-, in which case write $M \asymp_{e_-} M'$. Define:

> M is *optimal* at e iff M is a solution at e and for each strategy $M' \asymp_{e_-} M$ that is a solution at e, $M \leq_e M'$.

It is not enough for strategy M to be optimal at e. If the user of M is not to have reason to dispense with M later, it had best be the case that M is always optimal:

> M is *always optimal* iff for each $e \in F_\Gamma$, M is optimal at e.

When e is empty, say simply that M is *optimal*.

15 Unique Optimality of Normal Ockham Strategies

Here is the promised, non-circular argument, based entirely on truth-finding efficiency, for always following Ockham's razor. The results are relative to a fixed problem $\mathcal{P}_\Gamma$.

Theorem 6 (optimality) *If M is a normal Ockham strategy, then M is always an optimal solution.*

But that is not enough. One trouble with much of the standard literature on Ockham's razor is that it shows only that Ockham's razor is sufficient for, say, convergence to the truth, but what is required is an argument that Ockham's razor is *necessary* for optimal truth-conduciveness. Here is such an argument:

Theorem 7 (unique optimality) *Let $e \in F_\Gamma$. If M' is a convergent solution that violates Ockham's razor for the first time at e, then every strategy $M' \asymp_{e\_} M$ that is always normally Ockham is a more efficient solution than M' at e;*

The proof (cf. the Appendix) is closely analogous to the parable of the traveler discussed above, with extra cases added to allow for the possibility of branching freeway ramps. The two theorems jointly imply the following corollary, which summarizes the proposed argument for Ockham's razor.

Corollary 8 (Ockham efficiency characterization) *The following statements are equivalent:*

1. *M is always a normal, Ockham strategy;*
2. *M is always an optimal solution;*
3. *no solution M' is ever a more efficient solution than M.*

In other words, the normally Ockham methods are coextensive with the always efficient strategies and with the strategies such that no alternative strategy is ever more efficient.[15]

16 A General Definition of Empirical Complexity

In the preceding development, empirical effects were stipulated, by appeal to intuition, for each of the examples considered and the effects appealed to were quite different from case to case. Empirical effects will now be defined in a general way that explains the apparently ad hoc choices in the examples.[16] The idea is to locate effects and, hence, empirical complexity, in the power of nature to force an arbitrary, convergent method to change its answer to the problem to be solved. Thus, empirical complexity is a structural, semantic feature of the problem to be solved, rather than a matter of syntactic or computational brevity. As such, it is invariant under grue-like translations.

An empirical *problem* is a pair $\mathcal{P} = (K, \Pi)$ where K is now an arbitrary set of infinite sequences of *inputs* drawn from some arbitrary set I and Π is an arbitrary partition of K. The elements of I are just inputs (e.g., boolean bits in a binary coding scheme for meter readings or what-not). Answers are just arbitrary, mutually exclusive and exhaustive propositions over K. This is a very general conception of empirical problems. An empirical strategy takes finite, initial segments of elements of K as

[15] It is a further question whether it is always better to follow Ockham's razor even after violating it. The answer is negative: let $\Gamma = \{\{a\}, \{b\}, \{b, c\}\}$, let $M((\emptyset)) = M'((\emptyset)) = T_{\{b\}}$, and let $M((\emptyset, \emptyset)) =$ '?' whereas $M'((\emptyset, \emptyset)) = T_{\{b\}}$. Then M uses one extra retraction in reaching theory $T_{\{b,c\}}$ after seeing $e = (\emptyset, \emptyset)$, so $\lambda_e(M, 2) \not\leq \lambda_e(M', 2)$.

There is still something to say in favor of Ockham, however. Method M is *strongly Ockham* at e if M never favors an answer T_S such that some alternative S' compatible with e has a longer path through $\Gamma|e$. Then one can argue, along the same lines, that at each strong Ockham violation and at each violation of stalwartness by M, some alternative, convergent M' is more efficient.

[16] Preliminary versions of the following ideas can be found in (Kelly 2007).

inputs and outputs potential answers in $\Pi \cup \{\text{`?'}\}$ in response. Convergence and many other concepts like $M[w]$, $F|e$, $K|e$ extend to this more general setting in an obvious way. It remains to reconstruct Γ and the concepts that presuppose it.

Let some problem $\mathcal{P} = (K, \Pi)$ be understood to be fixed in the background. An *answer pattern* is a finite sequence of elements of Π without immediate repetitions (non-immediate repetitions are allowed). Pattern s is *forcible* by nature given e of length k if and only if for each convergent solution M, there exists $w \in K|e$ such that from stage k onward, M produces a sequence of answers of which s is a sub-sequence. In other words, no convergent solution can avoid producing the successive conclusions in s in the worst case, given e. Let Δ_e denote the set of all patterns forcible by nature given e. Restrict attention to problems $\mathcal{P}$ such that:

Axiom 1 (forcible path convergence) *for each $w \in K$, $\lim_{i\to\infty} \Delta_{w|i}$ exists, in the sense that the sequence $\{\Delta_{w|i} : i \geq 0\}$ stabilizes to a fixed set eventually.*

Define:

$$\begin{aligned} \Delta_w &= \lim_{i\to\infty} \Delta_{w|i}; \\ \Gamma|e &= \{\Delta_w : w \in K|e\}; \\ \Gamma &= \Gamma|(). \end{aligned}$$

Elements of $\Gamma|e$ serve the same purpose as before—they are the possible, permanent stopping places for nature given e because each element Δ_w of Γ is converged to in world w and world w is compatible with e. Call the elements of Γ *empirical problem states*. Define *epistemic accessibility* among states in the following way:

> $X \leq Y$ if and only if for each $e \in F$ such that $\Delta_e = X$, there exists $e' \in F$ such that $e' \geq e$ and $\Delta_{e'} = Y$.

