Simplicity, Truth, and Probability

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

October 17, 2009

[*] This material is based in part upon work supported by the National Science Foundation under Grant Number 0750681. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

Abstract

Simplicity has long been recognized as an apparent mark of truth in science, but it is difficult to explain why simplicity should be accorded such weight. This chapter examines some standard, statistical explanations of the role of simplicity in scientific method and argues that none of them explains, without circularity, how a reliance on simplicity could be conducive to finding true models or theories. The discussion then turns to a less familiar approach that does explain, in a sense, the elusive connection between simplicity and truth. The idea is that simplicity does not point at or reliably indicate the truth but, rather, keeps inquiry on the cognitively most direct path to the truth.

1 Introduction

Scientific theories command belief or, at least, confidence in their ability to predict what will happen in remote or novel circumstances. The justification of that trust must derive, somehow, from scientific method. And it is clear, both from the history of science and from the increasing codification and automation of that method both in statistics and in machine learning, that a major component of that method is *Ockham's razor*, a systematic bias toward *simple* theories, where "simplicity" has something to do with minimizing free parameters, gratuitous entities and causes, independent principles and ad hoc explanations and with maximizing unity, testability, and explanatory power.

Ockham's razor is not a bloodless, formal game that must be learned—it has a native, visceral grip on our credence. For a celebrated example, Copernicus was driven to move the earth to eliminate five epicycles from medieval astronomy (Kuhn 1957). The principal problem of positional planetary astronomy was to account for the apparently irregular, retrograde or backward motion of the planets against the fixed stars. According to the standard, Ptolemaic theory of the time, retrograde motion results from the planet revolving around an *epicycle* or circle whose center revolves, in turn, on another circle called the *deferent*, centered on the earth. Making the epicycle revolve in the same sense as the deferent implied that the planet should be closest or brightest at the midpoint of its retrograde motion, which agreed with observations. Copernicus explained retrograde motion in terms of the moving earth being lapped or lapping the other planets on a cosmic racetrack centered on the sun, which eliminates one epicycle per planet (figure 1). Copernicus still required many superimposed circles to approximate elliptical orbits, so the mere elimination of five such circles may not seem very impressive. But there is more to the story than just counting circles. It happens that the motions of Mars, Jupiter, and Saturn occur precisely when the respective

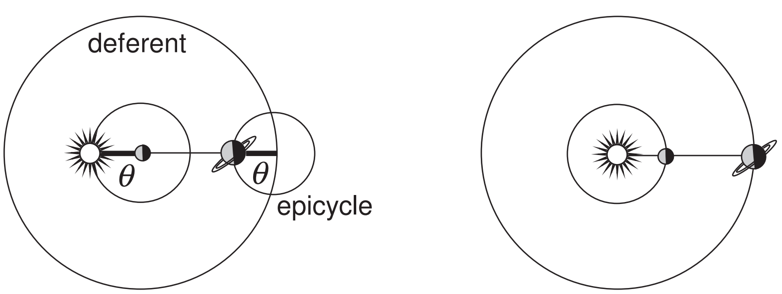

Figure 1: Ptolemy vs. Copernicus

planet is in solar opposition (i.e., is observed 180° from the sun) and that the retrograde motions of Mercury and Venus occur at solar conjuction (i.e., when the respective planet is 0° from the sun). Ptolemy's epicycles can be adjusted to recover the same effect, but only in a rather bizarre manner. Think of the line from the earth to the sun as the hand of a clock and think of the line from the center of Saturn's epicycle to Saturn as the hand of another clock. Then retrograde motion happens exactly at solar opposition if and only if Saturn's epicycle clock is *perfectly synchronized* with the sun's deferent clock. The same is true of Mars and Jupiter. Furthermore, Mercury and Venus undergo retrograde motion exactly at solar opposition just in case their deferent clocks are perfectly synchronized with the sun's deferent clock. In Ptolemy's theory, these perfect synchronies across vast distances in the solar system appear bizarre and miraculous. On Copernicus' theory, however, they are unavoidable banalities: a racer is lapped by a competitor on a circular racetrack only when the contestant is in opposition or conjunction with the center of the racetrack, depending on whether the competitor is in an inner or an outer lane. So Copernicus' theory crisply *explains* the striking synchronies. Copernicus' theory is also *severely tested* by the synchronies, since it would be refuted by any perceived deviation from exact synchrony, however slight. Ptolemy's theory, on the other hand, merely *accommodates* the data in an *ad hoc* manner by means of its adjustable parameters. It seems that Copernicus' theory should get some sort of reward for surviving a test shirked by its competitor. One could add clockwork gears to Ptolemy's theory to explain the synchronies, but that would be an *extra principle* receiving no *independent confirmation* from other evidence. Copernicus' explanation, on the other hand, recovers both retrograde motion and its correlation with solar position from the geometry of a circular racetrack, so it provides a *unified explanation* the two phenomena. Empirical simplicity is more than mere, notational brevity—it implies such red-blooded considerations as explanatory power (Harman 1965), unity (Kitcher 1982), independently confirmable principles (Friedman 1983, Glymour 1980) and severe testability (Popper 1968, Mayo 1996).

Another standard example of Ockham's razor in action concerns the search for empirical laws (figure 2). Any finite number of observations can be connected with a polynomial curve that passes through each, but we still prefer the straight line that comes close to each point. It is, perhaps, more tempting in this case to identify simplicity with syntactic length or complexity of the law, since $\alpha_0 x^0 + \alpha_1 x^1 + \ldots \alpha_n x^n$ is obviously more verbose than $\alpha_0 x^0 + \alpha_1 x^1$. But one can also say that the complex law merely accommodates the data by having an independent, adjustable parameter for each data point, whereas when the two parameters of the simple law can be estimated with a few data, providing an explanation of the remaining data points. The complex law is also less unified than the simple law (the several coefficients receive isolated support from the data points they are set to account for) and is less visually "uniform" than the simple law.

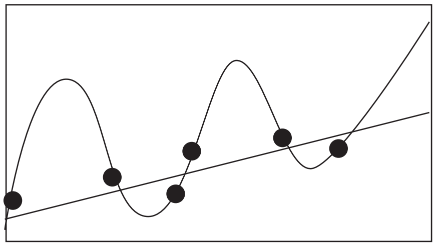

Figure 2: inferring polynomial degree

Ockham's razor does the heavy lifting in scientific theory choice, for no other principle suffices to winnow the infinite range of possible explanations of the available data down to a unique one. And whereas simplicity was once the theorist's personal prerogative, it is now a mathematically explicit and essential component of contemporary statistical and computational techniques for drawing conclusions from empirical data (cf. Mitchell 1977, Duda et al. 2001). The explicitness and indispensability of simplicity considerations in scientific theory selection raises a natural question about their justification. Epistemic justification is not just a word or a psychological urge or a socially sanctioned, exculpatory ritual or procedure. It should imply some sort of *truth-conduciveness* of the underlying process by which one's trust is produced. An attractively ambitious concept of truth-conduciveness is *reliable indication* of the truth, which means that the process has a high chance of producing the true theory, whatever the truth happens to be, the way a properly functioning thermometer indicates temperature. But Ockham's razor is more like a trick thermometer whose reading never changes. Such a thermometer cannot be said to indicate the temperature even if its fixed reading happens to be true and neither can a fixed bias toward simplicity immediately indicate the truth about nature even if the truth happens to be simple.[1]

[1] This point has been underscored in machine learning by the "no free lunch theorems"

Ockham's razor has a good excuse for failing to reliably indicate true theories, since theory choice requires inductive inference and no inductive inference rule can be a truth-indicator: each finite set of data points drawn with bounded precision from a linear law is also compatible with a sufficiently flat parabola, so no possible data-driven process could reliably indicate, in the short run, whether the truth is linear or quadratic. A more feasible concept of truth-conduciveness for inductive inference is *convergence in the limit*, which means that the chance that the method produces the true theory converges to one, no matter what the true theory might be.[2] Convergence to the truth in the limit is far weaker than short-run truth-indication, since it is compatible with the choice of any finite number of false theories with arbitrarily high chance before settling on the correct one. Each time a new theory T_{n+1} is produced with high chance, the chance of producing the previous candidate T_n must drop precipitously and one may say that the output is *retracted*. So convergence in the limit differs from reliable indication by allowing any finite number of arbitrarily precipitous retractions prior to "locking on" to the right answer. Assuming that the true theory is polynomial, Ockham's razor does converge in the limit to the true polynomial degree of $f(x)$—each polynomial degree lower than the true degree is ruled out, eventually, by the data (e.g., when new bumps in the true law become noticable), after which the true theory is the simplest theory compatible with experience. Think of successively more complex theories as tin cans lined up on a fence, one of which (the true one) is nailed to the fence. Then, if one shoots the cans from left to right, eventually the nailed can is the first can in line that has not yet been shot down. The familiar trouble with this explanation of Ockham's razor is that convergence in the long run is compatible reliance on any alternative bias for any finite duration (Salmon 1967). For example, guess an equation of degree 10 with the hope that the coefficient is so large that the thousand bumps will be noticed early—say in a sample of size 1000. If they aren't seen by then, revert back to Ockham's razor. Hence, convergence in the limit is feasible in theoretical inference, but it does not single out simple theories as the right theories to produce in the short run.

To summarize, the justification of Ockham's razor poses a puzzle. Ockham's razor can't reliably indicate the true theory in the short run, due to the problem of induction. And although Ockham's razor does converge to the truth in the ideal limit of inquiry, alternative methods producing very complex theories are also truth-conducive in that very weak sense as well (Salmon 1967). So short-run indication is too strong to be feasible and long-run indication is too weak

(Wolpert 1996).

[2] This concept is called *convergence in probability* in probability theory and *consistency* in statistics.

to single out Ockham's razor. It remains, therefore, to define a sense of truth-conduciveness according to which it can be argued, without circularity, that Ockham's razor helps one find the truth better than alternative methods that would produce arbitrarily complex theories *now*. Absent such a story, Ockham's razor starts to look like an exercise in wishful thinking—the epistemic sin of inferring that reality is simple because the true theory of a simple world would have pragmatic virtues (e.g., explanatory power) that we would prefer it to have. Such doubts motivate a skeptical or anti-realist attitude toward scientific theories in general (van Fraassen 1981).

This paper reviews the standard explanations of Ockham's razor, which fall into two main groups. The first group invokes a tacit, prior bias toward simplicity, which begs the question in favor of Ockham's razor. The second group substitutes a particular notion of predictive accuracy for truth, based on the surprising fact that a false theory may make more accurate predictions than the true one when the truth is complex. That evidently fails to explain how Ockham's razor finds true *theories*. Furthermore, when predictions concern the outcomes of interventions on the world, even the argument for predictive accuracy fails.[3] Since neither approach really explains how Ockham's razor leads to true theories or even to accurate policy predictions, the second part of the paper develops an entirely new explanation: Ockham's razor does not point at the truth, even with high probability, but it does help one arrive at the truth with uniquely optimal *efficiency*, where efficiency is measured in terms of such epistemically pertinent considerations as the total number of errors and retractions of prior opinions incurred before converging to the truth and the elapsed times by which the retractions occur. Thus, in a definite sense, Ockham's razor is demonstrably the uniquely most truth-conducive method for inferring general theories from particular facts—even though no possible method can be guaranteed to point toward the truth with high probability in the short run.

2 The Argument from Bayes Factors

Bayesian statisticians assign probability-valued degrees of belief to all the propositions in some language and then "rationally" update those degrees of belief by a universal rule called *conditionalization*.[4] If $p_t(T)$ is your prior degree of belief

[3] For candid discussions of the shortcomings of the usual explanations of Ockham's razor as it is used in machine learning, cf., for example, (Domingos 1999) and (Mitchell 1997).

[4] Not all Bayesians accept updating by conditionalization. Some Bayesians recommend accepting hypotheses altogether, in which case the degree of belief goes to one. Others recommend updating on partially believed evidence. Others recommend updating interval-valued degrees of belief, etc. Others reject its coherentist justification in terms of diachronic Dutch books.

that T at stage t and if E is new evidence received at stage $t+1$, then conditionalization says that your updated degree of belief that T at $n+1$ should be:

$$p_{t+1}(T) = p_t(T \mid E).$$

It follows from the conditionalization rule that:

$$p_{t+1}(T) = (p_t(T) \cdot p_t(E \mid T))/p_t(E).$$

An important feature of the rule is that one's updated degree of belief p_{t+1} depends on one's prior degree of belief $p_t(T)$, which might have been strongly biased for or against T prior to collecting any evidence about T whatever. That feature suggests an easy "justification" of Ockham's razor—just start out with prior probabilities biased toward simple theories. Then, if simple theories explain the data about as well as complex ones, the prior bias toward the simple theory will survive the updating procedure, implementing Ockham's razor. But to invoke a prior bias toward simplicity to explain a prior bias toward simplicity evidently begs the main question at hand.

