Ockham Efficiency Theorem for Random Empirical Methods

Kevin T. Kelly
Conor Mayo-Wilson

May 21, 2009

Abstract

Ockham's razor is the principle that, all other things being equal, rational scientists ought to prefer simpler theories to more complex ones. In recent years, philosophers have argued the simpler theories make better predictions, possess theoretical virtues like explanatory power, and have other pragmatic virtues like computational tractability. However, such arguments fail to explain why and how a preference for simplicity can help one find *true* theories in scientific inquiry, unless one already assumes that the truth is simple. One new solution to this problem is the Ockham efficiency theorem (Kelly 2002, 2004, 2007a-d, Kelly and Glymour 2004), which states that scientists who heed Ockham's razor retract their opinions less often and sooner than do their non-Ockham competitors. The theorem neglects, however, to consider competitors following random ("mixed") strategies and in many applications random strategies are known to achieve better worst-case bounds than deterministic strategies. In this paper, we describe an extension of the result to a very general class of random, empirical strategies.

1 Introduction

When confronted by a multitude of competing theories, all of which are compatible with existing experimental and observational evidence, scientists prefer theories that minimize free parameters, explanatory causes, independent hypotheses, and theoretical entities and that maximize unity, symmetry, testability, and explanatory power. Today, this systematic bias toward simpler theories—known popularly as "Ockham's razor"—is built into computer-statistical packages that have become everyday tools for working scientists. But why should one prefer simpler theories, and in particular, is there any relationship between simplicity and truth?

Some philosophers have argued that simpler theories are more *virtuous* than complex theories. Simpler theories, they claim, are more explanatory, more

easily falsified or tested, more unified, and more syntactically concise.[1] However, the scientific theory that truly describes the world might, for all we know in advance, involve multiple, fundamental constants or independent postulates; it might be difficult to test and/or falsify, and it might be "dappled" or lacking in underlying unity.[2] In short, since it is an empirical question about whether the truth possesses virtues such as unity and explanatory power, it seems that Ockham's razor cannot be supposed to point at the truth (Van Frassen 1980).

Recently, several philosophers have harnessed mathematical theorems from frequentist statistics and machine learning to argue that simpler theories make more accurate predictions (Forster and Sober 1994) (Vapnik 1998) (Hitchcock and Sober 2004) (Harman and Kulkarni 2007). There are two potential shortcomings with such arguments. First, the proposed motive for Ockham's razor is to maximize predictive accuracy, rather than to find true theories. In fact, simpler theories can improve predictive accuracy in the intended sense *even when it is known that the truth is complex* (Vapnik 1998). Thus, one is led to an anti-realist stance according to which the theories recommended by Ockham's razor should be used as predictive instruments rather than believed as true explanations (Hitchcock and Sober 2004). Second, the assumed notion of predictive accuracy does not extend to predictions of the effects of novel interventions on the system under study: for example, a regression equation may accurately predict cancer rates from the prevalence of ash-trays but might be extremely inaccurate at predicting the impact on cancer rates of a government ban on ash-trays.[3] Scientific realists are unlikely to agree that simplicity considerations have nothing to do with finding true explanations and even the most ardent instrumentalist would be disappointed to learn that Ockham's razor has nothing to do with the vital policy decisions faced by corporate and government policy makers on a daily basis. Hence, the question remains, "How can a systematic preference for simpler theories help one find potentially complex, true *theories*?"[4]

Bayesians and confirmation theorists have argued that simpler theories merit

[1]Nolan (1997), Baker (2003), and Baker (2007) claim that simpler theories are more explanatory. Popper (1959) and Mayo and Spanos (2006) both claim that simpler theories are more severely testable. Friedman (1983) claims unified theories are simpler, and finally, Li and Vitanyi (2001) and Simon (2001) claim that simpler theories are syntactically more concise.

[2]See Schlesinger (1961) for arguments concerning why falsifiability and simplicity are inversely related, and see Cartwright (1999) for a discussion of the apparent dis-unity of science.

[3]More precisely, in regression and density estimation, the predictive accuracy of the model-selection techniques endorsed by Forster, Sober, Harman, and Kulkarni are evaluated only with respect to respect to the distribution *from which the data is sampled.* Thus, for example, one can approximate, to arbitrary precision, the joint density of a set of random variables, and yet make arbitrarily bad predictions concerning the joint density when one or more variables is manipulated. Similarly, in regression, standard model-selection techniques can yield accurate predictions with respect to an underlying curve, but they provide no evidence whatsoever concerning values of the curve's derivative and/or integral. The objection can be overcome by estimating from experimental data, but such data are often too expensive or unethical to obtain precisely when policy predictions are most vital.

[4]We pass over a further concern—that the approach ties Ockham's razor to situations in which the data are sampled randomly, even though Ockham's razor seems just as intuitive for deterministic theories and discrete, non-stochastic data.

stronger belief in light of simple data than do complex theories. Such arguments, however, assume either explicitly or implicitly that simpler possibilities are more probable *a priori*.[5] But that argument is evidently circular—a prior bias toward complex possibilities yields the opposite result. So it remains to explain, without begging the question, why a prior bias toward simplicity is better for finding true theories than is a prior bias toward complexity.

One potential connection between Ockham's razor and truth is that a systematic bias toward true theories allows for convergence to the truth *in the long run* even if the truth is not simple (Sklar 1977, Friedman 1983, Rosenkrantz 1983). In particular, Bayesians argue that prior biases "wash out" in the limit as data accumulate (Savage 1972), resulting in a degree of belief arbitrarily close to 1 as the data accumulate. But prior biases toward complex theories also allow for eventual convergence to the truth (Reichenbach 1938, Hempel 1966, Salmon 1966), for one can dogmatically assert some complex theory until a specified time t_0, and then revise back to a simple theory after t_0 if the anticipated complexities have not yet been vindicated by the data. Hence, mere convergence to the truth does not single out simplicity as the right prior bias to have. So the elusive, intuitive connection between simplicity and theoretical truth is not explained by standard appeals to virtue, predictive accuracy, confirmation, or convergence in the limit.