Let $\pi_e(X, Y)$ denote the set of all $\leq$-paths between two states in Γ with respect to order $\geq$. Finally, define $c(w, e)$ and $c(P, e)$ in terms of these paths, as before, with Δ_w in place of S_w. Let the new, more general concepts so defined be marked with a prime, as in $c'(w, e)$, to distinguish them from the notions defined in terms of stipulated effects.[17]

The general concept of empirical complexity just defined agrees with the effect-based definition.

Proposition 9 (recovery) *Let $(\Gamma', \leq')$ be constructed from problem (K_Γ, Π_Γ) in the manner just described. Then for each $e \in F$, the mapping $\phi(S_w) = \Delta_w$ is well-defined and witnesses:*

$$(\Gamma|e, \subseteq) \text{ is order-isomorphic to } (\Gamma'|e, \leq').$$

[17]The structure $(\Gamma'|e, \leq')$ can be viewed as a model of an epistemic logic, in which elements of $\Gamma'|e$ are worlds and increasing information e "chops down" the set of worlds, in accordance with what is known as *dynamic epistemic logic* (Van Benthem 2006). What is new is a motivated constraint on accessibility and the idea that empirical complexity is a matter of maximum accessibility path length into a world.

Thus, for each $w \in K$:

$$\begin{aligned} c'(w, e) &= c(w, e); \\ c'(P, e) &= c(P, e). \end{aligned}$$

Something far more interesting is also true. Let (K, Π) be any one of the examples considered above (e.g., the conservation law problem) *prior* to being represented in the form (K_Γ, Π_Γ). The problem (K, Π) does not wear its empirical effects "on its sleeve"—the reactions may be presented in some obscure or even grue-like code that is highly misleading. But it is still the case that applying the preceding construction directly to (K, Π) results in $(\Gamma', \leq')$ order-isomorphic to $(\Gamma, \subseteq)$ and, hence, to the same empirical complexity concept $c(w, e)$. Depending on the structure of the problem $\mathcal{P}$, empirical complexity reflects extra parameters, extra conserved quantities, extra causes, etc., regardless of how gerrymandered the data-gathering process happens to be.[18] Therefore, the set Γ assumed in each case reflects more than mere notation, convention, or whim—it is an intrinsic, structural feature of the original problem that survives every sort of re-description that preserves the meanings of the background presupposition K and of the question Π.

17 A Word on Stochastic Applications

In real curve-fitting and causal discovery problems, the data are not merely inexact, but random. The above treatment of these problems in terms of collapsing open intervals around the true observations is intended only as an indication of how a fully statistical story might go (think of the intervals as idealizations of high probability quantiles). This section sketches some promising pieces of a fully statistical version of the theory.

A *world* is an objective probability distribution of interest (e.g., the distribution induced by a polynomial curve with normally distributed measurement error). A *question* is a partition of worlds. A *method* maps samples of arbitrary size to answers to the question. A method is *consistent* just in case the probability that the method produces the true answer converges to unity as sample size increases. The *retraction in chance* of answer T by method M at sample size $n + 1$ in distribution p is definable as the drop in chance that M outputs T from sample size n to sample size $n + 1$:

$$p^n(M = T) - p^{n+1}(M = T).$$

The total retractions in chance in p are the sums of the retractions in chance for all $T \in \Pi$, and for all sample sizes n.

A sequence of answers is *forcible in chance* if and only if nature can force an arbitrary, consistent method to produce the first answer in the sequence with arbitrarily high chance followed by the second answer in the sequence with arbitrarily high chance, etc. For a simplistic illustration of how this works (a similar argument applies in causal

[18] Contrast this result with the preceding discussion of algorithmic complexity, which is relative both to the choice of a computer language and to the encoding of observations.

discovery), let K consist of independent, bivariate normal means of fixed, known variance and let the possible answers correspond to the number of non-zero components of the true mean vector $\mu = (\mu_X, \mu_Y)$, so answer T_i is the set of all $p \in K$ such that exactly i components of μ are non-zero. Let M be a consistent method. Let $p_0 \in T_0$ and let $\epsilon > 0$ be as small as desired. Since M is consistent, there is a sample size n_0 such that

$$p_0^{n_0}(M = T_0) > 1 - \epsilon.$$

Since the chance of a fixed measurable event is continuous in μ, there exists $p_1 \in T_1$ such that

$$p_1^{n_0}(M = T_0) > 1 - \epsilon.$$

Since M is consistent, there exists sample size $n_1 > n_0$ such that

$$p_1^{n_1}(M = T_1) > 1 - \epsilon.$$

Again, by continuity, there exists $p_2 \in T_2$ such that:

$$\begin{aligned} p_2^{n_0}(M = T_0) &> 1 - \epsilon; \\ p_2^{n_1}(M = T_1) &> 1 - \epsilon. \end{aligned}$$

Again, by consistency, there exists $n_2 > n_1$ such that:

$$p_2^{n_1}(M = T_2) > 1 - \epsilon.$$

Hence, the sequence of answers (T_0, T_1, T_n) is forcible by nature in chance. The only premise required for this forcing argument is convergence to the truth, so a Bayesian's degree of belief in each successive answer can also be forced arbitrarily high. A Bayesian's retraction in chance of T at $n + 1$ in p can be measured in terms of the drop in his expected degree of belief in T at sample size $n+1$ in p.[19] Simplicity can be defined in terms of statistically forcible sequences of answers, just as in the deterministic case. It remains to recover a suitable analogue of corollary 8 in the setting just described.