A more promising Bayesian argument for Ockham's razor centers not on the prior probability $p_t(T)$, but on the term $p_t(E \mid T)$, which corresponds to the rational credence conferred to E by theory T. (cf. Jeffreys 1961, Rosenkrantz 1983, Myrvold 2003). According to this explanation, Ockham's razor does not demand that the simpler theory T_1 start out ahead of its complex competitor T_2; it suffices that T_1 pull ahead of T_2 when evidence E compatible with T_1 is received. That sounds impressive, for the conditional probability $p_t(E \mid T)$ is often thought to be more objective than the prior probability $p(T)$, because $p_t(E \mid T)$ reflects the degree to which T "explains" E. But that crucially overstates the case when T has free parameters to adjust, as when Ockham's razor is at issue. Thoroughly subjective Bayesians interpret "objective" probabilities as nothing more than relatively inter-subjective degrees of belief, but a more common view ties objectivity to *chances*. Chances are supposed to be natural, objective probabilities that apply to possible outcomes of random experiments. Chance will be denoted by a capital P, in contrast with the lower-case p denoting degrees of belief. Bayesian statisticians link chances to evidence and to action by means of the *direct inference principle* (Kyburg 1977, Levi 1977), which states that degrees of belief should accord with known chances, given only *admissible*[5] information E':

$$p_t(E \mid P(E) = r \ \wedge \ E') = r.$$

If theory T says exactly that the true chance distribution of X is P, then by the direct inference principle:

$$p_t(E \mid T) = P(E),$$

[5] Defining admissibility is a vexed question that will be ignored here.

which is, indeed, objective. But if T is complex, then T has adjustable parameters and, hence, implies only that the true chance distribution lies in some set, say: $\{P_1, \ldots, P_t\}$. Then the principle of direct inference yields the weighted average:

$$p_t(E \mid T) = \sum_{i=1}^{n} P_t(E) \cdot p_t(P_n \mid T),$$

in which the weights $p_t(P_n \mid T)$ are prior degrees of belief, not chances. So the objective-looking quantity $p_t(E \mid T)$ is *loaded* with prior opinion when T is complex and that potentially hidden fact is crucial to the Bayesian explanation of Ockham's razor.

A standard technique for comparing the posterior probabilities of theories is to look at the *posterior ratio*:

$$\frac{p_t(T_1 \mid E)}{p_t(T_2 \mid E)} = \frac{p_t(T_1)}{p_t(T_2)} \cdot \frac{p_t(E \mid T_1)}{p_t(E \mid T_1)}.$$

The first quotient on the right-hand-side is the *prior ratio*, which remains constant as new evidence E is received and the second quotient is the *Bayes factor*, which accounts for the entire impact of E on the relative credence of the two theories (Kass and Raftery 1995).

To guarantee that $p(T_1 \mid E) > p(T_2 \mid E)$, one must impose some further, material restrictions on coherent degrees of belief, but it can be argued that the constraints are presupposed by the very question whether Ockham's razor should be used when choosing between a simple and a complex theory. That places the Bayesian explanation of Ockham's razor in in the same class of a priori metaphysical arguments that includes Descartes' *cogito*, according to which the thesis "I exist" is evidently true each time its truth is questioned. First of all, a Bayesian wouldn't think of herself as choosing between T_1 and T_2 if she started with a strong bias toward one theory or the other: $p_t(T_1) \approx p_t(T_2)$. Second, she wouldn't be choosing between two theories compatible with E unless T_1 explains E, so that $P(E) \approx 1$. Third, she wouldn't say that T_2 is complex unless T_2 has a free parameter parameter i to adjust to save the data. She would not say that the parameter of T_2 is *free* unless she were fairly uncertain about which chance distribution P_i would obtain if T_2 were true: $p_t(P_i \mid T_2) \approx 1/n$. Furthermore, she would not say that the parameter must be *adjusted* to save E unless the chance of E is high only over a narrow range of possible chance distributions compatible with T_2: e.g., $P_0(E) \approx 1$ and for each alternative i such that $0 < i \leq n$, $p_i(E) \approx 0$. It follows from the above assumptions that the prior ratio is approximately 1 and the Bayes' factor is approximately n, so:

$$\frac{p_t(T_1 \mid E)}{p_t(T_2 \mid E)} \approx n.$$

Thus, the simple theory T_1 ends up way more probable than the complex theory T_2 in light of evidence E, as the complex theory T_2 becomes more "adjustable", which is the argument's intended conclusion. When the set of possible chance distributions $\{P_\theta : \theta \in \mathbb{R}\}$ is continuously parameterized, the argument is similar, except that the (discrete) weighted sum expressing $p_t(E \mid T_2)$ becomes a (continuous) integral:

$$p_t(E \mid T_2) = \int P_\theta(E) \cdot p_t(P_\theta \mid T_2) \;\; d\theta,$$

which, again, is clearly weighted by the subjective term $p_t(P_\theta \mid T_2)$.

Each of the above assumptions can be weakened. It suffices that the prior ratio not favor T_2 too much, that the explanation of E by T_1 not be too vague, that the explanation of E by T_2 not be too robust across parameter values and that the distribution of degrees of belief over free parameters of T_2 not be focused too heavily on the parameter values that more closely mimic the predictions of T_1.

The Bayes factor argument for Ockham's razor is closely related to standard paradoxes of indifference. Suppose that someone is entirely ignorant about the color of a marble in the box. Indifference over the various colors implies a strong bias against blue in the partition blue vs. non-blue, whereas indifference over blue vs. non-blue implies a strong bias against yellow. So uniformity over a coarse partition induces a strong bias in a refined partition and uniformity over a fine partition induces a strong bias over a coarser partition. The Bayes' factor argument amounts to plumping for the first bias. Think of the simple theory T_0 as "blue" and of the complex theory T_2 as "non-blue" with a "free parameter" ranging over red, green, yellow, etc. and assume, for example, that the evidence E is "either blue or red". Then, by the above calculation, the posterior ratio of "blue" over "non-blue" is the number n of distinguished non-blue colors. Now consider the underlying prior probability over the refined partition blue, red, green, yellow, etc. It is apparent that "blue" is assigned prior probability $1/2$, whereas each alternative color is assigned $1/2n$, where $n > 1$. Hence, the complex *theory* starts out even with the simple theory, but each complex *possibility* starts out with a large disadvantage. Thus, although "red" objectively "explains" E just as well as "blue" does, the prior bias for "blue" over "red" gets passed through the Bayesian updating formula and begs the question in favor of"blue". One could just as well choose to be "ignorant" over blue, red, green, yellow, etc., in which case "blue" and "red" end up locked in a tie after E is observed and "non-blue" remains more probable than "blue". So the Bayes factor argument again comes down to a question-begging prior bias in favor of simple possibilities.

One can attempt to single out the simplicity bias by expanding the Bayesian notion of rationality to include "objective" constraints on prior probability: e.g., by basing them on the length of Turing machine programs that would produce the data or type out the hypothesis (Jeffreys 1961, Rissannen 2007, Li and Vitanyi 1993). But objective Bayesianism is an epistemological red herring. Even if "rationality" is augmented to include an intuitively appealing, formal rule for picking out some prior biases over others, the real question regarding Ockham's razor is whether such a bias helps one find the truth better than alternative biases (cf. Mitchell 1997). To answer that question relevantly, one must explain, without circular appeal to the very bias in question, whether and in what sense Bayesians who start with a prior bias toward simplicity find the truth better than Bayesians starting with alternative biases would. There are two standard strategies for justifying Bayesian updating. Dutch Book arguments show that violating the Bayesian updating rule would result in preference for combinations of diachronic bets that result in a sure loss over time (Teller 1976). But such arguments do not begin to establish that Bayesian updating leads to higher degrees of belief in true theories in the short run. In fact, Bayesian updating can result in a huge short-run boost of credence in a false theory: e.g., when the the parameters of the true, complex theory are set very close to values that mimic observations fitting a simple alternative. Perhaps the nearest that Bayesians come to taking theoretical truth-conduciveness seriously is to argue that iterated Bayesian updating *converges* to the true theory in the limit, in the sense that $p(T \mid E_n)$ converges to the truth value of T as n increases.[6] But the main shortcoming with that approach has already been discussed: both Ockham and non-Ockham initial biases are compatible with convergent success in the long run. In sum, Bayesians either beg the question in favor of simplicity by assigning higher prior probability to simpler possibilities, or they ignore truth-conduciveness altogether in favor of arguments for coherence, or they fall back upon the insufficient strategy of appealing to long-run convergence.

3 The Argument from Over-fitting

Classical statisticians seek to justify scientific method entirely in terms of objective chances, so the Bayesian explanation of Ockham's razor in terms of Bayes factors and prior probabilities is not available to them. Instead, they pursue the third of the above strategies for explaining Ockham's razor: i.e., they maintain a firm focus on truth-conduciveness but lower their sights from choosing the true theory to choosing the theory that yields the most accurate predictions. If theory T is deterministic and observation is perfectly reliable and T has no free parameters, then prediction involves deducing what will happen from T. If T

[6] Even then, the convergence is guaranteed only with unit probability *in the agent's prior probability.* The non-trivial consequences of that slip are reviewed in (Kelly 1996).

has a free parameter θ, then one must use some empirical data to fix the true value of θ, after which one deduces what will happen from T (e.g., two observed points determine the slope and intercept of a linear law). More generally, fixing the parameter values of T results only in a chance distribution P_θ over possible experimental outcomes. In that case, it is natural to use past experimental data E to arrive at an empirical *estimate* $\widehat{\theta}(T, E')$ for parameter θ. A standard estimation technique is to define $\widehat{\theta}(T, E')$ to be the value of θ that maximizes $P_\theta(E')$. Then $\widehat{\theta}(T, E')$ is called the *maximum likelihood estimate* or MLE of T (given outcome E') and the chance distribution $P_{\widehat{\theta}(T,E')}$ predicts the probability of future experimental outcomes E. The important point is that theory T is not necessarily *inferred* or *believed* in this procedure; T is merely *used* to obtain a hopefully accurate approximation $P_{\widehat{\theta}(T,E')}$ to the true distribution P^* governing the random event E to be predicted. So the aim in choosing T is not to choose the true T but, rather, the T that maximizes the accuracy of the estimate $P_{\widehat{\theta}(T,E')}$ of P^*. Classical statisticians underscore their non-inferential, instrumentalistic attitude toward statistical theories by calling them *models*.

It may seem obvious that no theory predicts better than the true theory, in which case it would remain mysterious why a fixed bias toward simplicity yields more accurate predictions. However, if the data are random, the true theory is complex, the sample is small, and the above recipe for using a theory for predictive purposes is followed, then *a false, overly simplified theory can predict more accurately than the true theory*—e.g., even if God were to inform one that the true law is a degree 10 polynomial, one might prefer, on grounds of predictive accuracy, to derive predictions from a linear law. That surprising fact opens the door to an alternative, non-circular explanation of Ockham's razor in terms of predictive accuracy. The basic idea applies to accuracy in general, not just to accurate prediction. Consider, for example, a marksman shooting at a target. To keep our diagrams as elementary as possible, assume that the marksman is a Flatlander who exists entirely in a two-dimensional plane, so that the target is one-dimensional. There is a wall (line) in front of the marksman and the bull's eye is a distinguished point θ^* on that line. Each shot produced by the marksman hits the wall at some point $\hat{\theta}$, so it is natural to define the *squared error* of shot $\hat{\theta}$ as $(\hat{\theta} - \theta^*)^2$. (figure 3.a). Then for n shots, the average of the squared errors of the n points is a reflection of the marksman's accuracy, because the square function keeps all the errors positive, so none of them cancel.[7] If one thinks of the marksman's shots as being governed by a probability distribution reflecting all the stray causes that affect the marksman on a given shot, then one can explicate the marksman's dispositional *accuracy* as the expected or *mean*

[7] One could also sum the absolute values of the errors, but the square function is far more commonly used in statistics.

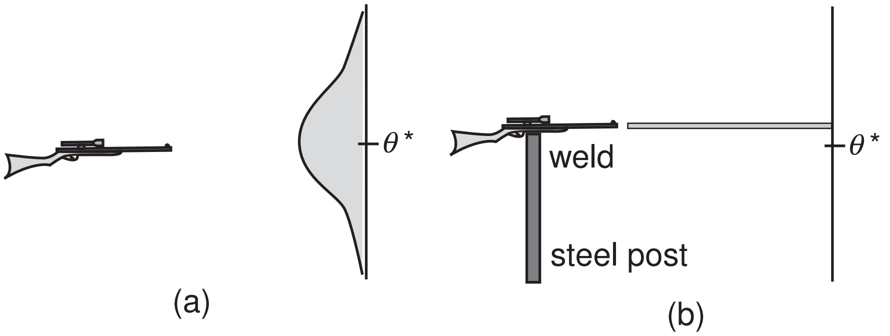

Figure 3: firing range

squared error (MSE) of a single shot with respect to distribution P:

$$MSE_P(\hat{\theta}, \theta^*) = \mathrm{Exp}_P(\hat{\theta} - \theta^*)^2.$$

The MSE is standardly factored in a revealing way into a formula known as the *bias-variance trade-off* (Wasserman 2004):

$$\mathrm{MSE}_P(\hat{\theta}, \theta^*) = \mathrm{Bias}_P(\hat{\theta}, \theta^*)^2 + \mathrm{Var}_P(\hat{\theta}),$$

where $\mathrm{Bias}_P(\hat{\theta}, \theta^*)$ is defined as the deviation of the marksman's average or expected shot from the bull's eye θ^*:

$$\mathrm{Bias}_P(\hat{\theta}, \theta^*) = \mathrm{Exp}_P(\hat{\theta}) - \theta^*;$$

and the *variance* $\mathrm{Var}_p(\hat{\theta})$ is defined as the expected distance of a shot from the average shot:

$$\mathrm{Var}_P(\theta) = \mathrm{Exp}_P((\hat{\theta} - \mathrm{Exp}_P(\hat{\theta}))^2).$$

Bias is a systematic tendency to hit to a given side of the bull's eye, whereas variance reflects spread around the marksman's expected or average shot. Even the best marksman is subject to some variance due to pulse, random gusts of wind, etc., and the variance is amplified systematically as distance from the target increases. In contrast, diligent aim, proper correction of vision, etc. can virtually eliminate bias, so it seems that a marksman worthy of the name should do everything possible to eliminate bias. But that argument is fallacious. Consider the extreme strategy of welding the rifle to a steel post to eliminate variance altogether (figure 3.b). In light of the bias-variance trade-off, the welded rifle is more accurate than honest aiming as long as the squared bias of the welded rifle is less than the variance of the marksman's unconstrained aim. If variance is sufficiently high (due to distance from the target, for example), the welded rifle can be more accurate, in the MSE sense, than skillful, unrestricted aim even if the weld *guarantees a miss.* That is the key insight behind the over-fitting argument.