It is, nonetheless, possible to explain, without circularity, how Ockham's razor finds true theories better than competing methods. It has been be demonstrated (Kelly 2002, 2004, 2007a-e, Kelly and Glymour 2004) that scientists who systematically favor simpler hypotheses converge to the truth in the long run *more efficiently* than scientists with alternative biases, where efficiency is measured in terms of such costs as the total number of errors committed prior to convergence, the total number of retractions performed prior to convergence, and the times at which the retractions occur. The argument is sufficiently general to connect Ockham's razor with the truth in such paradigmatic scientific problems such as curve-fitting, causal inference, and discovering conservation laws in particle physics.

One potential gap in the argument is that it restricts attention to *deterministic* scientific methods. Amongst game theorists, it is well-known that random strategies can achieve lower bounds on worst-case cost than can deterministic strategies, as in the game "rock-paper-scissors". Thus, an important question is: "Do scientists who heed Ockham's razor find true theories with optimal efficiency even when compared with arbitrary, *randomized* scientific strategies?" In this paper, we motivate and prove two new theorems that provide a positive answer to that question. We call the two theorems, *Ockham Efficiency Theorems*. Together, the theorems suggest that scientists who deterministically favor simpler hypotheses fare no worse than those who employ randomizing devices to select theories from data. Further, the notion of randomized strategy we consider is very general, requiring only that a scientific strategy's output depend on

[5]See (Jeffreys 1961) and (Rosenkrantz 1977), respectively, for arguments that explicitly and implicitly assume that simpler theories are more likely to true.

the total input and upon a randomly evolving, discrete, internal state according to an arbitrary, countably-additive probability measure. That includes, as special cases, the class of "mixed strategies" assumed in normal form games, the class of "behavior strategies" assumed in extensive form games and the class of Randomized Turing Machines (RTMs). A larger ambition for this project is to justify Ockham's razor as the optimal means for inferring true statistical theories, such as acyclic causal networks. It is expected that the techniques developed here for dealing with expected costs of convergence will prove to be an essential component of any such theory.

2 Stochastic Empirical Inquiry

Empirical worlds and theories. Scientific theory choice can depend crucially upon small, subtle, or arcane *effects* that can be impossible to detect without sensitive instrumentation, large numbers of observations, or sufficient experimental ingenuity and perseverance. For example, in curve fitting with inexact data[6] (Kelly and Glymour 2004, Kelly 2007a-e, 2008), a quadratic or second-order effect occurs when the data rule out linear laws and a cubic or third-order effect occurs when the data rule out quadratic laws, etc. When explaining particle reactions by means of conservation laws, an effect corresponds to a reaction violating some conservation law. When explaining patterns of correlation with a linear causal network, an effect corresponds to the discovery of new partial correlations that imply a new causal connection in the network (Spirtes et al. 2000, Schulte, Luo, and Greiner 2007, Kelly and Mayo-Wilson 2008). To model such cases, we assume that each potential theory is uniquely determined by the empirical effects it implies and we assume that empirical effects are phenomena that may take arbitrarily long to appear but that, once discovered, never disappear from scientific memory.

Formally, let E be a non-empty, countable (finite or countably infinite) set of *empirical effects*.[7] A *problem* is a set $K \subseteq 2^E$ that corresponds to an empirical constraint on which finite sets of effects one might see for eternity. An *empirical world* in K is an infinite sequence w (of order type ω) of disjoint subsets of E such that $\bigcup_{i\in\omega} w_i \in K$. Let W_K be the set of all empirical worlds. Let $w|n$ denote the finite initial segment $(w_0, \ldots, w_{n-1})$ of w, so that, in particular, $w|0 = ()$. Let F_K denote the set of all finite, initial segments of worlds in W_K. Let $e, e' \in W_K \cup F_K$. Let $l(e)$ denote the length of e and let $e < e'$ hold just in case e is a proper initial segment of e'. let e_- denote the result of deleting the last entry in e, if $e \neq ()$ and let e_- denote $()$ otherwise. Let $*$ denote sequence concatenation. The set of effects presented along e is denoted by S_e and let K_e

[6]It is usually assumed that the data are received according to a Gaussian distribution centered on the true value of Y for a given value of X. Since our framework does not yet handle statistical inference, we idealize by assuming that the successive data fall within ever smaller open intervals around the true value Y.

[7]In this paper, empirical effects are stipulated. It is also possible to define what the empirical effects are in empirical problems in which they are not presupposed (Kelly 2007b, c). The same approach could have been adopted here.

be the restriction of K to finite sets of effects extending S_e. The unique theory corresponding to effect set S is $T_S = \{w \in W : S_w = S\}$. For brevity, let T_w denote T_{S_w}. Let $\mathsf{Th}_K = \{T_S : S \in K\}$ and let $\mathsf{Th}_{K,e} = \{T_S : S \in K_e\}$, which respectively denote the set of all theories and the set of theories compatible with the effect sequence e. If $T \in \mathsf{Th}_K$, let S_T denote the unique $S' \in K$ such that $S_w = S'$, for each $w \in T$. Finally, let $\mathsf{Ans}_K = \mathsf{Th}_K \cup \{\text{`?'}\}$ be the set of answers available to the scientist, where '?' represents a refusal to commit to a given theory.

Discrete state stochastic methods. Let $\{X_e : e \in F_K\}$ be a collection of random variables on some underlying, countably additive probability space $(\Delta, \mathcal{D}, \mu)$. Think of X_e as the "random state" of method $\mathcal{M}$.[8] We assume, further, that the random states of $\mathcal{M}$ are *discrete*, by which we mean that the random variables assume values in a countable measure space $(\Sigma, \mathcal{S})$ such that for each state value $\sigma \in \Sigma$, the singleton event $\{\sigma\} \in \mathcal{S}$. Prior to receiving any inputs, $\mathcal{M}$ is initialized to its *start state* $X_{()} = \sigma_0 \in \Sigma$. As more data arrive, $\mathcal{M}$ enters subsequent random states. We impose no statistical independence assumptions whatever upon the random state evolution. When random state $X_e = \sigma$ is reached, $\mathcal{M}$ employs a uniform rule or procedure $\alpha_e(\sigma)$ for choosing an answer $A \in \mathsf{Ans}_K$. Define $\mathcal{M}_e = \alpha_e(X_e)$, so the probability $p(\mathcal{M}_e = A)$ is defined. When these conditions are met, say that $\mathcal{M}$ is a *discrete state stochastic empirical method* (or *method* for short).