18 Ockham, Fallibility, and "Information"

Like it or not, we do infer theoretical forms, they are subject to the problem of induction, and we may have to take them back. Indeed, there is no bound on the number of times science might have to change its tune as new layers of complexity are successively revealed in nature. Ockham's razor merely keeps science on the straightest path to the truth, crooked as it may be. For millennia, fallibility has been thought to *undermine* the justification of science, resulting in the usual, circular, metaphysical, or skeptically evasive justifications of Ockham's razor. The proposed account reverses the traditional

[19] Bayesians are sub-optimal: moving from ignorance (.5/.5) to knowledge (.99/.01) implies a retraction of nearly one half that could have been avoided by modeling ignorance as (0/0), as Schafer (1976) proposed.

reasoning—Ockham's razor is justified not because it points straight at the truth, but because its path to the truth, albeit crooked, is uniquely straightest. The Ockham path is straightest because its unavoidable kinks are due to the intrinsic fallibility of theory choice. Therefore, the ineluctable fallibility of theory choice justifies, rather than undermines, Ockham's razor. That is why the proposed account is not circular, metaphysical, or evasive of the connection between method and true theoretical structure.

Ockham's razor is, nonetheless, so firmly anchored in our animal spirits that it feels as if, somehow, simplicity *informs* us about the true theory in a way that the data alone do not, just as a compass needle augments the information provided by one's native sense of direction. Then there must be some benevolent cosmic cause behind the correlation of simplicity and truth—a mysterious, undetected agency that operates across evolutionary time and across domains from subatomic particles to cell metabolism to social policy—the irony of defending Ockham's razor with such hidden, metaphysical fancies notwithstanding (Koons 2000).

Therein lies a concern about the association of information-theoretic terminology with Ockham's razor, as in the MDL and SRM approaches. When information theory is applied to a telephone line, as originally intended, it really has something to do with informative signals from a source. If one wishes to minimize expected message length to maximize the line's capacity, it makes sense to adopt shorter codes for more frequently sent words. But applications of information theory to theory choice are not about sending information over a line. They are a formal recipe either for constructing short codes for plausible explanations or (contrariwise) for assigning high plausibility to short explanations. Either way, the ultimate connection between simplicity and truth is stipulated rather than explanatory. But since the stipulated connection is formulated in the language of "information", it is all too readily confused, in the popular mind, with a deep theoretical revelation that simplicity *does* provide a magical communication channel to the truth that amplifies the only real information available—the data. Better not to mention "information" at all than to kindle that perennial wish.

19 Acknowledgements

This work has benefitted from helpful comments from and patient discussions with (in alphabetical order), Pieter Adriaans, the very helpful Anonymous Reviewer, Seth Casana, Stephen Fancsali, Clark Glymour, Conor Mayo-Wilson, Joe Ramsey, Richard Scheines, Cosma Shalizi, Peter Spirtes, Larry Wasserman, and Jiji Zhang. The errors, of course, are my responsibility alone.

20 Bibliography

Adriaans, P. (2007) "The philosophy of learning, the cooperative computational universe", in *Handbook of the Philosophy of Information*, P. Adriaans, and J. van Benthem, eds. Dordrecht: Elsevier.

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

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

Chickering, D. (2003) "Optimal Structure Identification with Greedy Search", JMLR, 3: 507-554.

Dretske, F. (1981) *Knowledge and the Flow of Information*, Cambridge: M.I.T. Press.

Forster, M. R. 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.

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

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

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

Peter Gärdenfors (1988) *Knowledge in Flux*, Cambridge: MIT Press.

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

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

Goldman, A. (1986) *Epistemology and Cognition*, Cambridge: Harvard University Press.

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

Jeffreys, H. (1985) *Theory of Probability*, Third edition, Oxford: Clarendon Press.

Jain, S., Osherson, D., Royer, J. and Sharma, A (1999) *Systems That Learn: An Introduction to Learning Theory.* Camabridga: M.I.T. Press.

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

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

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

Kelly, K. (2007) "Ockham's Razor, Empirical Complexity, and Truth-finding Efficiency," *Theoretical Computer Science* 317: 227-249.

Kitcher, P. (1981) "Explanatory Unification," *Philosophy of Science*, 48, 507-31.

Koons, R. (2000) "The Incompatibility of Naturalism and Scientific Realism," In *Naturalism: A Critical Appraisal*, edited by J. Moreland and W. Craig, London: Routledge.

Kuhn, T. (1962) *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press.

Mayo, D. (1996) *Error and the Growth of Experimental Knowledge*, Chicago: University of Chicago Press.

Morrison, M. (2000) *Unifying Scientific Theories: Physical Concepts and Mathematical Structures*, Cambridge: Cambridge University Press.

Li, M. and Vitanyi, P. (1997) *An Introduction to Kolmogorov Complexity and Its Applications*, New York: Springer.

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

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

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

Richardson, T. and Spirtes, R. (2002) "Ancestral Graph Markov Models," *Annals of Statistics* 30: pp. 962-1030.