Welding the rifle to a post is draconian. One can imagine a range of options, from the welded rifle, through various, successively less constraining clamps, to unconstrained aim. For a fixed position θ^* of the bull's eye, squared bias goes down and variance goes up as the aim become less constrained (figure 4). The minimum MSE (among options available) occurs at the "sweet spot" where the squared bias and variance curves cross. Aiming options that are sub-optimal due to high bias are said to *under-aim* (as in "not trying hard enough") and aiming options that are sub-optimal due to high variance are said to *over-aim* (as in "trying too hard"). Note that most of the training provided to marksmen is directed at bias reduction and, hence, could easily result in "over-aiming".

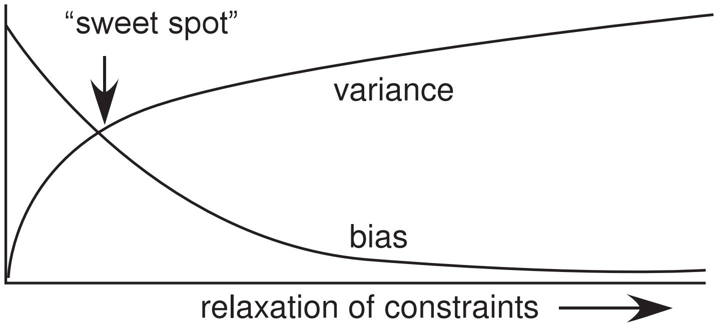

Figure 4: bias vs. variance

So far, the welded rifle strategy looks like a slam-dunk winner over all competing strategies—just hire an accurate welder to obtain a perfect score! But to keep the contest sporting, the target can be concealed behind a curtain until all the welders complete their work. Now the welded rifles still achieve zero variance or spread, but since bias depends on the bull's eye position θ^*, which might be anywhere, the welding strategy cannot guarantee any bound whatever on bias. The point generalizes to other, less draconian constraints on aim—prior to seeing the target there is no guarantee how much extra bias such constraints would contribute to the shot. One could lay down a prior probability reflecting about where the organizers might have positioned the target, but classical statisticians refuse to consider them unless they are grounded in knowledge of objective chance.

Empirical prediction of random quantities is closely analogous to a shooting contest whose target is hidden in advance. The maximum likelihood estimate $\hat{\theta}(T, E')$ is a function of random sample E' and, hence, has a probability distribution P^* that is uniquely determined by the true, sampling distribution P_{θ^*}. Thus, $\hat{\theta}(T, E')$ is like a stochastic shot $\hat{\theta}$ at bull's eye θ^*. When the MLE is taken with respect to the completely unconstrained theory $T_1 = \{P_\theta : \theta \in \Theta\}$, it is known in many standard cases that the MLE is unbiased: i.e., $\mathrm{Bias}_{P^*}(\hat{\theta}(T_1, E'), \theta^*) = 0$. Thus, the MLE based on the complex, unconstrained theory is like the marksman's free aim at the bull's eye. How can that be, when the scientist can't see the bull's eye θ^* she is aiming at? The answer is that *nature* aims the rifle straight at θ^*; the scientist merely chooses whether the rifle will be welded or not and then records the result of the shot. Similarly, the MLE with respect to constrained theory $T_0 = \{P_{\theta_0}\}$ is like shooting with the welded rifle—it has zero variance but no guarantee whatever regarding bias. For a fixed parameter value θ^* and for theories ordered by increasing complexity, there is a "sweet spot" theory T that maximizes accuracy by optimally trading bias for variance. Using a theory simpler than T reduces accuracy by adding extra bias and is called *under-fitting* whereas using a theory more complex or unconstrained than T reduces accuracy by adding variance and is called *over-fitting*. Note that "over-fitting" is defined in terms of the bias-variance trade-off, which is relative to sample size, and definitely *not* in terms of distinguishing structure from noise in the actual world, as some motivational discussions seem to suggest (e.g., Forster and Sober 1994).

To assume a priori that θ_0 is sufficiently close to θ^* for the MLE based on T_0 to be more accurate than the MLE based on T_1 is just another way to beg the question in Ockham's favor. But the choice between basing one's MLE on T_0 or on T_1 is a false dilemma—Ockham's razor says to presume no more complexity than necessary, rather than to presume no complexity at all, so it is up to Ockham's razor to say how much complexity *is* necessary to accommodate sample E'. To put the same point another way, Ockham's razor is not well-defined in statistical contexts until one specifies a formula that *scores* theories in a manner that rewards fit but taxes complexity. One such formula is the Akaike (1973) information criterion (AIC), which ranks theories (lower is better) relative to a given sample E' in terms of the remarkably tidy and suggestive formula:

$$\mathrm{AIC}(T, E) = \text{badness of fit of } T \text{ to } E + \text{complexity of } T,$$

where theoretical complexity is the number of free parameters in T and badness of fit is measured by: $-\ln(P_{\hat{\theta}(T,E')}(E'))$.[8]

Choosing T so as to minimize the AIC score computed from sample E' is definitely one way to strike a precise balance between simplicity and fit. The official theory behind AIC is that the AIC score is an unbiased estimate of a quantity whose minimization would minimize MSE (Wasserman 2004). That sounds remotely comforting, but it doesn't cut to the chase. Ultimately, what matters is the MLE of the whole strategy of using AIC to choose a model and then computing the MLE of the model so chosen. To get some feel for the MLE of the AIC strategy, itself, it is instructive to return to the firing line. Recall

[8] Recall that the MLE $\hat{\theta}(T, E')$ is the value of free parameter θ in theory T that maximizes $P_\theta(E')$, so $P_{\hat{\theta}(T,E')}(E')$ is the best likelihood that can be obtained from T for sample E'. Now recall that $-\ln$ drops monotonically from ∞ to 0 over the half-open interval $(0, 1]$.

that the MLE based on T_0 is like a shot from the welded rifle that always hits point θ_0 and the MLE based on T_1 is like honest, unbiased aiming at the bull's eye after the curtain rises. Using AIC to decide which strategy to employ has the effect of *funneling* shots that fall within a fixed distance r from θ_0 *exactly* to θ_0—call r the *funnel radius.* So on the firing range, AIC could be implemented by making a sturdy funnel of radius r out of battleship plate and mounting it on a firm post in the field so that its spout lines up with the point θ_0 (figure 5). The funnel is a welcome sight when the curtain over the target rises and θ_0 is seen to line up with the bull's eye θ^*, because all shots caught by the funnel are deflected to more accurate positions. In that case, one would like the funnel to have an infinite radius so as to redirect every shot to the bull's eye (which is decision-theoretically identical to welding the rifle to hit point θ_0). The funnel is far less welcome, however, if the intended target is barely obscured by the edge of the funnel, for then then accurate shots get deflected or biased away from the bull's eye, with possibly dire results if the target happens to be hostile (fig. 6). In that case, one would prefer the funnel to have radius 0 (i.e., one would prefer to have it vaporized).

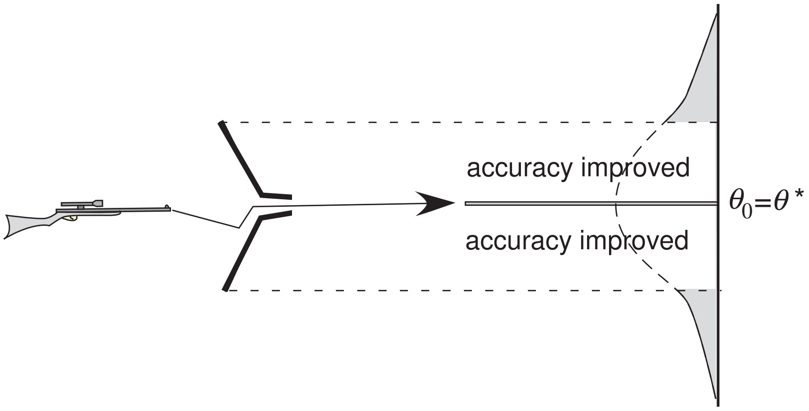

Figure 5: Ockham funnel, best case

More generally, for each funnel radius r from 0 to infinity, one can plot the funnel's MSE over possible bull's eye positions θ^* in order to portray the methods as decision-theoretic acts with MSE as the loss and θ as the state of the world (fig. 7).[9] How, then, does one choose a funnel radius r? Proponents of AIC sometimes speak of typical or anomalous performance, but that amounts to a tacit appeal to prior probabilities over parameter values, which is out of bounds for classical statisticians when nothing is known a priori about the prior location of the bull's eye. One prior-free decision rule is to eliminate *dominated* alternatives, but none

[9] For computer plots of such curves, cf. (Forster 2001).

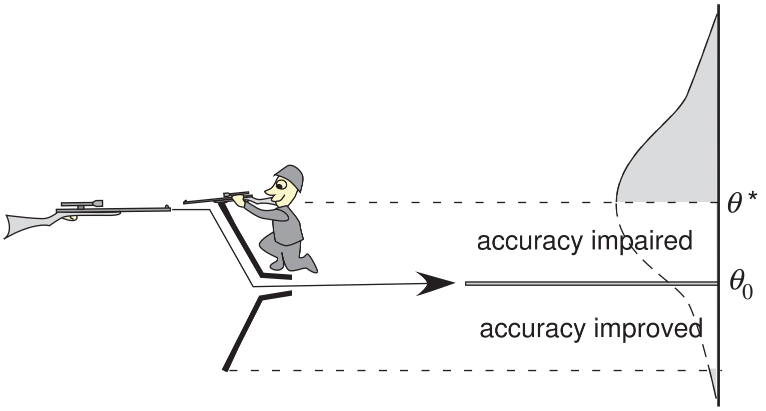

Figure 6: Ockham funnel, worst case

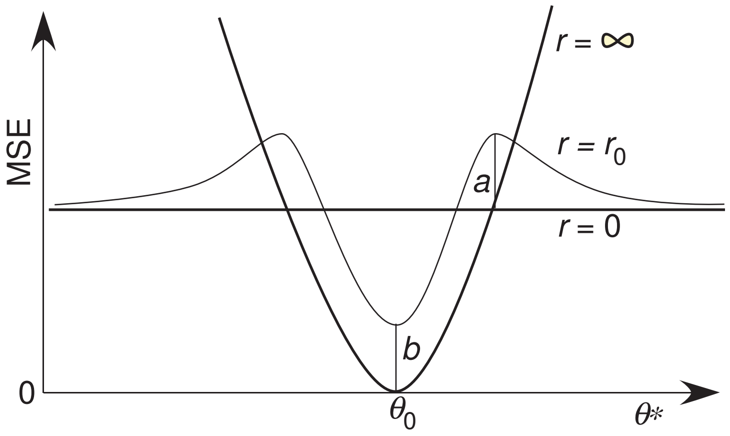

Figure 7: Ockham funnel decision problem

of the options in figure 7 is dominated—larger funnels do better as θ^* approaches θ_0 and smaller ones do better as θ^* diverges from θ_0. Another prior-free decision rule is to choose a *minimax* strategy, i.e., a strategy whose maximum MSE, over all possible values of θ^* is minimal, over all alternative strategies under consideration. Alas, from figure 7, it is clear that the unique minimax solution among the available options is $r = 0$, which corresponds to estimation using the most *complex* theory—hardly a ringing endorsement for Ockham's razor. There is, however, at least one prior-free decision rule that favors a non-extremal funnel diameter $0 < r < \infty$. The *regret* of an option at θ is the difference between the MSE of the option at θ and the minimum MSE over all alternative options available at θ. The *minimax regret* option minimizes worst-case regret. As r goes to infinity, regret a goes up against $r = 0$ and as r goes to 0 the regret b goes up against $r = \infty$. So there must be a "sweet" value r^* of r that minimizes a, b jointly and that yields a minimax regret solution based on the aim of accurate prediction. Then r^* can be viewed as the right balance between simplicity and fit, so far as minimax regret with respect to predictive inaccuracy is concerned. In some applications, it can be shown that AIC is approximately the same as the minimax regret solution when the difference in model complexity is large (Goldenschluger and Greenshtein 2000). AIC is just one representative of a broad range of funnel-like techniques motivated by the over-fitting argument, including cross-validation (Hjorth 1994), Mallows' (1973) statistic, minimum description length (Grünewald 2007), minimum message length, and structural risk minimization (Vapnik 1995).

There are, of course, some objections to the over-fitting argument. (1) The argument irrevocably ties Ockham's razor to randomness, but Ockham's razor doesn't seem to depend on randomness.[10] Intuitively, however, Ockham's razor has to do with uniformity of nature, conservation laws, symmetry, sequential patterns, and other features of the universe that may be entirely deterministic and discretely observable without serious concern about error.