Let $\mathcal{M}$ be a method with components as specified in the preceding section. When $e \in F_K$, the random state trajectory of $\mathcal{M}$ in response to e is the finite, random sequence $X_{[e]} = (X_{e|i} : i \leq l(e))$ and the random output trajectory of $\mathcal{M}$ in response to e is the finite, random sequence $\mathcal{M}_{[e]} = (\mathcal{M}_{e|i} : i \leq l(e))$. If $\mathcal{A} \subseteq \mathsf{Ans}_K^{\leq l(e)}$, the probability $p(\mathcal{M}_{[e]} \in \mathcal{A})$ is defined. Let $e \in F_K$, let $s \in \Sigma^{l(e)}$ and let $D \in \mathcal{D}$ satisfy $p(D) > 0$. Define the *conditional state support* of $\mathcal{M}$ on e *given* D as $\mathsf{Spt}(X_{[e]} \mid D) = \{s \in \Sigma^{l(e)+1} : p(X_{[e]} = s \mid D) > 0\}$.

3 Methodological Properties

A *methodological property* is a relation of form $\Phi(K, \mathcal{M}, e', e, s)$, which can be re-written mnemonically as $\Phi_K(\mathcal{M}, e' \mid X_{[e]} = s)$. It is not assumed that Φ depends upon all of its arguments. Say that methodological property Φ holds *henceforth* of $\mathcal{M}$ in K given $X_{[e]} = s$ if and only if $\Phi_K(\mathcal{M}, e' \mid X_{[e]} = s)$ holds for all $e' \in F_{K,e}$. A stronger notion of Φ "continuing to hold" is *perfection*, which varies the conditioning event along with the time. Say that methodological property Φ holds *perfectly* of $\mathcal{M}$ in K given $X_{[e]} = s$ if and only if $\Phi_K(\mathcal{M}, e' \mid X_{[e']} = s')$ holds, for all $e' \in F_{K,e}$ and $s' \in \mathsf{Spt}(X_{[e']})$. When Φ holds henceforth given $X_{[()]} = (\sigma_0)$, say that Φ holds *always* and when Φ holds perfectly given $X_{[()]} = (\sigma_0)$, say that Φ holds *perfectly*. Say that methodological

[8] More technically, the random states constitute a *stochastic process* indexed by F_K (cf. Lawlor 2006). That statement carries with it the relevant measurability properties.

property $\Phi_K(\mathcal{M}, e' \mid X_{[e]} = s)$ is *stable* just in case for each $e'' \in F_{K,e}$, the following conditions are equivalent:

a. $\Phi_K(\mathcal{M}, e' \mid X_{[e]} = s)$;

b. $\Phi_K(\mathcal{M}, e', \mid X_{[e'']} = s'')$, for each $s'' \in \mathsf{Spt}(X_{[e'']})$.

We now consider some examples of stable methodological properties.

Empirical Simplicity and Ockham's Razor. A *path* in K_e is a finite or infinite sequence of elements of K_e ordered by $\subset$. Let $\mathsf{path}_K(S \mid e)$ denote the set of all finite paths in K_e that terminate in S. Define the *empirical complexity* of S given e as:

$$c_{K,e}(S) = \max\{l(q) : q \in \mathsf{path}_K(S \mid e)\} - 1.$$

Then define $c_{K,e}(w) = c_{K,e}(S_w); c_{K,e}(T_S) = c_{K,e}(S)$; and $C_{K,n}(e) = \{w \in W_e : c_{K,e}(w) = n\}$. The set $C_{K,n}(e)$ is the nth *empirical complexity class* of worlds relative to problem K given e. Let $\mathsf{Ock}_{K,e}$ assume value $\{\text{`?'}, T\}$ if $T$ is the unique $T' \in \mathsf{Th}_{K,e}$ such that $c_{K,e}(T') = 0$, and let $\mathsf{Ock}_{K,e'}$ be $\{\text{`?'}\}$ if there is no unique such T'. Let $s \in \mathsf{Spt}(X_{[e]})$. It follows from the definitions that $C_{K,0}(e) \neq \emptyset$ and that the only way a theory can cease to be Ockham is to be refuted.[9] Say that $\mathcal{M}$ is *Ockham* at e' given $X_{[e]} = s$ if and only if $p(\mathcal{M}_{e'} \in \mathsf{Ock}_{K,e'} \mid X_{[e]} = s) = 1$.

Stalwartness. Ockham's razor proscribes answers other than the uniquely simplest, which leaves the option of returning the uninformative answer'?'. In the deterministic case, stalwartness insists that one never retract an informative Ockham answer until it is no longer Ockham. The statistical generalization of that idea is: if you ever have a chance of producing an informative answer, produce it with unit chance conditional on having just produced it. More precisely, $\mathcal{M}$ is *stalwart* for T at e' given $X_{[e]} = s$ if and only if $l(e') > 0$ and $p(\mathcal{M}_{e'\_} = T \mid X_{[e]} = s) > 0$ and $T \in \mathsf{Ock}_{K,e'}$ imply that $p(\mathcal{M}_{e'} = T \mid \mathcal{M}_{e'\_} = T \& X_{[e]} = s) = 1$. It follows that a stalwart method can statistically "mix" at most one informative answer with '?' and can do even that at most once before leaping all the way to the informative answer. Thus, stalwart, Ockham methods are trivial variants of deterministic methods.