Rissanen, J. (1983) "A universal prior for integers and estimation by inimum description length," *The Annals of Statistics*, 11: 416-431.

Robins, J., Scheines, R., Spirtes, P., and Wasserman, L. (1999) "Uniform Consistency in Causal Inference," *Biometrika* 90:491-515.

Rosenkrantz, R. (1983) "Why Glymour is a Bayesian," in *Testing Scientific Theories*, J. Earman ed., Minneapolis: Universitiy of Minnesota Press.

Salmon, W. (1967) *The Logic of Scientific Inference*, Pittsburgh: University of Pittsburgh Press.

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

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

Schulte, O. (2001) "Inferring Conservation Laws in Particle Physics: A Case Study in the Problem of Induction," *The British Journal for the Philosophy of Science*, 51: 771-806.

Schulte, O., Luo, W., and Griner, R. (2007) "Mind Change Optimal Learning of Bayes Net Structure". Unpublished manuscript.

Schwarz, G. (1978) "Estimating the Dimension of a Model," *The Annals of Statistics*, 6:461-464.

Shimizu, S., Hoher, P., Hyv/"arinen, A., and Kerminen, A. (2006) "A Linear Non-Gaussian Acyclic Model for Causal Discovery", *Journal of Machine Learning Research* 7: pp. 2003-2030.

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

Valdez-Perez, R. and Zytkow, J. (1996) "Systematic Generation of Constituent Models of Particle Families, " *Physical Review*, 54:2102-2110.

van Benthem, J. (2006) "Epistemic Logic and Epistemology, the state of their affairs," *Philosophical Studies* 128: 49 - 76.

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

Vapnik, V. (1998) *Statistical Learning Theory*, New York: John Wiley and Sons, Ltd.

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

Wasserman, L. (2003)*All of Statistics: A Concise Course in Statistical Inference.* New York: Springer.

Zhang, J. and Spirtes, P. (2003) Strong Faithfulness and Uniform Consistency in Causal Inference, in *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, 632-639. Morgan Kaufmann.

21 Appendix: Proofs and Lemmas

Proof of proposition 1. Immediate. $\dashv$

Proof of proposition 2. For (1), suppose that $w \in K_\Gamma$ and that $c(w, e) = 0$. Then each $p \in \pi_e(*, S_w)$ has unit length and terminates with S_w, so since $S_w \in \Gamma|e$, $\pi_e(*, S_w) = \{(S_w)\}$. Hence, $S_w \in (\Gamma|e)_{\min}$. Conversely, suppose that $S_w \in (\Gamma|e)_{\min}$.

Then for each $R \in \Gamma|e$, $R \not\subset S_w$. So $\pi_e(*, S_w) = \{(S_w)\}$, so $c(w, e) = 0$.

For (2), suppose that $T_S \in \Pi_\Gamma|e$ and that $c(T_S, e) = 0$. Then there exists $w \in T_S$ such that $c(w, e) = 0$. So by (1), $S = S_w \in (\Gamma|e)_{\min}$. Conversely, suppose that $S \in (\Gamma|e)_{\min}$. Let $w = e * (S \setminus S_e) * \emptyset^\infty$. Then $w \in T_S$ and $S_w = S \in (\Gamma|e)_{\min}$. So by part (1), $c(w, e) = 0$. So $c(T_S, e) = 0$. $\dashv$

Proof of proposition 3. The equivalence (1) $\Leftrightarrow$ (2) is by part 2 of proposition 2. Equivalence (2) $\Leftrightarrow$ (3) is an elementary property of $\subseteq$ over a collection of finite sets. $\dashv$

Proof of proposition 4. Suppose that $w \in K_\Gamma|e$. Let $k \geq \mu(w)$ so that $S_w = S_{w|k}$. Then $(\Gamma|(w|k))_{\min} = \{S_w\}$, so by proposition 3, T_{S_w} is Ockham given e. Since M is eventually informative from e onward, M produces some answer from Π_Γ after e in w. Since M is Ockham from e onward, the answer M chooses is T_{S_w}. Since M is stalwart from e onward, M never drops T_{S_w} thereafter. So $\lim_{i\to\infty} M(w|i) = T_{S_w}$. $\dashv$

Proof of proposition 5. Immediate. $\dashv$

Proof of theorem 6. Let M be a strategy that is always normally Ockham. Hence, M is a solution, by proposition 4. Let $e \in F_\Gamma$ have length k. Let M' be an arbitrary solution given e such that $M' \asymp_{e_-} M$. Let d denote the maximum, over all $w \in C_e(0)$, of the number of errors committed by both M and M' along e_- and let the retraction times for M, M' along e_- be $(r_1, \ldots, r_m)$. Consider the case in which M retracts at e. Since M is always stalwart and Ockham, it follows that $T_S = M(e_-)$ is Ockham at e_- but not at e, so by proposition 3, $(\Gamma|e_-)_{\min} = \{S\}$ and $(\Gamma|e)_{\min} \neq \{S\}$. So by lemma 9, $S \notin (\Gamma|e)_{\min}$.