(2) Over-fitting arguments are sometimes presented vaguely in terms of "minimizing" MSE, without much attention to the awkward decision depicted in figure 7 and the consequent need to invoke either prior probabilities or minimax regret as a decision rule.[11] In particular, figure 7 should make it clear that computer simulations of Ockham strategies at "typical" parameter values should not be taken seriously by classical statisticians, who reject prior probabilistic representations of ignorance.

(3) MSE can be challenged as a correct explication of accuracy in some applications. For an extreme example, suppose that an enemy soldier is aiming directly at you. There happens to be a rifle welded to a lamp post that would

[10] It is not correct to insist that all scientific observations are noisy (Sober and Forster 1994), since SRM is routinely applied to dichotomous classification problems in which the binary value (e.g., black vs. white pixels in digitized image processing) is assumed to be observed without noise.

[11] Readers familiar with structural risk minimization (SRM) may suspect otherwise, because SRM theory is based on a function $b(\alpha, n, c)$ such that with worst-case chance $1 - \alpha$, the true MSE of using model T of complexity c for predictive purposes is less than $b(\alpha, n, c)$ (Vapnik 1995). The SRM rule starts with a fixed value $\alpha > 0$ and sample size n and a fixed sequence of models of increasing complexity and then chooses for predictive purposes (at sample size n) the model whose worst-case MSE bound $b(\alpha, n, c)$ is least. Note, however, that the bound is valid only when the model in question is selected and used for predictive purposes *a priori*. Since b can be expressed as a sum of a measure of badnes of fit and a term taxing complexity, SRM is just another version of an Ockham funnel (albeit with a diameter larger than that of AIC). Therefore, the MSE of SRM will be higher than that of the theory SRM selects at the"bumps" in MSE depicted in figure 7. So the (short-run) decision theory for SRM ultimately poses the same problems as the decision theory for AIC. In the long run, SRM converges to the true model and AIC does not but, as has already been explained, long-run convergence does not explain Ockham's razor.

barely miss your opponent and another, perfectly good rifle is lying free on the ground. If you value your life, you will pick up the rifle on the ground and aim it earnestly at your opponent even if you know that the welded rifle has lower MSE with respect to the intended target. For that reason, perhaps, military marksmanship is scored in terms of hits vs. misses on a human silhouette (U.S. Army 2003) rather than in terms of MSE from a geometrical bull's eye.

(4) Finally, the underlying sense of accurate prediction does not extend to predicting the results of novel policies that alter the underlying sampling distribution and, therefore, is too narrow to satisfy even the most pragmatic instrumentalist. That important point is developed in detail in the following section on causal discovery and prediction.

4 Ockham's Causal Razor

Suppose that one employs a model selection technique to accurately estimate the incidence of lung cancer from the concentration of nicotine on teeth and suppose that a strong statistical "link" is found and reported breathlessly in the evening news. Nothing in the logic of over-fitting entails that the estimated correlation would accurately predict the cancer-reducing efficacy of a public tooth-brushing subsidy, for enactment of the policy might *change* the underlying sampling distribution so as to sever the "link". Getting the underlying causal theory wrong can make even the most accurate predictions about the actual population useless for predicting the counterfactual results of enacting new policies on that population.

A possible response is that causal conclusions require controlled, randomized trials, in which case the sample is already taken from the modified distribution and the logic of over-fitting once again applies. But controlled experiments are frequently too expensive or too immoral to perform. Happily, there is an alternative to the traditional dilemma between infeasible experiments and causal skepticism: recent work on causal discovery (Spirtes et al. 2000, Verma and Pearl 1991) has demonstrated that there *is*, after all, a sense in which *patterns* of correlations among several (at least four) variables can yield unambiguous causal conclusions. The essential idea is readily grasped. Let $X \rightarrow Y$ abbreviate the claim that X is a direct cause of Y. Consider the causal situations depicted in figure 8. It is helpful to think of variables as measurements of flows in pipes and of causal relation $X \rightarrow Y$ as a pipe with water flowing from flow meter X to flow meter Y (Heise 1973). In the causal chain $W \rightarrow Y \rightarrow X$, we have three meters connected by a straight run of pipe, so it is clear that information about one meter's reading would provide some information about the other meter readings. But since W informs about X only in virtue of providing information about Y, knowledge of X provides no *further* information about W than Y does—in

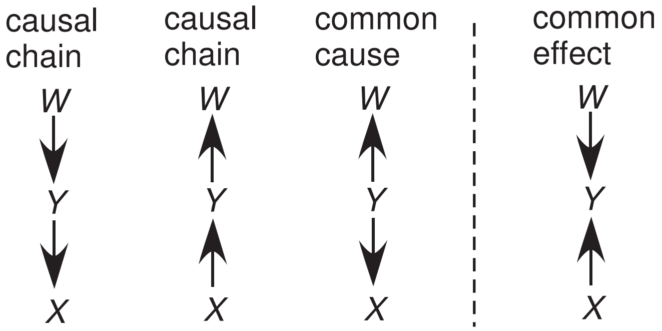

Figure 8: causal situations

jargon, X is independent of W *conditional on* Y. By symmetrical logic, the same holds for the inverted chain $X \rightarrow Y \rightarrow W$. The common cause situation $W \leftarrow Y \rightarrow X$ is the same: W provides information about X only in virtue of providing information about the common cause Y, so conditional on Y, W is independent of X. So far, the situation is pretty grim—all three situations imply the same conditional dependence relations. But now consider the common effect $W \rightarrow Y \leftarrow X$. In that case, W provides no information about X, since the two variables are causally independent and could be set in any combination. But *conditional* on Y, the variable X does provide some information about W because both W and X must collaborate in a specific manner to produce the observed value of Y. Thus, the common effect implies a pattern of dependence and conditional dependence distinct from the pattern shared by the remaining three alternatives. Therefore, common effects and their consequences can be determined from observable conditional dependencies holding in the data.

There is more. Another standard source of causal skepticism is the possibility that apparent causal relation $W \rightarrow X$ is actually produced by a latent or unobserved common cause $W \leftarrow C \rightarrow X$ (just as a puppeteer can make one puppet appear to interact directly with another). Suppose, for example, that Z is a direct effect of common effect Y. Consider the skeptical alternative in which $Y \rightarrow Z$ is actually produced by a hidden common cause C of Y and Z (fig. 9). But the skeptical alternative leaves a footprint in the data, since in the confounded situation Z and W are dependent given Y (since W provides some information about C given Y and C, as a direct cause of Y, provides some information about Y. In the non-confounded situation, the reverse pattern of dependence obtains: W is independent of Z given Y because Z yields information about W only in virtue of the information Z yields about Y. So it is possible, after all, to obtain non-confoundable causal conclusions from non-experimental data.

Given the *true causal theory* relating some variables of interest and given an accurate estimate of the free parameters of the theory, one can obtain accurate

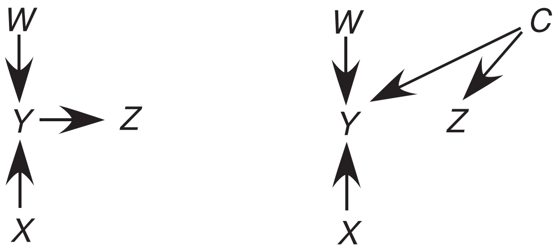

Figure 9: confounding hidden cause

counterfactual predictions according to a natural rule: to predict the result of *forcing* variable X to assume value x, first *erase* all causal arrows into X, holding other theory parameters fixed at their prior values and now use the modified theory to predict the value of the variable of interest, say Y. Thus, for example, if X is itself an effect, forcing X to assume a value will break all connections between X and other variables, so the values of other variables will be predicted not to change, whereas if X is a cause, forcing X to assume a value will alter the values of the effects of X. The moral is that accurate counterfactual predictions depend on inferring the causal model corresponding to the true causal relations among the variables of interest—causal hypotheses are not merely a way to constrain noise in actual empirical estimates.

Causal discovery from non-experimental data depends crucially on Ockham's razor in the sense that causal structure is read off of patterns of conditional correlations and there is a bias toward assuming that a conditional correlation is zero. That is a version of Ockham's razor, because non-zero conditional correlations are free parameters that must be estimated in order to arrive at predictions. Absent any bias toward causal theories with fewer free parameters, one would obtain no non-trivial causal conclusions, since the most complex theory entails a causal connection between each pair of variables and all such causal networks imply exactly the same patterns of conditional statistical dependence. But since the over-fitting argument does not explain how such a bias conduces to true causal structure, it fails to justify Ockham's razor in causal discovery from non-experimental data. The following, novel, alternative explanation does.

5 Efficient Pursuit of the Truth

To summarize the preceding discussion, the puzzle posed by Ockham's razor is to explain how a fixed bias toward simplicity is conducive to finding true *theories*. The crux of the puzzle is to specify a concept of truth-conduciveness according to which Ockham's razor is more truth-conducive than competing strategies. The trouble with the standard explanations is that the concepts of truth-conduciveness they presuppose are either too weak or too strong to single out Ockham's razor as the most truth-conducive inferential strategy. Mere convergence to the truth is too weak, since alternative strategies would also converge to the truth. Reliable indication or tracking of the truth in the short run, on the other hand, is so strict that Ockham's razor can be shown to achieve it only by circular arguments (the Bayes factor argument) or by substituting accurate, non-counterfactual predictions for theoretical truth (over-fitting argument).

There is, however, a third option. A natural conception of truth-conduciveness lying between reliable indication of the truth and mere convergence to the truth is *effective pursuit* of the truth. Effective pursuit is not necessarily direct or even bounded in time or complexity (e.g., pursuit through a labyrinth). But neither is effective pursuit entirely arbitrary—gratuitous course reversals and cycles should evidently be avoided. Perhaps, then, Ockham's razor is the best possible way to pursue the true theory, even though simplicity cannot point at or indicate the true theory in the short run and even though alternative methods would have converged to the truth eventually.

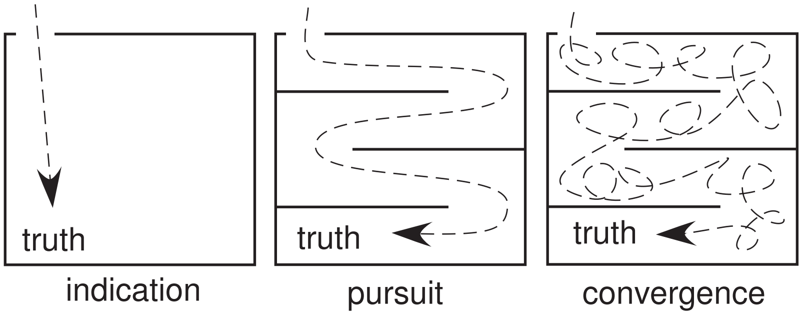

Figure 10: three concepts of conduciveness

In the pursuit of truth, a course reversal occurs when one *retracts* or takes back an earlier belief, as when Ptolemaic theory was replaced by Copernican theory. It caused a sensation when Thomas Kuhn (1962) argued that scientific change essentially involves losses or retractions of content as well as steady accretion and invoked the tremendous cognitive cost of retooling entailed by such changes to explain retention of one's theoretical ideas in the face of anomalies. Emphasis on cognitive retooling may suggest that retractions are a merely "pragmatic" cost, but deeper considerations point to their epistemic relevance. (1) Potential retractions have been invoked in philosophical analyses of the knowledge concept since ancient times. Plato traced the essential difference between knowledge and true belief to the *stability* of knowledge in his dialogue *Meno* and subsequent authors have expanded upon that theme in attempts to provide indefeasibility accounts of knowledge. For example, suppose that one has good but inconclusive evidence E that Jones owns a Ford when, in fact, Smith has one and that one believes, on the basis of E that either Smith or Jones owns a Ford (Gettier 1963). It seems that the inferred belief is not known. Indefeasibility analyses of knowledge (e.g., Lehrer 1990) attempt to explain that judgment in terms of the the potential for retracting the disjunctive belief when the grounds for the false belief are retracted. (2) Deductive logic is *monotonic*, in the sense that additional premises never yield fewer conclusions. Inductive logic is non-monotonic, in the sense that additional premises (new empirical evidence) can undermine conclusions based on earlier evidence. Non-monotonicities are retractions of earlier conclusions, so to minimize retractions as far as finding the truth allows is to approximate deduction as closely as finding the truth allows. (3) Part of the official motivation of belief revision theory (Gardenfors 1988), which is supposed to be a properly epistemic analysis of rational belief change, is that belief change should be deductive whenever new information does not contradict one's current beliefs, which is a synchronic consequence of the the general principle that inference should be as deductive as finding the truth permits. Retraction minimization may be thought of as a diachronic extension of that principle. (4) In mathematical logic, a formal proof system provides just an effective, positive test for theorem-hood—i.e., a Turing machine that *halts* with "yes" if and only if the given statement is a theorem. The halting condition essentially bounds the power of sound proof systems. But nothing other than convention requires a Turing machine to halt when it produces an answer—like human scientists and mathematicians, the Turing machine can be allowed to output a sequence of revised answers upon receipt of further inputs, in an unending loop. Hilary Putnam (1965) showed that Turing machines that are allowed to retract prior answers at most $n + 1$ times prior to convergence to the truth can do more than Turing machines that are allowed to retract at most n times. Furthermore, formal verifiability (halting with "yes" if and only if ϕ is a theorem) is computationally equivalent to finding the right answer with one retraction starting with "no" (say "no" until the verifier halts with "yes" and then retract to "yes"), refutation is computationally equivalent to finding the right answer with one retraction starting with "yes" and formal decidability is computationally equivalent with finding the right answer with no retractions. So retraction bounds are a natural and fundamental *generalization* of the usual computational concepts of verifiability, refutability, and decidability (Kelly 2004). The idea is so natural from a computational viewpoint that theoretical computer scientists interested in inductive inference have developed an elaborate theory of inductive retraction complexity (Case and Smith 1983, Freivalds and Smith 1993). (5) Finally, and most importantly, the usual reason for distinguishing epistemic from merely pragmatic considerations is that the former are truth-conducive and the latter conduce to some other concern (e.g., wishful thinking is happiness-conducive but not truth-conducive). On the contrary, retraction-minimization (i.e., optimally direct *pursuit* of the truth) is part of what it *means* for an inductive inference procedure to be truth-conducive, so retractions are an epistemic consideration.