Statistical consistency. In statistical usage, a *consistent* method is a method that converges in probability to the truth. Let $e \in F_K$ and let $s \in \mathsf{Spt}(X_{[e]})$ and let $e' \in F_K$. Say that $\mathcal{M}$ is *consistent* over K given $X_{[e]} = s$ if and only if $\lim_{i \to \infty} p(\mathcal{M}_{w|i} = T_w \mid X_{[e]} = s) = 1, \text{ for each } w \in W_{K,e}$.

Eventual informativeness. Say that $\mathcal{M}$ is *eventually informative* over K given $X_{[e]} = s$ if and only if $\lim_{i \to \infty} p(\mathcal{M}_{w|i} = \text{`?'} \mid X_{[e]} = s) = 0, \quad \text{for all } w \in W_{K,e}$. Eventual informativeness is entailed by consistency and implies that $\mathcal{M}$

[9]The latter property fails if disjunctive theories are entertained.

cannot keep producing '?' infinitely often with non-vanishing probability.

One argument in favor of using an eventually informative, stalwart Ockham method is that such methods converge to the truth (the proof is in appendix 5.3):

Proposition 1 *Suppose that $\mathcal{M}$ is both henceforth stalwartly Ockham and eventually informative given $X_{[e]} = s$. Then $\mathcal{M}$ is consistent over K given $X_{[e]} = s$.*

But that is just one way to converge to the truth. One could just as well guess a complex theory for a thousand stages and revert to a stalwart, Ockham strategy thereafter. So it remains to explain why one should follow Ockham's razor now. We will argue that Ockham's razor is the most efficient possible route to the truth.

4 Efficiency of Empirical Inquiry

Cognitive loss. Efficiency is a matter of avoiding cumulative loss. A *loss function* is a mapping $\lambda : \mathsf{Ans}^{\omega} \times W_K \to \Re$. A *local loss function* is a mapping: $\gamma : \mathsf{Ans}^{<\omega} \times W_K \to \Re$. Let $c \in \mathsf{Ans}^{<\omega}$. Let the local *error* loss function $\epsilon(c, w)$ charge one unit of cost if the last entry in c is a theory false of w and 0 units of cost otherwise.[10] Define the local *retraction* loss function $\mathsf{Ret}(c)$ to charge a unit of loss if the last entry of c differs from the second to last entry *and* the second to last entry is not '?.' Retractions are an unavoidable consequence of inductive, or non-monotonic inference. Everyone prefers deductive (monotonic) inferences when they suffice for finding the truth, but when induction is required one can, at least, insist upon methods that approximate monotonicity as closely as possible—i.e., that minimize retractions en route to the truth. Thus, one needs to compare *cumulative* losses accrued by methods, where cumulative loss is defined to be the sum $\gamma(c, w)[^{\beta}_{m} = \sum_{i=m}^{\beta} \gamma(c|i, w)$, where β may be finite or infinite.

An Ockham efficiency theorem can be obtained for cumulative errors and cumulative retractions alone, but a stronger Ockham efficiency theorem can be obtained if one considers, as well, the lag time to each retraction, the idea being that if a retraction is going to happen, it is best to get it out of the way as soon as possible—both to minimize the number of applications that must be flushed along with the theory and to alleviate the insouciance implied by adherence to views one is destined to reject. If γ is a local loss function, define the *lag time* prior to aggregated cost r to be the least n such that $\gamma(c, w)[^{i}_{0} \geq r$, if there is such an n, and 0 otherwise. Thus, the lag time to the nth retraction is given by $\tau_{\rho \geq n}$. Since the cumulative losses and their incursion times can all be shown to be measurable, it makes sense to speak of expected errors, retrctions, and lag time to incursion of the nth retraction. It is convenient to abbreviate:

[10] That is crude, but the overall argument continues to work as long as the cost of producing a theory in error is invariant over worlds in which the theory is false, as in the epistemic utility theory of I. Levi (1972).

$\gamma^{<n}_{c,w} = \gamma_{c,w}[^{n-1}_{0}$; $\gamma^{\leq n}_{c,w} = \gamma_{c,w}[^{n}_{0}$; $\gamma^{>n}_{c,w} = \gamma_{c,w}[^{\omega}_{n+1}$; $\gamma^{\geq n}_{c,w} = \gamma_{c,w}[^{\omega}_{n}$. In appendix 5.5, it is shown that the losses in question are measurable, so that their (possibly infinite) expectations exist.

States of inquiry. We wish to rank stochastic methods $\mathcal{M}, \mathcal{M}'$ in terms of cumulative loss in light of finite input history e, but it isn't that simple because $\mathcal{M}$ has, by that time, already traversed some state trajectory $s \in \mathsf{Spt}(X_{[e]})$ and $\mathcal{M}'$ has traversed some state trajectory $s' \in \mathsf{Spt}(X'_{[e]})$. The two state spaces could be entirely disjoint. Therefore, we will begin by ranking *states of inquiry* for K at e, by which we mean pairs $(\mathcal{M}, s)$ such that $\mathcal{M}$ is a stochastic empirical method for K and $s \in \mathsf{Spt}(X_{[e]})$. Let $\mathsf{Inq}_{K,e}$ denote the set of all states of inquiry for K at e.

Worst-case cumulative loss. One approach to comparing methods is to compare their worst-case loss bounds. But in problems of unbounded empirical complexity, $W_{K,e}$ is infinite, which would result in equivalence of all methods. On the other hand, some methods can achieve finite retraction bounds in each empirical complexity class, which explains both the reason why we consider retractions as a loss function and why we will consider rankings defined in terms of worst-case loss taken not over all of $W_{K,e}$, but over complexity classes $C_{K,e}(i)$, for $i \in \omega$. For local loss function γ, define: $(\mathcal{M}, s) \leq^{\gamma}_{K,e,n} (\mathcal{M}', s')$ to hold if and only if:

$$\sup_{w \in C_{K,e}(n)} \mathsf{Exp}(\gamma^{\geq 0}_{\mathcal{M},w} \mid X_{[e]} = s) \leq \sup_{w \in C_{K,e}(n)} \mathsf{Exp}(\gamma^{\geq 0}_{\mathcal{M},w} \mid X'_{[e]} = s'),$$

where Exp denotes expected value. It is immediate from the definition that $\leq^{\gamma}_{K,e,n}$ is a pre-order (reflexive and transitive) over $\mathsf{Inq}_{K,e}$.