Suppose that $w \in C_e(0)$. By parts (1) and (2) of lemma 2, M never retracts or commits an error from stage $k + 1$ onward in w. Hence:

$$\lambda_e(M, 0) \leq (d, (r_1, \ldots, r_m, k)).$$

Since M retracts at e, there exists $S \subseteq E$ such that

$$M(e_-) = M'(e_-) = T_S \neq M(e).$$

Since $e \in F_\Gamma$, lemma 6 implies that there exists $w' \in C_e(0)$ such that $S_{w'} \neq S$. Since $M'(e_-) = T_S$ and M' is a solution, it follows that M' retracts after e_- along w'. So since M' commits at least d errors in some world in $C_0(e)$ (they do not have to be committed in w' to affect the worst-case lower bound):

$$\lambda_e(M, 0) \leq (d, (r_1, \ldots, r_m, k)) \leq \lambda_e(M', 0).$$

Now suppose that $w \in C_e(n + 1)$. By part 1 of lemma 2, M retracts at most $n + 1$ times in w from $k + 1$ onward, so allowing for the extra retraction at k:

$$\lambda_e(M, n + 1) \leq (\omega, (r_1, \ldots, r_m, k, \underbrace{\omega, \ldots, \omega}_{n+1})).$$

Suppose that there exists $w \in C_e(n+1)$. By lemma 1, there exists path $(S_0, \ldots, S_{n+1})$ and $w' \in C_e(n)$ such that (1) $S_0 \in (\Gamma|e)_{\min}$ and (2) $S_{w'} = S_w$ and for each $i \leq n+1$, M' produces T_{S_i} at least b times in immediate succession in w' after the end of e. It was shown above that $S \notin (\Gamma|e)_{\min}$. So, since $S_0 \in (\Gamma|e)_{\min}$, it follows that $S_0 \neq S$. So M' retracts from T_S to T_{S_0} no sooner than stage k. By incrementing b, w can be chosen so that the retractions of M' between answers $T_{S_0}, \ldots, T_{S_{n+1}}$ along w' occur arbitrarily late and M' produces arbitrarily many errors along w', so:

$$\lambda_e(M', n+1) \geq (\omega, (r_1, \ldots, r_m, k, \underbrace{\omega, \ldots, \omega}_{n+1}))) \geq \lambda_e(M, n).$$

For the case in which M does not retract at e, simply erase the k's from the bounds in the preceding argument. $\dashv$

Proof of theorem 7. Let $e \in$ be of length k. Let M' be given and let $M \asymp_e M'$ be a strategy that is normally Ockham from e' onward. Hence, M is a solution given e', by proposition 4. Let $b \geq 0$. Let d denote the maximum, over $w \in C_e(0)$ of the number of errors committed by both M and M' along e_- and let the retraction times for M, M' along e_- be $(r_1, \ldots, r_m)$.

Suppose that solution M' violates Ockham's razor at $e \in F_\Gamma$ of length k but not at any proper, initial segment of e. So $T_S = M'(e)$ is not Ockham at e. By proposition 3, $(\Gamma|e)_{\min} \neq \{S\}$. Thus, there exists $S' \neq S$ such that $S' \in (\Gamma|e)_{\min}$. Let $w \in C_e(0)$. Since M is Ockham from e onward, if $M(e) = T_{S'}$ then $T_{S'}$ is Ockham at e so, by proposition 3, $(\Gamma|e)_{\min} = \{S'\}$ and, hence, $w \in T_{S'}$ so M commits no error at e in w. So by lemma 2:

$$\lambda_e(M, 0) \quad \leq \quad (d, (r_1, \ldots, r_m, k)).$$

Let $w \in T_{S'}$. Then $w \in C_0(e)$, since $S' \in (\Gamma|e)_{\min}$. Since M' is a solution, M' converges to $T_{S'}$ in w, so M' retracts T_S properly later than stage k in w. Since M' commits at least d errors in some world in $C_0(e)$ (they do not have to be committed in w' to affect the worst-case lower bound):

$$\lambda_e(M', 0) > (d, (r_1, \ldots, r_m, k)) \geq \lambda_e(M, 0).$$

Suppose that there exists $w \in C_e(n+1)$. As in the proof of proposition 6,

$$\lambda_e(M, n+1) \leq \lambda_e(M', n+1).$$

Next, suppose that M' violates stalwartness at e of length k. Then $T_S = M'(e_-) = M(e_-)$ is Ockham at e, so by proposition 3, $(\Gamma|e)_{\min} = \{S\}$. Since M is stalwart from e onward, $M(e) = T_S$, so M does not retract at e. Let $w \in C_e(0)$. Then, by proposition 2, $S_w \in (\Gamma|e)_{\min}$, so $S_w = S$. So M commits no error at e. So by lemma 2:

$$\begin{aligned} \lambda_e(M, 0) &\leq (d, (r_1, \ldots, r_m)); \\ \lambda_e(M, n+1) &\leq (\omega, (r_1, \ldots, r_m, \underbrace{\omega, \ldots, \omega}_{n+1}))). \end{aligned}$$

By lemma 6, there exists $w \in C_e(0)$. Since M' retracts at e:

$$\lambda_e(M', 0) \geq (d, (r_1, \ldots, r_m, k)) > \lambda_e(M, 0).$$

Suppose that there exists $w \in C_e(n+1)$. By lemma 1, there exists $w' \in C_e(n+1)$ such that $S_{w'} = S_w$ and in w', M' produces $n+2$ distinct blocks of answers in $K_\Gamma|e$ after the end of e, each block having length at least b. So in w', M' retracts at e and at the end of each block prior to the terminal block. By incrementing b, w can be chosen so that the retractions occur arbitrarily late and there are arbitrarily many errors, so including the assumed retraction at k,

$$\lambda_e(M', n+1) \quad \geq \quad (\omega, (r_1, \ldots, r_m, k, \underbrace{\omega, \ldots, \omega}_{n+1})) > \lambda_e(M, n+1).$$