Additional costs of inquiry may be considered in addition to retractions: e.g., the number and severity of erroneous conclusions are a natural epistemic cost, and the times elapsed until errors and/or retractions are finally avoided. But retractions are crucial for elucidating the elusive, truth-finding advantages of Ockham's razor, for reasons that will become apparent below.

6 An Ockham Efficiency Theorem

Here is a simplistic, logical model of theory choice; important generalizations will be discussed below.[12] A *stream of experience* is a finite or infinite sequence of inputs to science. Infinite streams of experience will be called *empirical worlds*. A given theory T may or may not be *compatible* with a given empirical world.[13] The *empirical content* of a theory is the set of all empirical worlds compatible with the the theory. An *empirical problem* consists of an *empirical presupposition* K which delimits the set of possible worlds and a *theoretical question* Q, which is a collection of mutually incompatible theories whose disjunction is entailed by K. Attention is restricted to empirical problems for which there exists a method that converges to the true theory in each empirical world. That implies that no two theories in Q have overlapping empirical contents.[14]

Recall that testability is a familiar way of understanding simplicity. In that spirit, say that T is *simpler* than theory T' when each finite stream of experience compatible with T is also compatible with T' (twiddle the free parameters of T' until T' fits the data) but each empirical world that would arise if T' were correct becomes incompatible with T eventually.[15] For example, Newton tested

[12] For formal presentations of the model, cf. (Kelly 2007, 2008).

[13] In the philosophy of science literature, compatibility with an infinite stream of experience is called *empirical adequacy.* A theory may be incompatible with an empirical world ϵ while being compatible with each finite initial segment $\epsilon|n$ of ϵ: e.g., the constantly 0 stream of experience is incompatible with "0 infinitely often" but each finite run of zeros is compatible with that hypothesis.

[14] Ptolemy's theory can be tuned to duplicate Copernican observations for eternity. The proposed framework does not apply to that case unless it is assumed that a Ptolemaic universe would not duplicate Copernican appearances for eternity. Even ruling out that possibility, a Ptolemaic universe might duplicate Copernican appearances for an arbitrarily long stretch of time. One good reason for ruling out the possibility of an eternally perfect illusion is that no possible method could converge to the truth in such an empirical world, so even —em optimally truth-conducive methods fail in such worlds.

[15] Cf. also (Schulte 2000, Luo and Schulte 2006).

the simple theory that gravitational mass is identical with inertial mass by filling pendula with distinct materials and failing to observe them out of phase for a long period. Each finite period of time without a noticeable difference is compatible with observing a difference later, but an infinite period with no sign of difference is compatible only with there being no difference.

This concept of simplicity, which can be developed in greater generality than it is here (Kelly 2007, 2008), depends essentially on the structure of the empirical presupposition K and the theoretical question Q one is trying to answer. In that sense, it contrasts with alternative explications of empirical simplicity, such as syntactic or computational compressibility (Li and Vitanyi 1993), which are relative to conventions in natural or computing *languages*. The motive for question-relativity is familiar from the performance analysis of algorithms in computer science. Optimal truth-conduciveness depends on the structure of the truth-finding problem (K, Q) one is trying to solve—on the empirical possibilities K one must succeed over and on the question Q one is trying to answer. Therefore, if Ockham's razor is optimally truth-conducive over a wide range of possible, empirical problems (K, Q), then simplicity must somehow conform itself to the structure of (K, Q).

Now let $T_0, T_1, \ldots$ be a sequence of theories strictly ordered by simplicity in the sense just defined, let K state that some T_i is true, and let Q ask which T_i is true. For example, suppose that there is a black box that occasionally emits discrete, detectable particles and assume that at most finitely many types of particles are emitted for eternity (fig. 11). Nothing further is known, such as when to expect new particle types to be emitted. Theory T_i says that exactly i types of particles will be emitted. The theories are all mutually exclusive and exhaustive, so believing such a theory on the basis of evidence regarding past particle emissions constitutes a genuine inductive inference. Furthermore, any finite run of experience compatible with i particle types is also compatible with $i + 1$ particle types, but not conversely—seeing $i + 1$ types rules out i as the right count. For another example, suppose that the truth is some polynomial law $y = f(x)$ and that the scientist receives a tighter interval around $f(x)$ each time $f(x)$ is measured (cf. Glymour 2001). Let theory T_i say that the polynomial

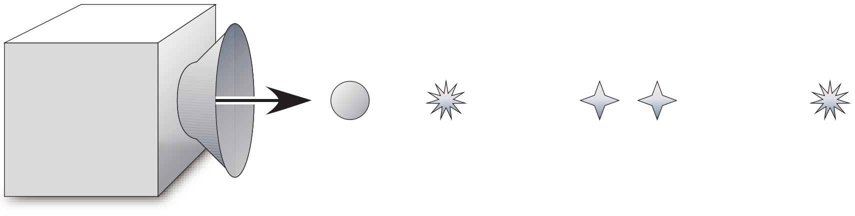

Figure 11: particle emittor

degree of f is exactly $i-1$, so that T_0 says that $f(x) = 0$ and T_{n+1} says that there exist coefficients $\theta_0, \ldots, \theta_n$ with $\theta_n \neq 0$ such that $f(x) = \sum_{i \leq n} \theta_i x^i$. Each finite set of open intervals compatible with T_n is also compatible with T_{n+1}—e.g., each finite collection of open intervals around a linear plot is also compatible with a parabola of very low curvature (fig. 12 (a-b)), but there are three sufficiently small open intervals around the parabola that rule out every line (fig. 12 (c)).[16]

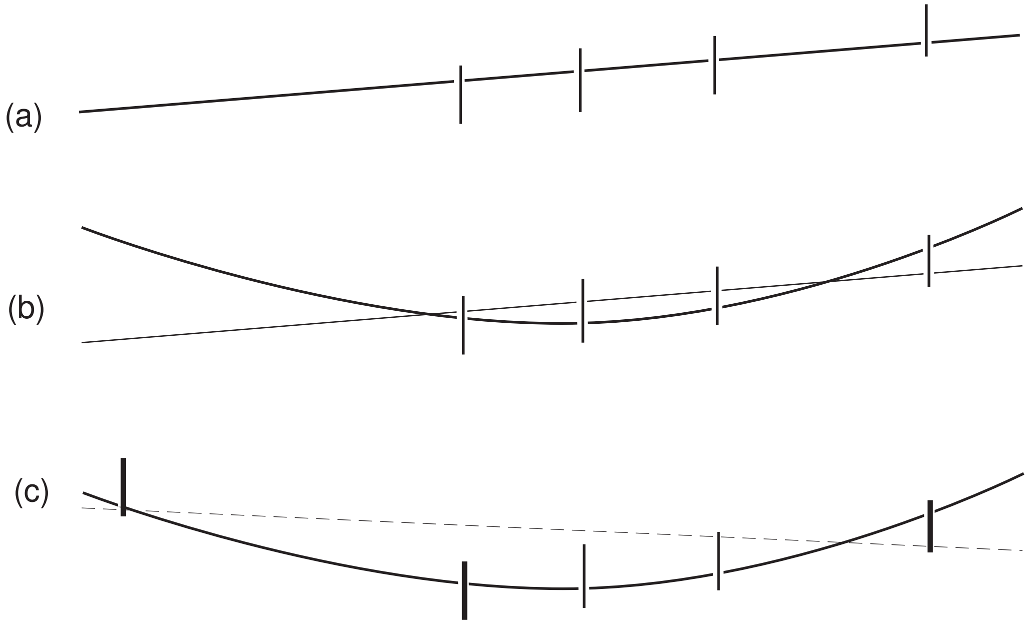

Figure 12: curve-fitting

Let the scientist's strategy for pursuing the truth be a function M from finite sequences of input data to theories (or to '?', which indicates refusal to choose among theories at the current time). *Ockham's razor* says to choose no theory unless it is the uniquely simplest theory compatible with experience. Strictly speaking, Ockham's razor allows for any number of counter-intuitive vacillations between some theory T_i and '?'. A natural companion principle, called *stalwartness* for want of a better name, requires that one hang onto one's current theory T_i as long as T_i remains uniquely simplest among the theories compatible with experience.[17] Two stalwart, Ockham methods then differ only in the amount of

[16] This is very close to Karl Popper's discussion of degrees of falsifiability, except that his approach assumed exact data rather than intervals. The difference is crucial to the following argument.

[17] Since theories are linearly ordered by empirical complexity in this introductory sketch, uniqueness is trivial, but the argument can be extended to the non-unique case, with interesting consequences discussed below.

time it takes to "leap" from '?' to the newly simplest theory after the previously simplest theory is refuted.[18]

Stalwart, Ockham methods are guaranteed to converge to the true theory. For let w be an empirical world satisfying K that is compatible with answer T_i to theoretical question Q. By some finite stage, w refutes each theory simpler than T_i, at which point Ockham's razor licenses selection of T_i. Thereafter, T_i remains uniquely simplest, so stalwartness demands that T_i be retained.

But to resolve the simplicity puzzle, one must show, in addition, that Ockham methods are optimally truth-conducive. Let M be a convergent, stalwart, Ockham method. It is evident that M retracts at most n times if the true theory is T_n since, at worst, such a strategy chooses theories in ascending order (by Ockham's razor) and never retracts a given theory more than once (by stalwartness). So far, that sounds banal—like saying that one always finds what one seeks in the last place one looks (fig. 13). But here is the crucial part: no convergent method can achieve a better worst-case retraction bound than n when T_n is true, so one may say that convergent, stalwart, Ockham methods are *retraction-efficient*. To see why, let M' be an arbitrary strategy that converges, in each empirical world w compatible with empirical presupposition K, to the unique answer to question Q that is compatible with w. Nature is free to present an empirical world w_0 compatible with T_0 until M' converges to T_0. If M' never takes the bait and leaps to conclusion T_0, then nature continues to present w_0 and M' fails to converge to the true answer to Q, contrary to assumption; so M' must leap to T_0 by some finite stage n along w_0. The finite sequence of data presented along w_0 by stage n can be extended to a new empirical world w_1 compatible with T_1. Nature is free to continue presenting data from w_1 until M' leaps to conclusion T_1, and so forth, up to T_n.

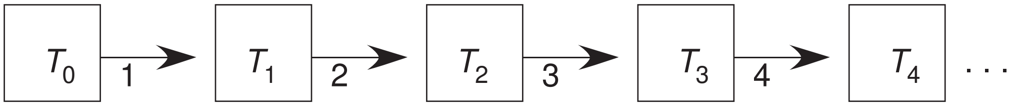

Figure 13: stalwart, Ockham retractions = forcible retractions

To clinch the argument, it will now be shown that all non-Ockham strategies are retraction-*inefficient* in a strong sense—they achieve higher worst-case retraction bounds than Ockham strategies no matter which theory T_i compatible with experience at the time of the violation is true. It is pretty obvious that a truth-seeker should at least converge to the truth and should never produce refuted theories (they can't be true), so restrict attention to convergent, *consis-*

[18] In a similar spirit, Rudolf Carnap's inductive logic (1950) left learning speed as a free parameter λ to be set by the user.

tent methods,[19] which never produce theories that have already been refuted.[20] Assuming that M' converges to the truth, suppose that M' violates Ockham's razor upon seeing the last entry in input sequence e by, say, producing theory T_4 when simpler theory T_2 is still compatible with e (fig. 14). Suppose that M' has retracted r times prior to receipt of the last entry in e (in figure 14, $r = 1$). Nature can continue to present experience compatible with T_2 until M' takes the bait and revises to T_2, on pain of failing to converge to T_2 when T_2 is true. Thereafter, nature can force another $i - 2$ retractions while presenting data making T_i true, for a total of at least $r + i - 1$ retractions. Alternatively, one could switch to a convergent method M that agrees with M' up to the last entry in e and that is Ockham and stalwart from the last entry in e onward. Method M uses at most $r + i - 2$ retractions if T_i is true, which is at least one retraction better than the bound achieved by M' if T_i is true (compare figures 13 and 14). Say, in that case, that M' is *retraction-beaten* by M in T_i. A similar efficiency argument can be given for stalwartness—a convergent method that gratuitously drops the simplest theory can be forced by nature to choose it again, which is an extra retraction on top of all the others nature is in a position to force.