There is a slight wrinkle in the worst case cost comparison concept when it comes to expected retraction times. The most obvious way to compare worst-case expected retraction times for alternative methods is to compare the expected time of the nth retraction, for each n. But that isn't right, intuitively. Consider the sequence $(T_0, ?, T_0, ?, T_0, ?)$ and the sequence $(T_0, T_0, T_0, T_0, T_0, ?)$. The former seems worse than the latter, but the *first* retraction in the former comes earlier than the first retraction in the latter. It is more natural to ignore the first two retractions in the first sequence and to note that the first sequence is then still as bad as the second in terms of expected retraction times. So in the special case of τ, define $(\mathcal{M}, s) \leq^{\tau}_{K,e,n} (\mathcal{M}', s')$ to hold if and only if for each $w \in C_{K,e}(n)$, there exist $w' \in C_{K,e}(n)$ and a local loss function γ bounded everywhere by ρ such that for each $j \leq \omega$, the inequality $\mathsf{Exp}(\tau^{\rho \geq j}_{\mathcal{M},w} \mid X_{[e]} = s) \leq \mathsf{Exp}(\tau^{\gamma \geq j}_{\mathcal{M}',w'} \mid X'_{[e]} = s')$ holds. Then:

Proposition 2 *$\leq^{\tau}_{K,e,n}$ is a pre-order over $\mathsf{Inq}_{K,e}$.*

Pareto-Rankings. It remains to assemble the various worst-case rankings under consideration into a single ranking. We do so in the least controversial way, by ordering two states of inquiry just in case all the individual rankings

agree. That is known as the *Pareto* ranking. Think of $\gamma \in \{\rho, \epsilon, \tau\}$ as a formal parameter picking out relation $\leq^{\gamma}_{K,e,n}$. Let $\Gamma \subseteq \{\rho, \epsilon, \tau\}$. Then define:

$$(\mathcal{M}, s) \leq^{\Gamma}_{K,e,n} (\mathcal{M}', s') \quad \text{iff} \quad (\mathcal{M}, s) \leq^{\gamma}_{K,e,n} (\mathcal{M}', s'), \text{ for each } \gamma \in \Gamma;$$
$$(\mathcal{M}, s) <^{\Gamma}_{K,e,n} (\mathcal{M}', s') \quad \text{iff} \quad (\mathcal{M}, s) \leq^{\Gamma}_{K,e,n} (\mathcal{M}', s') \text{ and } (\mathcal{M}', s') \not\leq^{\Gamma}_{K,e,n} (\mathcal{M}, s);$$
$$(\mathcal{M}, s) \ll^{\Gamma}_{K,e,n} (\mathcal{M}', s') \quad \text{iff} \quad (\mathcal{M}, s) <^{\gamma}_{K,e,n} (\mathcal{M}', s'), \text{ for each } \gamma \in \Gamma.$$

These rankings are complexitiy-relative. To avoid grounding our results on subjective weights over complexity classes, we again restrict attention to Pareto comparisons that agree in every complexity dimension. First define the global upper complexity bound for sub-problem K_e as: $c_{K,e} = \sup\{i + 1 : i \in \omega \text{ and } C_{K,e}(\beta) \neq \emptyset\}$.[11] Then define:

$$(\mathcal{M}, s) \leq^{\Gamma}_{K,e} (\mathcal{M}', s') \quad \text{iff} \quad (\mathcal{M}, s) \leq^{\Gamma}_{K,e,n} (\mathcal{M}', s'), \text{ for each } n \in \omega;$$
$$(\mathcal{M}, s) <^{\Gamma}_{K,e} (\mathcal{M}', s') \quad \text{iff} \quad (\mathcal{M}, s) \leq^{\Gamma}_{K,e} (\mathcal{M}', s') \text{ and } (\mathcal{M}', s') \not\leq^{\Gamma}_{K,e} (\mathcal{M}, s);$$
$$(\mathcal{M}, s) \ll^{\Gamma}_{K,e} (\mathcal{M}', s') \quad \text{iff} \quad (\mathcal{M}, s) \ll^{\Gamma}_{K,e,n} (\mathcal{M}', s'), \text{ for each } n < c_{K,e}.$$

The relation $<^{\Gamma}_{K,e}$ is *weak* Pareto-dominance and $\ll^{\Gamma}_{K,e}$ is *strong* Pareto-dominance.

Switching Methods in Midstream. Let $\mathcal{M}$ be a stochastic method for K. Suppose that one has been using $\mathcal{M}$ and the current, finite input sequence is e. Given that $X_{[e]} = s$, where $e > ()$, the past outputs of $\mathcal{M}$ along e_- cannot be changed, so one is stuck with the output sequence $c = \mathcal{M}_{[e_-]}(s)$ and with the cumulative loss $\gamma(c)|_0^{l(c)-1}$. Now consider alternative stochastic method $\mathcal{M}'$ with state variables $\{X'_e : e \in F_K\}$. Given that $X'_{[e]} = s'$, one has the option to switch methods from $\mathcal{M}$ to $\mathcal{M}'$ with state history s' from e onward. But one is still stuck with the costs from having used $\mathcal{M}$. So, when switching from $\mathcal{M}$ to $\mathcal{M}'$ at e, one must consider not the overall resource consumption of $\mathcal{M}'$ given $X_{[e']} = s'$, but the cost of $\mathcal{M}'$ given $X_{[e']} = s'$ from $l(e)$ onward, added to the resource consumption of $\mathcal{M}$ given $X_{[e]} = s$ along e_-. It is convenient to conceive of the switch from $\mathcal{M}$ to $\mathcal{M}'$ at e as having *always* followed hybrid method $\mathcal{M} \star^s_e \mathcal{M}'$, which acts like $\mathcal{M}$ given $X_{[e]} = s$ along e_- and like $\mathcal{M}'$ thereafter. That is readily accomplished simply by modifying the output function α' of $\mathcal{M}'$. Define the hybrid output function:

$$(\alpha \star^s_e \alpha')_{e'}(\sigma) = \begin{cases} \alpha_{e'}(s(l(e'))) & \text{if } e' < e; \\ \alpha'_{e'}(\sigma) & \text{otherwise.} \end{cases}$$

Then define the hybrid method: $\mathcal{M} \star^s_e \mathcal{M}'$ to be the result of replacing α' with $(\alpha \star^s_e \alpha')$ in $\mathcal{M}'$.