$\dashv$

Proof of corollary 8. The equivalence (1) $\Rightarrow$ (2) is by proposition 4 and theorem 6. Equivalence (2) $\Rightarrow$ (3) is immediate from the definitions. (3) $\Rightarrow$ (1) is by proposition 5 and theorem 7. $\dashv$

Proof of proposition 9. When the problem is (K_Γ, Π_Γ), the following relations hold:

$$\begin{array}{lll} \text{i.} & S_w = S_{w'} & \text{if and only if} \quad \Delta_w = \Delta_{w'}; \\ \text{ii.} & S_w \subseteq S_{w'} & \text{if and only if} \quad \Delta_w \leq' \Delta_{w'}. \end{array}$$

For (i), suppose that $S_w = S_{w'}$. Suppose that $s = (T_{S_1}, \ldots, T_{S_k}) \in \Delta_w$. So $S_1, \ldots, S_k$ are distinct elements of Γ and for each m, nature can force the successive, distinct answers $T_{S_1}, \ldots, T_{S_k}$ from an arbitrary, convergent method M starting from $w|m$. Hence, $S_{w'} = S_w \subseteq S_1 \subset \ldots \subset S_k$. So for each m, nature can force $S_1 \subset \ldots \subset S_k$ from M starting with $w'|m$, so $s \in \Delta_{w'}$. Thus, $\Delta_w \subseteq \Delta_{w'}$. For the converse inclusion, reverse the roles of w and w'. For the converse implication, suppose that $\Delta_w = \Delta_{w'}$. Suppose that $s = (T_{S_1}, \ldots, T_{S_k}) \in \Delta_w$. Then for each m, nature can force M to produce s wtarting from $w|m$. Since M is a convergent solution, there exists $m' \geq n$ such that $M(w|m') = T_{S_w}$. Nature can still force M to produce $T_{S_w} * s$ starting from $w|m'$. Hence, (a) for each $s \in \Delta_w$, for each m, $T_{S_w} * s$ is forcible by nature starting from $w|m$. By a similar argument, (a) also holds as well for w'. Call that statement (a$'$). Since for each m, nature can force (T_{S_w}) given $w|m$, $(T_{S_w}) \in \Delta_w$. Suppose, for reductio, that $S_w \neq S_{w'}$. Then by (a$'$), $(T_{S_{w'}}, T_{S_w}) \in \Delta_{w'}$ and by (a), $(T_{S_w}, T_{S_{w'}}, T_{S_w}) \in \Delta_w$ and, hence, is forcible. So $S_w \subseteq S_{w'} \subseteq S_w$, so $S_w = S_{w'}$. Contradiction.

For (ii), suppose that $S_w \subseteq S_{w'}$. Suppose that $s \in \Delta_w$. Then for each m, sequence $s = (T_{S_1}, \ldots, T_{S_k})$ is forcible starting with $w'|m$. Let $\Delta_e = \Delta_w$. Recall from case (i) that $(T_{S_w}) \in \Delta_w$, so T_{S_w} is forcible starting with e and, hence, $S_e \subseteq S_w$. Since $S_w \subseteq S_{w'}$, choose $e' \geq e$ such that $S_{e'} = S_{w'}$. Since forcibility in $\mathcal{P}_\Gamma$ depends only on presentation of effects, $\Delta_{e'} = \Delta_{w'}$. Hence, $\Delta_e \leq \Delta_{e'}$. Conversely, suppose that

$S_w \not\subseteq S_{w'}$, so let effect $a \in S_w \setminus S_{w'}$. Choose e such that $S_e = S_w$, since S_w is finite. Since forcibility in $\mathcal{P}_\Gamma$ depends only on presentation of effects, $\Delta_e = \Delta_w$. Recall from part (i) that $(T_{S_{w'}}) \in \Delta_{w'}$. But $a \in S_e \setminus S_{e'} = S_{w'}$, so Nature cannot force $(T_{S_{w'}})$ at arbitrary $e' \geq e$. Hence, $\Delta_w \not\leq \Delta_{w'}$, completing the proof of (ii).

Let $e \in F$. The mapping $\phi : \Gamma|e \to \Gamma'|e$ defined by:

$$\phi(S_w) = \Delta_w$$

is well-defined, by property (i). To see that ϕ is onto, let $\Delta_w \in \Gamma'|e$. Then $w \in K_\Gamma|e$, so $S_w \in \Gamma|e$, so $\phi(S_w) = \Delta_w$. To see that ϕ is injective, let $S_w \neq S_{w'}$. Without loss of generality, let $a \in S_w \setminus S_{w'}$. Then $(T_{S_w}) \in \Delta_w \setminus \Delta_{w'}$, so $\phi(S_w) = \Delta_w \neq \Delta_{w'} = \phi(S_{w'})$. Finally, ϕ is order-preserving by (ii), so ϕ is the required order-isomorphism. It follows immediately that $c(w, e) = c'(w, e)$ and $c(P, e) = c'(P, e)$. $\dashv$

Lemma 1 (lower cost bound for arbitrary solutions) *If M is a convergent solution given $e \in F_\Gamma$ and $w \in C_e(n)$, and $b > 0$, then there exists $w' \in C_e(n)$ and there exists path $(S_0, \ldots, S_n) \in \pi_e(*, S_{w'})$ such that*