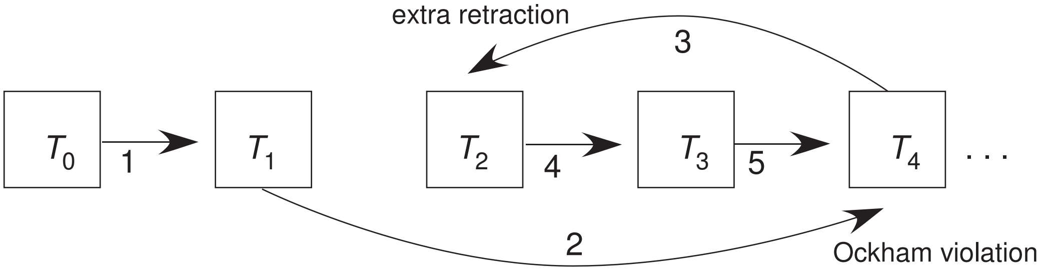

Figure 14: Ockham violator's extra retraction

The preceding discussion proves the following theorem:

Theorem 1 (Kelly 2004) *Assume that the possible theories $T_0, T_1, \ldots$ are sequentially ordered by empirical simplicity and that each theory loses its maximal simplicity status only when it is refuted. Then:*

I. each consistent method M that is henceforth stalwart and Ockham is retraction-efficient over all competing methods that agree with M until now;

[19] Statisticians refer to convergence to the truth as "consistency". Here, "consistency" is used in the logical, rather than the statistical sense. Note that the AIC method, discussed above, is not consistent in that sense (cf. Wasserman 2004).

[20] For computable scientists, there is a possible motive for inconsistency, for there exist some well-defined theoretical questions for which only inconsistent computable methods can converge to the truth (Kelly and Schulte 1995). The crucial non-computability considerations that drive that result do not apply in the elementary sorts of examples considered in this paper, however.

II. each convergent, consistent method M that violates Ockham's razor upon receipt of the last entry in finite input sequence e is beaten in each theory T_i compatible with e by a method that agrees with M prior to receiving the last datum in e and that is stalwart, Ockham thereafter.

Unlike over-fitting explanations, the Ockham efficiency theorem applies to deterministic questions. Unlike the Bayes factor explanation, the Ockham efficiency theorem presupposes no question-begging prior bias in credence toward simple worlds—every world is as important as every other. The crux of any non-circular epistemic argument for Ockham's razor is to explain why leaping to a needlessly complex theory makes one a bad truth-seeker *even if that theory happens to be true*. To see how the hard case is handled in the Ockham efficiency theorem, note that even if T_4 is true in figure 14, leaping straight to T_4 when experience refutes T_1 provides nature with a strategy to force one through the sequence of theories T_4, T_2, T_3, T_4, which not only adds an extra retraction to the optimal sequence T_2, T_3, T_4 but also involves an embarrassing cycle away from T_4 and back to T_4. In terms of the metaphor of pursuit, it is as if the heat-seeking missile *passed* its target and had to make a hairpin turn back to it—a performance likely to motivate some re-engineering. That point makes it clear that the logic of the argument essentially involves worst-case scenarios about the timing of the data—in *some* worlds T_2 and T_3 are refuted before M' loses confidence in T_4, in which case M' would retract less than Ockham. Balancing the worst case against the best case suggests consideration of the expected case, but the expected case involves a circular appeal to simplicity-biased prior probabilities. The point of the worst-case reasoning is to explain Ockham's razor without circles.

It is natural to wonder how the consideration of further costs would affect the argument. For example, eliminating an error requires a retraction, so it might seem that minimizing the total number of errors would compete with minimizing retractions. In general, yes, but in the worst case, no. Assuming the truth of the simplest theory compatible with experience, Ockham's razor commits no errors. In each more complex theory, the total number of errors committed by an arbitrary convergent method is unbounded.[21] So stalwart, Ockham methods are also error-efficient. Furthermore, violating Ockham's razor means that one commits at least one error given that the currently simplest theory is true. So every error-efficient strategy is Ockham.[22] The same is true if one includes elapsed times to retractions or to the last error. What is special about retractions is

[21] Nature can present experience compatible with the currently simplest theory until arbitrary convergent method M produces that theory T and then continue to produce such evidence for another k stages before refuting T to force k errors out of M.

[22] In other words, Ockham methods are weakly Pareto-dominant with respect to total retractions and total errors.

that non-Ockham methods retract strictly more in *every* theory compatible with experience, not just the simplest. A stronger version of Ockham's razor follows if one charges for expansions of belief or for elapsed time to choosing the true theory, for in that case one should avoid agnosticism and select the simplest theory at the very outset to achieve zero loss in the simplest theory compatible with experience. That conclusion seems too strong, however, confirming the intuition that when belief changes, the painful part is retracting the old belief rather than adopting up the new one.

7 Extension to Branching Simplicity

Sometimes, the theories of interest are not ordered sequentially by simplicity, in which case there may be more than one simplest theory compatible with experience. For example, suppose that the question is to determine the set of all particle types that will ever be emitted (figure 11). For a more pressing example, let T_S be the theory that the true causal structure is compatible with exactly the partial statistical dependencies in set S. In the inference of linear causal structures with Gaussian error, the branching simplicity structure over models with three variables is exactly the lattice depicted in figure 15 (cf. Chickering 2003, Meek 1995). In such cases, the empirical complexity of theory T in light of finite input sequence e is the length of the longest simplicity path of theories compatible with e that terminates in T and the empirical complexity of an empirical world w is

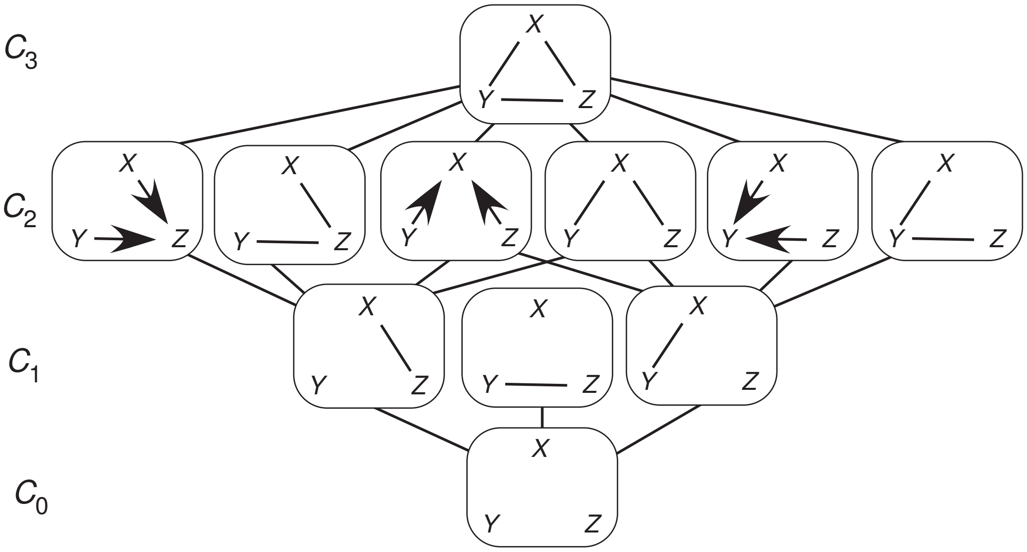

Figure 15: simplicity for acyclic linear causal models

defined to be the empirical complexity of the unique theory w is compatible with. The *complexity set* $C_{K,e}(i)$ is the set of all empirical worlds satisfying background presupposition K that have empirical complexity i in light of e.

When there is more than one simplest theory compatible with experience, Ockham's razor seems to demand that one suspend judgment with '?' until nature winnows the field down to a unique theory. For example, if one were to hear an unusual noise in the particle emitter indicating that a new particle type is about to be discovered, it seems that one should wait for the particle to appear (in contrast, when Ockham's razor licenses the leap to the conclusion that no more particle types are coming, there is no "sign" to wait for and that is why leaping is retraction-efficient in that case).

Suppose that, as in the causal case (figure 15), no path of increasingly complex theories is shorter than any other.[23] Call that the *no short path* assumption. Then violating Ockham's razor by choosing one simplest theory over another incurs an extra retraction in every complexity set, since nature is free to make the *other* simplest theory appear true, forcing the scientist into an extra retraction. Thereafter, nature can force the usual retractions along a path of that visits each non-empty complexity set $C_{K,e}(j)$, by the assumption that no path is short.

Theorem 2 (Kelly 2007) *Assume that there are no short simplicity paths and that each theory loses its maximal simplicity status only when it is refuted. Then:*

I. *each consistent method M that is henceforth stalwart and Ockham is retraction-efficient over all competing methods that agree with M until now;*

II. *each convergent, consistent method M that violates Ockham's razor upon receipt of the last entry in finite input sequence e is beaten in each non-empty complexity set $C_{K,e}(i)$ by a method that agrees with M prior to receiving the last datum in e but is stalwart, Ockham thereafter.*

When the no short path condition is dropped, it is still the case that staying on the Ockham path forever is better than ever to stray. For suppose, in the preceding example, that M outputs '?' and contemplates leaping to T_1, in violation of Ockham's razor. Sticking with the Ockham method guarantees 0 retractions in $C_{K,e}(0)$ and 1 retraction in $C_{K,e}(1)$, whereas the violator can be forced into 1 retraction in $C_{K,e}(0)$ and one retraction in $C_{K,e}(1)$. A more general argument of that kind establishes:

[23] In the case of acyclic linear causal models with independently distributed Gaussian noise, it is a consequence of (Chickering 2003) that the only way to add a new implied conditional dependence relationship is to add a new causal connection. Hence, each causal network with n causal connections can be extended by adding successive edges, so there are no short paths in that application and the strong argument for Ockham's razor holds.

Theorem 3 (Kelly 2007) *Assume that each theory loses its maximal simplicity status only when it is refuted. Then:*

I. methods that are always stalwart and Ockham are retraction-efficient over all consistent, convergent methods;

II. each consistent, convergent method M that violates Ockham's razor or stalwartness on at the end of finite input sequence e is retraction-beaten in at least complexity set $C_{K,e}(0)$ by each stalwart, Ockham method that agrees with M until the end of e.

Without the no short path assumption, methods that refuse to return to the Ockham path are no longer beaten by methods that do. Suppose that T_0 and T_1 are equally simple and that T_2 is more complex than T_1 but not more complex than T_0. Then T_0 and T_1 both receive empirical complexity degree 0 and T_2 is assigned complexity degree 1. Suppose that method M has already violated Ockham's razor by choosing T_0 when T_1 is still compatible with experience. Alas, sticking with the Ockham violation *beats* Ockham's retreating strategy in terms of retractions. For Ockham's retreat counts as a retraction in $C_{K,e}(0)$. Nature can still lure Ockham to choose T_0 and can force a further retraction to T_1 for a total of 2 retractions in $C_{K,e}(1)$. But strategy M retracts just once in $C_{K,e}(0)$ and once in $C_{K,e}(1)$.

One response to the preceding limitation is to question whether the short path really couldn't be extended—with all infinite paths, there are no short paths. A second response is to weaken Ockham's razor to allow for favoring simplest theories on longer paths, in which case M does not violate Ockham's razor and is retraction-efficient. A third response is to consider extra costs like total number of errors or elapsed time to each retraction, for the newly Ockham method commits fewer errors in $C_{K,e}(0)$ than M does and both methods commit arbitrarily many errors or arbitrarily late retractions in each non-empty class $C_{K,e}(i+1)$. A fourth response is that the simplicity degrees assigned to theories along a short path are arbitrary as long as they preserve order along the path. The proposed definition of simplicity degrees ranks theories along a short complexity path as low as possible, but one might have ranked them as high as possible (e.g., putting T_0 in $C_{K,e}(1)$ rather than in $C_{K,e}(0)$, in which case the preceding counterexample no longer holds.[24] That option is no longer available, however, if some path is infinite in length, but then, perhaps, the simplicity degrees of theories along the short path are simply indeterminate.

[24] That approach is adopted, for example, in earlier work by (Freivalds and Smith 1993).

8 Ockham Efficiency when Defeat does not Imply Refutation

The preceding theorems all assume that only refutation can demote a theory from the status of being simplest in light of the data, so that logical consistency with the data forces science to give up the uniquely simplest theory as soon as it is no longer simplest. But that assumption is sometimes violated. For a rather artificial example, consider the question whether the total number of fundamental particle types is even or odd. In that case, it seems plausible to say that "even" is simplest when the current count is 2, but not when the count rises to 3. The appearance of a third particle type deposes "even" from the status of simplest theory compatible wth experience without refuting "even" once for all.

A more pressing example concerns the status of a single causal relation $X \to Y$. One very interesting sequence of causal theories nature can force every convergent method to produce is presented in figure 16. Focus on the causal relation between X and Y. Note that the orientation of the edge flips when the inferred common effect at Y is canceled through discovery of new causal connection $V - Z$ and is flipped in the opposite direction by the inference of a common effect at X. The process can be iterated by canceling the new common effect and re-introducing one at Y, etc. So nature can force an arbitrary, convergent method to cycle any number of times between the opposite causal conclusions $X \to Y$ and $Y \to X$.[25] The causal flips depicted in figure 16 have been elicited (in probabil-

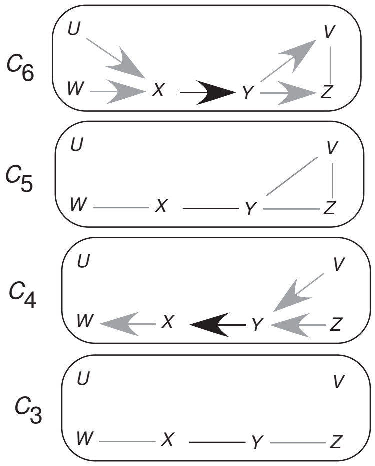

Figure 16: causal flipping

[25] In fact, it can be demonstrated that arbitrarily long causal chains can be flipped in this way.

ity) from the PC causal discovery algorithm (Spirtes et al. 2000) using computer simulated random samples of increasing size from a fixed causal model.