Efficiency. Say that $\mathcal{M}$ *is* Γ-*efficient* for K given $X_{[e]} = s$ if and only if for each $(\mathcal{M}', s') \in \mathsf{Inq}_{K,e}$, if $\mathcal{M}'$ is consistent given $X'_{[e]} = s'$ then $(\mathcal{M}, s) \leq^{\Gamma}_e$

[11] Adding 1 makes the bound strict both in the case of finitely bounded and finitely unbounded orders of complexity.

$(\mathcal{M} \star_e^s \mathcal{M}',\ s')$. Next, say that $\mathcal{M}$ *is weakly* Γ*-dominated* given $X_{[e]} = s$ if and only if there exists $(\mathcal{M}', s') \in \mathsf{Inq}_{K,e}$ such that $\mathcal{M}'$ is consistent from e given $X'_{[e]} = s'$ and $(\mathcal{M} \star_e^s \mathcal{M}',\ s') <_e^\Gamma (\mathcal{M}, s)$. Finally, say that $\mathcal{M}$ *is strongly* Γ*-dominated* given $X_{[e]} = s$ if and only if there exists $(\mathcal{M}', s') \in \mathsf{Inq}_{K,e}$ such that $\mathcal{M}'$ is consistent from e given $X'_{[e]} = s'$ and $(\mathcal{M} \star \mathcal{M}',\ s') \ll_e^\Gamma (\mathcal{M}, s)$. Note that these concepts are relative to e and that such a property holds *perfectly* just in case it holds at every $e \in F_K$. Thus, one may speak of perfect Γ-efficiency, perfect non-Γ dominance and perfect non-Γ-strict-dominance.

4.1 Ockham Efficiency Theorems

It is now possible to state the main results, whose proofs are available in (Kelly and Mayo-Wilson 2009).

Theorem 1 (Ockham Efficiency Characterization) *Assume that* $\mathcal{M}$ *is consistent and that:*

$$\begin{array}{rcccl} \{\epsilon, \rho\} & \subseteq & \Gamma & \subseteq & \{\rho, \epsilon, \tau\} \text{ or} \\ \{\tau\} & \subseteq & \Gamma & \subseteq & \{\rho, \epsilon, \tau\}. \end{array}$$

Then following are equivalent:

1. $\mathcal{M}$ *is always Ockham and stalwart;*
2. $\mathcal{M}$ *is perfectly* Γ*-efficient;*
3. $\mathcal{M}$ *is perfectly weakly* Γ*-undominated.*

Regarding the main question posed in the introduction, it follows from the theorem that random methods cannot do better than deterministic Ockham methods and that most randomized strategies do worse. Recall that every stalwart strategy must produce Ockham answer T with chance 1 immediately after producing T with any non-zero probability as long as T remains Ockham. Thus, the only "mixtures" of answers such a method can produce involve a single informative theory T and the uninformative answer '?' and after producing such a mixture once, the method must produce T with probability 1 thereafter, until T is refuted.

It is immediate from the definitions that weak Γ-dominance at some e implies Γ-inefficiency at that e, but the converse, implied by the preceding theorem, is not at all trivial: it holds only because of the asymptotic character of the costs considered and because of nature's ability to force one arbitrarily late retraction for each degree of empirical complexity from an arbitrary, consistent method. Thus, the consistent methods are neatly partitioned into the efficient, stalwart, Ockham ones and the weakly dominated ones, with no awkward cases remaining to be resolved by subjective weights on costs or complexity classes. In that important sense, Ockham's razor is a matter of logical structure rather than of practical taste.

The simple idea behind the proof is that nature has a strategy to *force* an expected retraction for each step along a path in K_e, which is the worst that

a stalwart, Ockham strategy would do, since Ockham strategies always allow Nature to lead the way through the paths in K_e. Hanging onto a refuted theory does not add extra retractions, but does add extra errors and elapsed time to the first retraction in complexity class 0. Violating stalwartness adds an extra retraction in each complexity class.

Theorem 1 does *not* imply that a strategy is dominated each time it violates Ockham's razor. For example, suppose that $K_e = \{\{a\}, \{a, b\}, \{c\}\}$ and that $\mathcal{M}$ produces $T_{\{a\}}$ with probability 1 both at e_- and e, which are Ockham violations because $T_{\{c\}}$ is equally simple. Then it is a bad idea to retract to '?' at e, as Ockham demands, since a competitor who sticks with $T_{\{c\}}$ at e would retract at most once in complexity class $C_{K,1}(e)$ and zero times in $C_{K,2}(e)$, whereas Ockham would retract at least once in both non-empty complexity classes. Note that the path $(\{a\}, \{a, b\})$ in this example is longer than the path $(\{c\})$. A much stronger Ockham efficiency theorem holds under the special hypothesis that K has *no short paths*, which means that for each $e \in K$ and for each $S \in K_e$ there is a path p of maximum length over all paths in K_e such that p begins with S. For example, K has no short paths if each maximal path in K is infinite, as in standard examples like curve fitting. When there are no short paths, each method that fails to be always Ockham and stalwart does *strictly* worse in *every* complexity class at *each* violation. That is to say, Theorem 2 proves that there is a always a good reason never to deviate from the behavior of an Ockham, stalwart strategy.