1. *$S_0 \in (\Gamma|e)_{min}$;*
2. *$S_{w'} = S_w$;*
3. *M produces T_{S_i} at least b times in immediate succession after the end of e (if $n = 0$) or after the end of e_{i-1} (if $n > 0$) in w';*

Proof. Let $e \in F_\Gamma$ and $w \in C_e(n)$. Then there exists a path $p = (S_0, \ldots, S_n) \in \pi_e(*, S_w)$ of length $n+1$ whose length is maximal among paths in $\pi_e(*, S_w)$. Property (1) follows from lemma 3. Let $w' = (e_n * \emptyset^\infty)$. For property (2), note that by part 2 of lemma 4, $S_{e_n} = S_n$. By construction, $S_{w'} = S_{e_n}$. Since $p = (S_0, \ldots, S_n) \in \pi_e(*, S_w)$, $S_n = S_w$. So $S_w = S_{w'}$. Property (3) is part 3 of lemma 4. $\dashv$

Lemma 2 (upper cost bound for normal Ockham strategies) *Suppose that $e \in F_\Gamma$ and $e \in K_\Gamma|e$ and M is normally Ockham from e onward, where length$(e) = k$. Then for each $n \geq 0$:*

1. *if $c(w, e) \leq n$, then M retracts at most n times in w from stage $k + 1$ onward.*
2. *if $c(w, e) = 0$, then M commits no error in w from stage k onward.*

Proof. For statement (1), suppose that M retracts $> n$ times along $w \in K_\Gamma|e$ from stage $k + 1$ onward, say at stages $j_0 + 1 < \ldots < j_n + 1$, where $k \leq j_0$. Let index i range between 0 and n. Then there exists $(S_0, \ldots, S_n)$ such that $M(w|j_i) = T_{S_i}$ and $M(w|j_i + 1) \neq T_{S_i}$. Since M is a normal Ockham method, proposition 3 implies that

$$\begin{array}{lrcl} \text{i.} & (\Gamma|(w|j_i)))_{\min} & = & \{S_i\}; \\ \text{ii.} & (\Gamma|(w|(j_i + 1)))_{\min} & \neq & \{S_i\}. \end{array}$$

Also, by part 1 of lemma 10, there exists $j > j_{n+1}$ such that $(\Gamma|(w|j))_{\min} = \{S_w\}$. Then, by (i) and lemma 8, $S_0 \subseteq \ldots \subseteq S_n \subseteq S_w$. So by (ii) and lemma 9, $S_0 \subset \ldots \subset S_n \subset S_w$, so $p = (S_0, \ldots, S_n, S_w) \in \pi_e(*, S_w)$. Observe that length$(p) = n + 2$. So $c(w, e) > n$. For statement (2), suppose that $c(w, e) = 0$. Then by part 1 of proposition 2, $S_w \in (\Gamma|e)_{\min}$. Let $k' \geq k$. By part 2 of lemma 10, $S_w \in (\Gamma|(w|k'))_{\min}$. So by proposition 3 and the fact that M is Ockham from $w|k$ onward, either $M(w|k) = T_{S_w}$ or $M(w|k) =$ '?', neither of which is an error. $\dashv$

Lemma 3 (minimal beginning of maximal path) *Suppose $e \in F_\Gamma$ and $S \in \Gamma|e$ and $p \in \pi_e(*, S)$ has maximal length in $\pi_e(*, S)$. Then $p(0) \in (\Gamma|e)_{min}$.*

Proof. Suppose that $p(0) \notin (\Gamma|e)_{\min}$. Since $p \in \pi_e(*, S)$, $p(0) \in \Gamma|e$, so there exists $S' \subset p(0)$ such that $S' \in \Gamma|e$. Then $S' * p \in \pi_e(*, S)$, so p does not have maximal length in $\pi_e(*, S)$. $\dashv$

Lemma 4 (optimal strategy for nature) *If M is a convergent solution given $e \in F_\Gamma$ and if $w \in C_e(n)$, and $p = (S_0, \ldots, S_n) \in \pi_e(*, S_w)$ and $b > 0$, then there exists sequence $(e_0, \ldots, e_n)$ of elements of $F_\Gamma|e$ such that for each i such that $0 \leq i \leq n$ and for each j such that $0 \leq j < n$:*

1. *$e < e_j < e_{j+1}$;*
2. *$S_{e_i} = S_i$;*
3. *M produces T_{S_i} at least b times in immediate succession in e_i after the end of e (if $n = 0$) or after the end of e_{i-1} along w (if $n > 0$);*
4. *$(e_n * \emptyset^\infty) \in K_\Gamma \cap C_e(n)$.*

Proof by induction on length(p). Base case: $p = ()$. Then the lemma is trivially true, because $() \notin \Pi_e(*, S_w)$. Induction: let $p = (S_0, \ldots, S_n) \in \Pi_e(*, S_w)$. Let $e_{-1} = e$ and let $S_{-1} = S_e$. Let $R_n = S_n \setminus S_{e_{n-1}}$. Let e_n be the least initial segment of $w_n = (e_{n-1} * R_n * \emptyset^\infty)$ extending $e_{n-1} * R_n$ by which M selects theory T_{S_n} at least b times without interruption after the end of e_{n-1}. There exists such an e_n, since M is a convergent solution and $S_{w_n} = S_n \in \Gamma|e$, so $w_n \in K_\Gamma|e$. Properties (1-3) are immediate by construction and by the induction hypothesis. For property (4), observe that $(e_n * \emptyset^\infty) = (e_{n-1} * R_n * \emptyset^\infty) = w_n$. By the induction hypothesis, $S_{w_n} = S_{e_n} = S_{e_{n-1}} \cup R_n = S_{n-1} \cup (S_n \setminus S_{n-1}) = S_n$. So since $(S_0, \ldots, S_n)$ is maximal in $\pi_e(*, S_w) = \pi_e(*, S_n)$, $w_n \in K_\Gamma \cap C_e(n)$. $\dashv$