When the question at hand concerns only the causal connection $X - Y$, one can no longer rely on consistency with the data to explain why it is wrong to hang on to a theory after it ceases to be uniquely simplest in light of the data. Furthermore, hanging on "too long" to the causal hypothesis $X \rightarrow Y$ incurs no extra retractions. Hanging onto the wrong orientation too long does incur extra errors in the simplest complexity set compatible with experience, because the Ockham method produce no errors in that complexity set, but it would be more satisfactory if Ockham violators were to do worse in every complexity set. One way to obtain that result is to consider not only retractions but the *times* at which they occur—delaying a scientific revolution only makes the retooling of paradigms and reprinting of textbooks harder and for purely epistemic reasons it is desirable to escape from the hubris preceding a destined retraction as soon as possible. Represent the total timed retractions of (convergent) method M in empirical world w as the (finite) sequence $(r_1, \ldots, r_n)$ of times at which the n retractions performed by M occur prior to convergence. Then for two such sequences γ, γ', say that $\gamma \leq \gamma'$ if and only if there exists a sub-sequence γ'' of γ' whose length is the same as that of γ such that each entry in γ is no greater than the corresponding entry in γ''. Define $\gamma < \gamma'$ to hold if and only if $\gamma \leq \gamma'$ but not conversely. The point is that late retractions in $C_{K,e}(0)$ get "carried up" to higher complexity classes, allowing one to recover the strong conclusion of theorem 2 without assuming consistency with the data.

Theorem 4 (Kelly 2006) *If retraction times are considered in addition to the number of retractions, then theorems 2 and 3 continue to hold without the requirement that each theory loses its maximal simplicity status only when it is refuted.*[26]

[26] Here is a sketch of the proof of the no short path case. Suppose that M' violates Ockham's razor by hanging onto "even" when the third particle type is discovered, say, upon receiving the last entry in finite input sequence e of length m. Nature can present no further particle types until M' retracts "even" in favor of "odd", which happens no sooner than stage $n + 1$. So in complexity set $C_{K,e}(0)$, M' incurs timed retractions at least as bad as $(r_1, \ldots, r_k, m)$, where $r_1, \ldots, r_n$ are the previous retractions of M'. Let M be just like M' except that M switches to a convergent, stalwart, Ockham strategy at stage n, which forces M to retract "even" immediately. So in complexity set $C_{K,e}(0)$, method M' incurs timed retractions $(r_1, \ldots, r_k, m) < (r_1, \ldots, r_k, m + 1)$. In complexity class $C_{K,e}(i)$, nature can force i more retractions arbitrarily late (by withholding new particle types as long as she pleases), so the worst-case timed retraction bounds for M' and M are respectively: $(r_1, \ldots, r_k, m, \omega, \ldots, \omega) < (r_1, \ldots, r_k, m + 1, \omega, \ldots, \omega)$, where ω is repeated i times.

9 Extension to Randomized Scientific Strategies

The preceding theorems assume that the scientist's method is a deterministic function of the input data. It is frequently the case, however, that randomized or "mixed" strategies achieve lower worst-case cost than deterministic strategies. For example, if the problem is to guess which way a coin lands inside of a black box and the loss is 0 or 1 depending on whether one is right or wrong, guessing randomly achieves a worst-case expected loss bound of 1/2, whereas the lowest worst-case loss bound achieved by either pure (deterministic) strategy is 1. Nonetheless, the Ockham efficiency argument can be extended to show that deterministically stalwart, Ockham strategies are efficient with respect to all convergent mixed scientific strategies, where convergence efficiency is defined in terms of *expected* retractions is now understood as convergence *in probability*, meaning that the objective *chance* (grounded in the method's internal coin-flipper) that the method produces the true theory goes to one as experience increases (Kelly and Mayo-Wilson 2009).

Theorem 5 (Kelly and Mayo-Wilson 2009) *All of the preceding theorems extend to random empirical methods when retractions are replaced with expected retractions and retraction times are replaced with expected retraction times.*[27]

The extension of the Ockham efficiency theorem to random methods and expected retraction times strongly suggests a further extension to probabilistic theories and evidence (i.e., statistical theoretical inference), but that step raises further issues akin to those that arise in the over-fitting argument. A statistical theory choice method still looks like one of the funnels discussed earlier. Funnel-like methods tend to pick up extra retractions in chance when the actual world is near the edge of the funnel, so it is no longer possible to achieve a worst case bound of n expected retractions in $C_{K,e}(n)$, so the upper bounds no longer meet the lower bounds in the efficiency argument. That issue remains to be addressed.

[27] Here is the proof strategy. A method is said to retract T_i *in chance* to degree r at stage $k+1$ if the chance that T_i produces T_i goes down by r from k to $k+1$. Total retractions in chance are summed over theories and stages of inquiry, so as the chance of producing one theory goes up, the chance of producing the remaining theories goes down. Therefore, nature is in a position to force a convergent method to produce total retractions arbitrarily close to i by presenting an infinite stream of experience ϵ making T_i true. It is readily shown that the total retractions in chance along ϵ are a lower bound on expected total retractions along *epsilon*. It is also evident that for deterministic strategies, the total expected retractions are just the total deterministic retractions. So, since deterministically Ockham strategies retract at most i times given that T_i is true, they are efficient over all mixed strategies as well, and violating either property results in inefficiency.

10 Disjunctive Beliefs, Retraction Degrees, and a Gettier Example

Use of '?' to indicate refusal to choose a particular theory is admittedly crude. When there are two simplest theories T_1, T_2 compatible with the data, it is more realistic to allow retreat to the disjunction $T_1 \vee T_2$ than to a generic refusal to choose at all—e.g., uncertainty between two equally simple orientations of a single causal arrow does not necessarily require (or even justify) retraction of all the other causal conclusions settled up to that time. Accordingly, method M will now be allowed to produce finite *disjunctions* of theories in Q. Suppose that there are mutually exclusive and exhaustive theories $\{T_i : i \leq n\}$ and let $\mathbf{x}$ be a Boolean n-vector. Viewing $\mathbf{x}$ as the indicator function of finite set $S_{\mathbf{x}} = \{i \leq n : x_i = 1\}$, one can associate with $\mathbf{x}$ the disjunction:

$$T_{\mathbf{x}} = \bigvee_{i \in S_{\mathbf{x}}} T_i.$$

A retraction now occurs whenever some disjunct is added to one's previous conclusion, regardless how many disjuncts are also removed. Charging one unit per retraction, regardless of the total content retracted, amounts to the following rule:

$$\rho_{\mathrm{ret}}(T_{\mathbf{x}}, T_{\mathbf{y}}) = \max_i \; y_i - x_i.$$

One could also charge one unit for each disjunct added to one's prior output, regardless how many disjuncts are removed, which corresponds to the slightly modified rule:

$$\rho_{\mathrm{dis}}(T_{\mathbf{x}}, T_{\mathbf{y}}) = \sum_i y_i \mathbin{\dot{-}} x_i,$$

where the cutoff subtraction $y \dot{-} x$ assumes value 0 when $x \geq y$.[28] Assuming no short simplicity paths, charging jointly for the total number of disjuncts added and the times at which the disjuncts are added allows one to derive stronger versions of Ockham's razor and stalwartness from retraction efficiency. The strengthened version of Ockham's razor is that one should never produce a disjunction stronger than the disjunction of all currently simplest theories (disjunctions take the place of '?') and the strengthened version of stalwartness is that one should never disjoin a theory T to one's prior conclusion unless T is among the currently simplest theories.[29]

[28] The same formula takes a finite value for a countable infinity of dimensions as long as each disjunction has at most finitely many disjuncts.

[29] It is still the case that nature can force at least n retractions in complexity set C_n and stalwart, Ockham methods retract no more than that. If M violates the strengthened version

When there are short simplicity paths, the Ockham efficiency argument can fail for both of the proposed retraction measures. It is interesting that the counterexample concerns Ockham's razor, but is also reminiscent of Gettier's (1963) counterexample to the justified true belief analysis of knowledge (fig. 17). Suppose that T_0 is simpler than T_1 and T_2 and that T_2 is simpler than T_3. Suppose that experience e is compatible with T_0 and that M produces the disjunction of $(T_0 \vee T_2)$ in response to e "because" M believes that T_0 on the basis of Ockham's razor and the disjunction follows from T_0. If T_1 true, then M has true belief $(T_0 \vee T_2)$ "for the wrong reason"—a Gettier case. Suppose that T_0 is refuted. An Ockham method should now retract to $(T_1 \vee T_2)$, but M *expands* to T_2 "because" M believed $(T_0 \vee T_2)$ and learned that $\neg T_0$. If the truth is T_1, then both methods have 1 retraction on either retraction measure and Ockham incurs the retraction earlier, so Ockham (barely) wins in $C_{K,e}(0)$ after T_0 is refuted. But M wins by retracting only once in $C_{K,e}(0)$, when T_3 is true.[30] Possible responses to the issue of short simplicity paths include those discussed above in section 7.

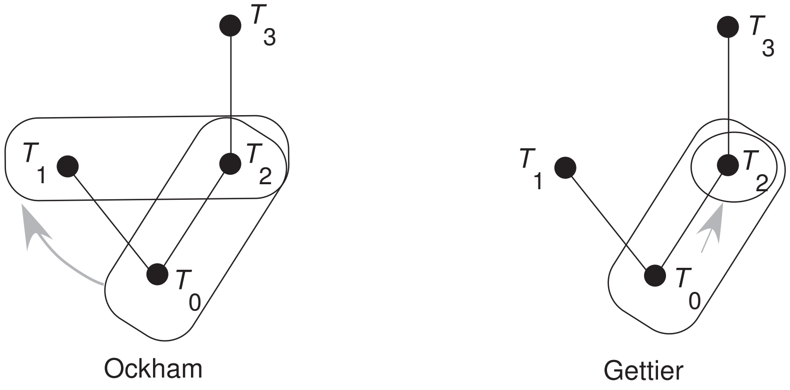

Figure 17: Gettier counterexample to Ockham efficiency

of Ockham's razor, M produces a disjunction missing some simplest theory T. Nature is now free to force M down a path of increasingly complex theories that begins with T. By the no short paths assumption, this path passes through each complexity set, so M incurs at least one extra retraction in each complexity set. If M violates the strengthened version of stalwartness, then M retracts by adding a complex disjunct T. Nature is free to present a world of experience for a simplest world, forcing M to retract disjunct T.

[30] To see why short paths are essential to the example, suppose that there were a theory T_4 more complex than T_1. Then M would also retract twice in complexity set C_2 and Ockham would complete the retraction in C_1 earlier.

11 Extension to Degrees of Belief

Bayesian agents may use their degrees of belief to choose among potential theories (Levi 1983), but they may also regard updated degrees of belief as the ultimate product of scientific inquiry. It is, therefore, of considerable interest to extend the logic of the Ockham efficiency theorems from problems of theory choice to problems of degree of belief assignment. Here are some recent ideas in that direction.

Suppose that the theories under consideration are just T_1, T_2, T_3, in order of increasing complexity. Then each prior probability distribution p over these three theories can be represented uniquely as the ordered triple $\mathbf{p} = (p(T_1), p(T_2), p(T_3))$. The extremal distributions are the basis vectors $\mathbf{i}_1 = (1, 0, 0)$, $\mathbf{i}_2 = (0, 1, 0)$, and $\mathbf{i}_3 = (0, 0, 1)$ and all other coherent distributions lie on the *simplex* or triangle connecting these points in three-dimensional Euclidean space. A standard argument for distributing degrees of belief as probabilities (de Finetti 1975, Rosenkrantz 1983, Joyce 1998) is that each point $\mathbf{x}$ off of the simplex is farther from the true corner of the simplex (whichever it might be) than the point $\mathbf{p}$ on the simplex directly below $\mathbf{x}$, so agents who seek immediate proximity to the truth should stay on the surface of the simplex—i.e., be coherent (fig. 18 (a)).

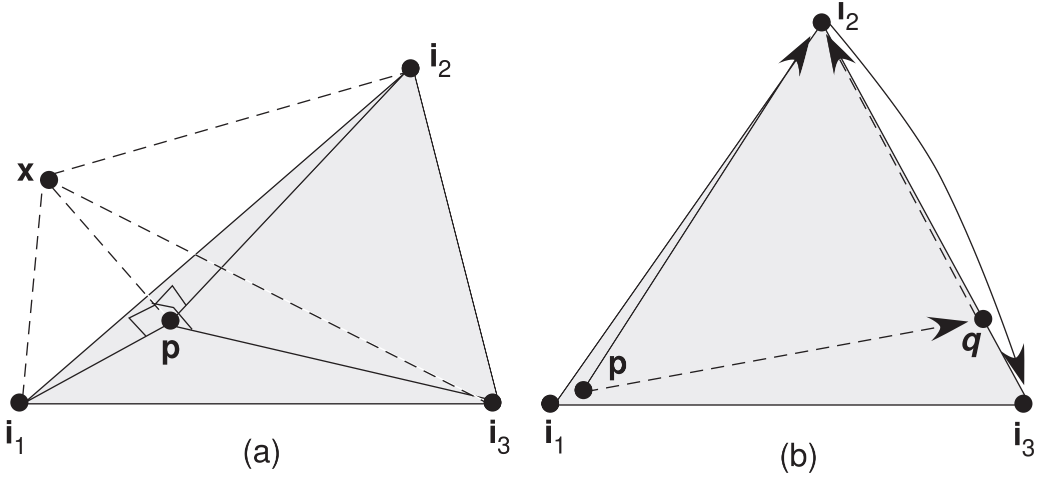

Figure 18: distance from the truth vs. efficient pursuit of the truth

It is natural to extend that static argument to the active *pursuit* of truth in terms of total Euclidean distance traversed on the surface of the simplex prior to convergence to the truth (fig. 18 (b)). As in section 6, nature has a strategy to force each convergent Bayesian arbitrarily close to $\mathbf{i}_1$, then arbitrarily close to $\mathbf{i}_2$ and then all the way to $\mathbf{i}_3$. Each side of the triangular simplex has length $\sqrt{2}$, so if one adopts $\sqrt{2}$ as the unit of loss, then nature can force retraction bound k

in complexity set $C_{K,e}(k)$, just as in the discussion of theory choice. Therefore, the path $(\mathbf{p}, \mathbf{i}_2, \mathbf{i}_3)$ is efficient, since it achieves that bound. Furthermore, suppose that method M favors complex theory T_2 over simpler theory T_1 by moving from $\mathbf{p}$ to $\mathbf{q}$ instead of to $\mathbf{i}_2$. Then nature can force M back to $\mathbf{i}_2$ by presenting simple data. So the detour through $\mathbf{q}$ results, in the worst case, in the longer path $(\mathbf{p}, \mathbf{q}, \mathbf{i}_2, \mathbf{i}_3)$ that hardly counts as an efficient pursuit curve ($\mathbf{q}$ is passed *twice*, which amounts to a needless cycle).