Theorem 2 (Strong, Stable Ockham Efficiency Characterization) *Suppose that K has no short paths. Let $e \in F_K$, $s \in \mathsf{Spt}(X_{[e]})$ and let $\mathcal{M}$ be consistent given $X_{[e]} = s$. Assume that $\{\tau\} \subseteq \Gamma \subseteq \{\rho, \epsilon, \tau\}$. Then following are equivalent:*

1. *$\mathcal{M}$ is henceforth Ockham and stalwart given $X_{[e]} = s$;*
2. *$\mathcal{M}$ is perfectly Γ-efficient given $X_{[e]} = s$;*
3. *$\mathcal{M}$ is perfectly strongly Γ-undominated given $X_{[e]} = s$.*

5 Discussion and Future Work

The Ockham efficiency theorems establish that scientists who systematically favor simpler theories minimize errors and the number and timeliness of retractions of opinion. Intricate randomized strategies, moreover, are of little help: the proofs show that scientists who *deterministically* select the simplest theory compatible with the data still minimize costs of empirical inquiry. Therefore, there is a deep sense in which simplicity marks the "shortest path" to the true scientific theory governing a particular phenomenon.

Because our concept of a "stochastic empirical method" is closely related to that of a "mixed strategy," it is, perhaps, of some interest to frame the results within a more game-theoretic perspective. In fact, we conjecture that Ockham

methods form the scientist's half of every Nash equilibrium in a strictly competitive game with Nature, where Nash equilibrium is understood with respect to a generalized sense of preference to be clarified below.

To analyze our learning model as a two-person game, one must decide what sort of "strategies" ought to be available to a second player, Nature. It is natural to view Nature not as a real player but as a personification of the scientist's prior probabilities over the set W_K of possible worlds. Thus, each pure strategy for Nature is a world and a mixed strategy for Nature is a possible prior probability of the scientist. Preliminary results suggest that there are no equilibria in which science fails to play an Ockham strategy and, furthermore, that there exist Ockham equilibria in which Nature's mixture favors *complex* theories.[12] Thus, the game-theoretic argument may serve as the basis for a *non-circular* Bayesian vindication of Ockham's razor.

To prove the preceding conjecture, one must surmount at least four obstacles. The preceding argument essentially addresses the first three obstacles, but the fourth remains open for future research. First, game theorists typically assume that players' preferences over outcomes of the game are *totally* ordered; in our game, outcomes are often incomparable with one another (for both players), as when one method retracts fewer times but later than another method. In the absence of totally-ordered preferences, standard game-theoretic results, such as Nash's theorem, fail to guarantee the existence of equilibria. It is possible to show (Mayo-Wilson 2009) that Nash equilibria often exist in games in which player's preferences are merely pre-orders (i.e. transitive and reflexive), so we expect that the Pareto-orderings on cost discussed in this paper are not an impediment to representing our model as a game.

Second, in extensive-form games, game theorists standardly assume that a player's disposition to perform a particular action at a move in the game is probabilistically independent of the actions taken by players in the past. In many cases, this amounts to assuming that players ignore important information from the past, which is especially unrealistic when modeling empirical inquiry as a game. Here, our results make no independence assumptions whatsoever, and no complications arise when Nature is introduced as a second player.

Third, game theorists routinely employ *discount* functions or other asymptotic artifices to ensure that cumulative costs in (repeated or extensive-form) games are bounded. Such discount factors are ad hoc even in practical applications and it would be all the more odd if the validity of Ockham's razor were to depend upon the selection of such a factor. In contrast, we follow practice in the theory of computational complexity (Garey and Johnson 1979), in which cumulative computational costs such as the total number of steps of computation are also unbounded over all possible world states (inputs to a given algorithm).

[12]We also conjecture that there are no equilibria in which Nature's mixture is countably additive. The idea is that if Nature's mixture were countably additive, the scientist could reduce expected retractions by producing '?' for a longer expected time and then Nature would regret not having withheld effects for even longer expected time. In a finitely additive mixed strategy, Nature can present effects "infinitely late". For a discussion of the connection between finite additivity and skeptical arguments, cf. (Kelly 1996, chapter 13).

The idea in computational complexity theory is to partition possible states (inputs) according to *length*, so that the worst-case computational time over each length class exists and is finite. In the case of inquiry, inputs never cease, so we plausibly substitute empirical complexity for length. Then we seek methods that are *admissible* not with respect to all states of the world, but with respect to worst case bounds over all empirical complexity classes. The resulting comparisons are richer than admissibility with respect to world states and are also less objectionably pessimistic than worst-case (maximin) comparisons.

The fourth and major hurdle in representing our theorems as game-theoretic equilibria is the development of a more general theory of simplicity. The definition of simplicity stated in this paper is very narrow, allowing only for prior knowledge about which finite sets of effects might occur—knowledge about timing and order of effects is not allowed for. But nothing prevents nature from choosing a mixed strategy that implies knowledge about timing or order of effects (recall that nature's mixture is to be understood as the scientist's prior probability. Such knowledge may essentially alter the structure of the problem: e.g., if nature chooses a mixing distribution according to which effect a is always followed immediately by effect b, then the sequence a, b ought properly to be viewed as a single effect rather than as two separate effects.[13] But if simplicity is altered by nature's choice of a mixing distribution, then so is Ockham's razor and, hence, what counts as an Ockham strategy for the scientist. Therefore, in order to say what it means for Ockham's razor to be a "best response" to Nature, it is necessary to define simplicity with sufficient generality to apply to every possible restriction of an effect accounting problem W_K to a narrower domain $W' \subseteq W_K$ of worlds. More general theories of simplicity than the one presented in this paper have been proposed and have been shown to support Ockham efficiency theorems (Kelly 2007d, 2008), but those concepts are still not general enough to cover *all* possible restrictions of W_K. Of course, a general Ockham efficiency theorem based on a general concept of simplicity would be of considerable interest quite independently of this exploratory discussion of game theory. Current work on that important problem is promising but, as yet, inconclusive.

6 Bibliography

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

Baker, A. (2003) "Quantitative Parsimony and Explanatory Power", *British Journal for the Philosophy of Science* 54: 245-259.

[13]The difficulties are exacerbated when scientist's prior probability (i.e. Nature's mixed strategy) is only finitely additive, as there is no obvious concept of "support" in that case, even over countable sets of worlds.