Lemma 5 (non-triviality) *Let $e \in F_\Gamma$. Then $(\Gamma|e)_{min} \neq \emptyset$.*

Proof. Since $e \in F_\Gamma$, there exists $w \in K_\Gamma|e$ such that $e < w$. Hence, $S_w \in \Gamma|e$. Since each member of $\Gamma|e$ is finite and $\Gamma|e \neq \emptyset$, let $S' \in \Gamma|e$ be $\subseteq$-minimal in $\Gamma|e$, so $S' \in (\Gamma|e)_{\min}$. $\dashv$

Lemma 6 (simple alternative world) *Suppose that $e \in F_\Gamma$ and $(\Gamma|e)_{min} \neq \{S\}$. Then there exists $w \in C_e(0)$ such that $S_w \neq S$.*

Proof. Since $(\Gamma|e)_{\min} \neq \{S\}$, lemma 5 implies that there exists $S' \in (\Gamma|e)_{\min}$ such that $S' \neq S$. Let $w = (e*(S' \setminus S_e)*\emptyset^\infty)$. By construction, $S_w = S' \in (\Gamma|e)_{\min}$ and $w \in K_\Gamma|e$, so $w \in C_e(0)$, by proposition 2. $\dashv$

Lemma 7 (simple world existence) *Let $e \in F_\Gamma$. Then there exists $w \in K_\Gamma|e$ such that $c(w, e) = 0$.*

Proof. Let $S \in (\Gamma|e)_{\min}$, by lemma 5. Let $w' = (e * (S \setminus S_e) * \emptyset^\infty)$. Then $S_w = S \in (\Gamma|e)_{\min}$. So by proposition 2, $c(w', e) = 0$. $\dashv$

Lemma 8 (monotonicity) *Suppose that $e, e' \in F_\Gamma$. Then:*

if $e \leq e'$ and $(\Gamma|e)_{min} = \{S\}$ and $(\Gamma|e')_{min} = \{S'\}$, then $S \subseteq S'$.

Proof. Since $S' \in \Gamma|e'$ and $e \leq e'$, $S' \in \Gamma|e$. Since $(\Gamma|e)_{\min} = \{S\}$, S is minimal in $\Gamma|e$ by proposition 3, so $S \subseteq S'$. $\dashv$

Lemma 9 (down and out) *Suppose that $e, e', e'' \in F_\Gamma$. Then:*

if $e < e' \leq e''$ and $(\Gamma|e)_{min} = \{S\}$ and $(\Gamma|e')_{min} \neq \{S\}$, then $S \notin \Gamma|e''$.

Proof. Suppose that $e, e', e'' \in F_\Gamma$ and $e < e' \leq e''$ and $(\Gamma|e)_{\min} = \{S\}$ and $S \in \Gamma|e''$. It suffices to show that $(\Gamma|e')_{\min} = \{S\}$. Since $S \in \Gamma|e''$ and $e < e'$, $S \in \Gamma|e'$. Suppose $S' \in \Gamma|e'$. Then $S' \in \Gamma|e$, since $e < e'$. Since $(\Gamma|e)_{\min} = \{S\}$, proposition 3 yields that $S \subseteq S'$. So $S' \not\subset S$, so $S \in (\Gamma|e')_{\min}$. Now, suppose that $R \in (\Gamma|e')_{\min}$. Then $R \in \Gamma|e$, since $e < e'$, so by lemma 3, $S \subseteq R$. But since $R \in (\Gamma|e')_{\min}$, $S \not\subset R$. So $S = R$. So $(\Gamma|e')_{\min} = \{S\}$. $\dashv$

Lemma 10 (stable arrival) *Suppose that $w \in K_\Gamma$ and $k \geq \mu(w)$. Then*

1. *if $k \geq \mu(w)$, then $(\Gamma|(w|k))_{min} = \{S_w\}$;*
2. *if $S_w \in (\Gamma|(w|i))_{min}$ and $i \leq i'$, then $S_w \in (\Gamma|(w|i'))_{min}$.*

Proof. For (1), let $k \geq \mu(w)$. Then $S_w = S_{w|k}$, so $S_w \in \Gamma|(w|k)$ and for each $R \in \Gamma|(w|k)$, $S_w = S_{w|k} \subseteq R$, so $S_w \in (\Gamma|(w|k))_{\min} = \{S_w\}$. For (2), suppose that $i \leq i'$ and $S_w \in (\Gamma|(w|i))_{\min}$. Note that $S_w \in \Gamma|(w|i')$. Suppose that there exists $R \in \Gamma(w|i')$ such that $R \subset S_w$. Then $R \in \Gamma(w|i)$, so $S_w \notin (\Gamma(w|i))_{\min}$, which is a contradiction. So $S_w \in (\Gamma(w|i'))_{\min}$. $\dashv$