Baker, A. (2007) Occam's Razor in Science: A Case Study from Biogeography. *Biology and Philosophy.* 22: 193215.

Cartwirght, N. (1999) *The Dappled World: A Study of the Boundaries of Science*, Cambridge: Cambridge University Press.

Garey, M. and Johnson, D. (1979) *Computers and Intractability: A Guide to the Theory of NP-Completeness*, New York: W. H. Freman.

Forster, M. (2001) The New Science of Simplicity. (In A. Zellner, H. Keuzenkamp, and M. McAleer (Eds.) *Simplicity, Inference and Modelling.* (pp. 83-119). Cambridge: Cambridge University Press)

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

Friedman, M. (1983) *Foundations of Spacetime Theories: Relativistic Physics and Philosophy of Science.* Princeton University Press.

Harman, G. and Kulkarni, S. (2007) *Reliable Reasoning: Induction and Statistical Learning Theory* (Cambridge: MIT Press).

Hempel, C. (1966) *Philosophy of Natural Science* (Englewood-Cliffs: Prentice Hall).

Hitchcock, C. and Sober, E. (2004) "Prediction Versus Accommodation and the Risk of Overfitting." *British Journal for the Philosophy of Science.* 55: pp.

Jeffreys H. (1961) *Theory of Probability.* Oxford: Clarendon Press.

Kelly, K. (1996) *The Logic of Reliable Inquiry.* Oxford University Press.

Kelly, K. (2002)"Efficient Convergence Implies Ockham's Razor," *Proceedings of the 2002 International Workshop onvComputational 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. (2007a)"A New Solution to the Puzzle of Simplicity", *Philosophy of Science* 74: 561-573.

Kelly, K. (2007b)"How Simplicity Helps You Find the Truth Without Pointing at it", V. Harazinov, M. Friend, and N. Goethe, eds. *Philosophy of Mathematics and Induction*, Dordrecht: Springer, pp. 321-360.

Kelly, K. (2007c) "Ockham's Razor, Empirical Complexity, and Truth-finding Efficiency," *Theoretical Computer Science*, 383: 270-289.

Kelly, K. (2007d) "Simplicity, Truth, and the Unending Game of Science", *Infinite Games: Foundations of the Formal Sciences V*, S. Bold, B. Löwe, T. Rsch, J. van Benthem eds, Roskilde: College Press 2007 pp. 223-270.

Kelly, K. (2008) "Ockhams Razor, Truth, and Information", in *Philosophy of Information*, Van Benthem, J. Adriaans, P. eds. Dordrecht: *Elsevier*, 2008 pp. 321-360.

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

Kelly, K. and Lin, H. (2009) "A (Topo)Logical Theory of Empirical Simplicity." Unpublished Draft.

Kelly, K. and Mayo-Wilson, C. (2007) "Causal Discovery, Retractions, and Their Minimization." Unpublished Draft.

Kelly, K. and Mayo-Wilson, C. (2009) "Ockham Efficiency Theorem for Random Empirical Methods." Unpublished Draft. Available electronically at ⟨ http://www.andrew.cmu.edu/user/kk3n/ockham/Mixed Strategies upload.pdf⟩

Lawlor, G. (2006) *Introduction to Stochastic Processes*, 2nd ed., New York: Chapman and Hall.

Levi, I. (1976) *Gambling with Truth*, Cambridge: M.I.T. Press.

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

Li, M. and Vitanyi, P. (2001) "Simplicity, Information, and Kolmogorov Complexity", in A. Zellner, H. Keuzenkamp, and M. McAleer eds., *Simplicity, Inference and Modelling*, Cambridge: Cambridge University Press, pp. 83-119.

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

Mayo, D. and Spanos, A. (2006) "Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction", *British Journal for the Philsophy of Science*, 57: 323-357.

Mayo-Wilson, C. (2009) "Ockham's Shaky Razor: Efficient Convergence by Random Methods." In Preparation. Master's Thesis, Department of Philosophy, Carnegie Mellon University..

Nolan, D. (1997) "Quantitative Parsimony", *British Journal for the Philosophy of Science,* 48: 329-343.

Popper, K. (1959) The Logic of Scientific Discovery. London: Hutchinson.

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

Reichenbach, H. (1938) *Experience and Prediction.* University of Chicago Press.

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

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

Rosenkrantz, R. (1977) *Inference, Method, And Decision: Towards A Bayesian Philosophy Of Science*, Boston: Reidel.

Salmon, W. (1966) *The Foundations of Scientific Inference*, Pittsburgh: University of Pittsburgh Press.

Savage, L. (1972) *The Foundations of Statistics*, New York: Dover.

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 Principles in Particle Physics: A Case Study in the Problem of Induction. *The British Journal for the Philosophy of Science,* 51, 771-806.

Schulte, O. (2008) The Co-Discovery of Conservation Laws and Particle Families. *Studies in History and Philosophy of Modern Physics* 39: 288-314.

Schulte, O., Luo, W., and Griner, R. (2007) Mind Change Optimal Learning of Bayes Net Structure. *20th Annual Conference on Learning Theory* (COLT), San Diego.

Schlesinger, G. (1961) "Dynamic Simplicity." *The Philosophical Review,* 70: 485-499.

Sklar, L. (1977) *Space, Time, and Spacetime.* Berkeley: University of California Press.

Simon, H. (2001) Science Seeks Parsimony, not Simplicity: Searching for Pattern in Phenomena. (In A. Zellner, H. Keuzenkamp, and M. McAleer eds., *Simplicity, Inference and Modelling.* pp. 83-119, Cambridge: Cambridge University Press.

Spirtes, P., Glymour, C. and Scheines, R. (2001) *Causation, Prediction, and Search*, 2nd. ed., Cambridge: M.I.T. Press.

Vapnik, V. (1998) *Statistical Learning Theory*, New York: Wiley.

van Fraassen, B. (1980) *The Scientific Image.* Oxford University Press.

Wrinch, D. and Jeffreys, H. (1923) On Certain Fundamental Principles of Scientific Inquiry. *Philosophical Magazine,* 45: 368-374