An ironic objection to the preceding argument is that the conclusion is too *strong*—efficiency measured by total distance traveled demands that one start out with full credence in the simplest theory and that one leap immediately and fully to the newly simplest theory when the previous simplest theory is refuted. Avoidance of that strong conclusion was one of the motives for focusing on retractions as opposed to expansions of belief in problems of theory choice, since movement from a state of suspension to a state of belief is not counted as a retraction. Euclidean distance charges equally for expansions and retractions of Bayesian credence, so it is of interest to see whether weaker results can be obtained by charging only for Bayesian retractions.

One obvious approach is to define Bayesian retractions as increases in *entropy*, defined as:

$$M(q) = -\sum_i q_i \log_2 q_i.$$

That is wrong, however, since the circuit path $(\mathbf{i}_1, \mathbf{i}_2, \mathbf{i}_1)$ seems to incur two large retractions, but entropy remains constantly 0. A more sophisticated idea is to tally the cumulative increases in entropy along the entire path from p to q, rather than just at the endpoints. But that proposal still allows for "retraction-free" circuits around the entropy peak at the midpoint $(1/3, 1/3, 1/3)$ along a path of constant entropy. The same objection obtains if entropy is replaced with any alternative scalar field that plausibly represents informativeness.

Another idea is to measure the retractions from p to q in terms of a popular measure of separation for probability distributions called the *Kullback Leibler (KL) divergence* from p to q:

$$KL(q|p) = \sum_i q_i \log_2 \frac{q_i}{p_i}.$$

KL divergence is commonly applied to measure motions on the simplex in Bayesian experimental design, where the idea is to design the experiment that maximizes the KL divergence from the prior distribution p to the posterior distribution q (Chaloner and Verdinelli 1995). It is well known that KL divergence is not a true distance measure or *metric* because it is asymmetrical and fails to satisfy the triangle inequality. It is interesting but less familiar that the asymmetry amounts to a bias against retractions: e.g., if $\mathbf{p} = (1/3, 1/3, 1/3)$ and $\mathbf{q} = (.999, .0005, .0005)$ then $KL(p|q) \approx .5.7$ and $KL(q|p) \approx 1.6$. Unfortunately, KL divergence cannot be used to measure retractions after a theory is refuted because it is undefined (due to taking $\log(0)$) for any motion terminating at the perimeter of the simplex. But even if one approximates such a motion by barely avoiding the perimeter, KL divergence still charges significantly more for hedging one's bets than for leaping directly to the current simplest theory. For example, if $\mathbf{p} = (.999, .0005, .0005)$, $\mathbf{q} = (.0001, .5, .4999)$, $\mathbf{r} = (.0005, .9995, .0005)$, then the KL divergence along path $(\mathbf{p}, \mathbf{r})$ is nearly 10.9, whereas the total KL divergence along path $(\mathbf{p}, \mathbf{q}, \mathbf{r})$ is around 17.7.

Here is an entirely different approach, motivated by a fusion of logic and geometry, that yields Ockham efficiency theorems closely analogous to those in the disjunctive theory choice paradigm.[31] The simplex of coherent probability distributions over T_0, T_1, T_2 is just the intersection of the unit cube with a plane touched by each of the unit vectors (fig. 19). The Boolean vectors labeling vertices of the unit cube are the labels of the possible disjunctions of theories (the origin $\mathbf{0}$ = (0,0,0) corresponds to the empty disjunction or contradiction). To extend that picture to the entire unit cube, think of $T_\mathbf{x}$ as a *fuzzy* disjunction

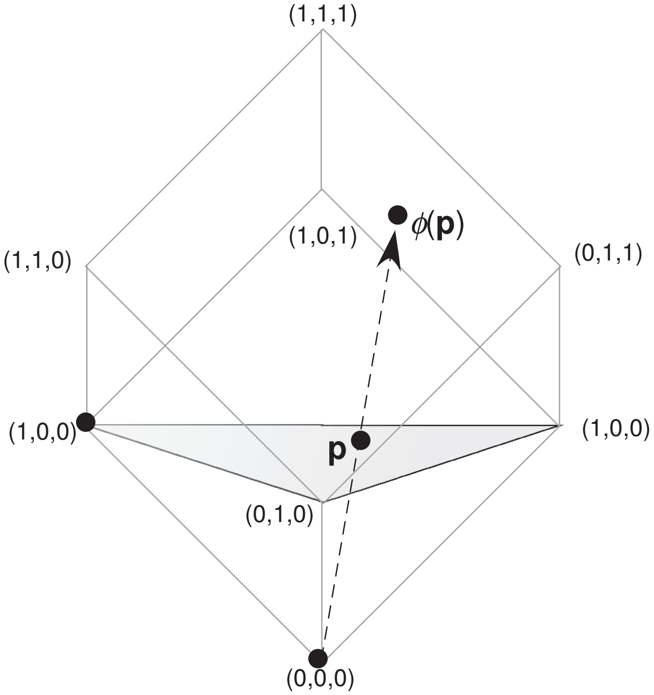

Figure 19: simplex and unit cube

[31] The following definitions and results were developed in collaboration with Hanti Lin.

in which theory T_i occurs to degree x_i (Zadeh 1965). Say that $T_{\mathbf{x}}$ is *sharp* when $\mathbf{x}$ is Boolean and say that $\mathbf{y}$ is *sharp* when $\mathbf{y}$ is a unit vector. Each vector $\mathbf{y}$ in the unit cube can also be viewed as a fuzzy assignment of semantic values to the possible theories. Define the *valuation* of $T_{\mathbf{x}}$ in $\mathbf{y}$ to be the inner product: $\tau_{\mathbf{y}}(T_{\mathbf{x}}) = \mathbf{y} \cdot \mathbf{x} = \sum_i \mathbf{y}_i \cdot \mathbf{x}_i$. If $\mathbf{y}$ and $T_{\mathbf{x}}$ are both sharp, then $\tau_{\mathbf{y}}(T_{\mathbf{x}})$ is the classical truth value of $T_{\mathbf{x}}$ in $\mathbf{y}$ and if p is a probability and $T_{\mathbf{x}}$ is sharp, then $\tau_{\mathbf{p}}(T_{\mathbf{x}}) = p(T_{\mathbf{x}})$.[32] Entailment is defined by: $T_{\mathbf{x}} \models T_{\mathbf{y}}$ if and only if $\tau_{\mathbf{z}}(T_{\mathbf{x}}) \leq \tau_{\mathbf{z}}(T_{\mathbf{y}})$, for each vector $\mathbf{z}$. Thus, $T_{\mathbf{x}} \models T_{\mathbf{y}}$ holds if and only if $x_i \leq y_i$, for each i. The entire Euclidean unit cube may now be viewed as a Boolean algebra under the standard intersection, union, and complementation operations of fuzzy set theory. The *fully consistent* disjunctions are the fuzzy disjunctions that evaluate to 1 in some sharp assignment. They comprise exactly the upper three faces of the unit cube. The vertices of those faces are the consistent, sharp disjunctions of classical logic.

The formulas for retraction measures ρ_{ret} and ρ_{dis} are already defined over the entire unit cube and, hence, may be applied directly to probability assignments. That is not the right idea, however, for it is natural to view the move from $(0, 1/2, 1/2)$ to $(0, 1, 0)$ as a pure expansion of credence, but both retraction measures assign retraction 1/2 in this case. As a result, efficiency once again demands that one move immediately to full credence in T_1 when T_0 is refuted.

Here is a closely related idea that works. The grain of truth behind probabilistic indifferentism is that the sharp disjunction $T_{(1,1,0)} = T_1 \vee T_2$ more faithfully summarizes or expresses the uniform distribution $(1/2, 1/2, 0)$ than the biased distribution $(1/3, 2/3, 0)$; a view that can be conceded without insisting, further, that uniform degrees of belief should be adopted. One explanation of the indifferentist intuition is geometrical—the components of $\mathbf{p} = (1/2, 1/2, 0)$ are proportional to the components of $\mathbf{x} = (1, 1, 0)$ in the sense that there exists constant c such that $\mathbf{x} = c\mathbf{p}$. To be assertible, a proposition should be fully consistent. T_p satisfies the proportionality condition for p but is not fully consistent. Accordingly, say that $T_{\mathbf{x}}$ *expresses* p just in case $T_{\mathbf{x}}$ is fully consistent and $\mathbf{x}$ is proportional to $\mathbf{p}$. Sharp propositions cannot express non-uniform distributions, but fuzzy propositions can: e.g., $T_{(1/2,1,0)}$ expresses $(1/3, 2/3, 0)$ in much the same, natural way that $T_{(1,1,0)}$ expresses $(1/2, 1/2, 0)$.[33] Each fully consistent disjunction has a unit component, which fixes the constant of proportionality at

[32] It is tempting, but not necessary for our purposes, to define $p(T_{\mathbf{x}}) = p \cdot \mathbf{x}$ for non-sharp $T_{\mathbf{x}}$ as well.

[33] A disanalogy: $\tau_{(1/2,1/2,0)}(T_{(1,1,0)}) = 1$, but $\tau_{(1/3,2/3,0)}(T_{(1/2,1,0)}) = 5/6$, so the expression of a uniform distribution is also the support of the distribution, but that fails in the non-uniform case.

$1/\max_i p_i$. Thus, the unique, propositional expression of p is $T_{\phi(\mathbf{p})}$, where:

$$\phi(\mathbf{p})_i = \mathbf{p}_i / \max_i \mathbf{p}_i.$$

Geometrically, $\phi(\mathbf{p})$ can be found simply by drawing a ray from $\mathbf{0}$ through $\mathbf{p}$ to the upper surface of the unit cube (fig. 19).

One can now define probabilistic retractions as the logical retractions of the corresponding, propositional expressions:

$$\begin{aligned}\rho_{\text{ret}}(\mathbf{p}, \mathbf{q}) &= \rho_{\text{ret}}(T_{\phi(\mathbf{p})}, T_{\phi(\mathbf{q})});\\ \rho_{\text{dis}}(\mathbf{p}, \mathbf{q}) &= \rho_{\text{dis}}(T_{\phi(\mathbf{p})}, T_{\phi(\mathbf{q})}).\end{aligned}$$

In passing, one can also define Bayesian *expansions* of belief by permuting $\mathbf{p}$ and $\mathbf{q}$ on the right-hand-sides of the above formulas. Revisions are then the sum of the expansions and retractions. Thus, one can extend the concepts of belief revision theory (Gärdenfors 1988) to Bayesian degrees of belief—an idea that may have useful applications elsewhere, such as in Bayesian experimental design.

Both retraction measures have the natural property that if T_i is the most probable theory under p, then for each alternative theory T_j, the move from p to the conditional distribution $p(.|\neg T_j)$ incurs no retractions (Lin 2009). Moreover, for purely retractive paths (paths that incur 0 expansions), the disjunctive measure is attractively *path-independent*:

$$\rho_{\text{dis}}(p, r) = \rho_{\text{dis}}(p, q) + \rho_{\text{dis}}(q, r).$$

Most importantly, both measures entail simplicity biases that fall short of the implausible demand that one must leap to the currently simplest theory immediately (fig. 20). For ρ_{ret}, the zone of efficient moves from $\mathbf{p}$ to the next simplest vertex $\mathbf{j}$ when nearby vertex $\mathbf{i}$ is refuted is constructed as follows. Let $\mathbf{c}$ be the center of the simplex, let $\mathbf{i}$ be the vertex nearest to $\mathbf{p}$, let $\mathbf{m}$ be the mid-point of the side nearest $\mathbf{p}$ and let $\mathbf{m}'$ be the midpoint of the side farthest from $\mathbf{p}$ (ties don't matter). Let $\mathbf{v}$ be the intersection of line $\overline{\mathbf{pm}'}$ with line $\overline{\mathbf{cm}}$. Let $\mathbf{o}$ be the intersection of line $\overline{\mathbf{iv}}$ with the side of the simplex farthest from $\mathbf{p}$. Then assuming that credence in the refuted theory drops to 0 immediately, retraction-efficiency countenances moving anywhere on the line segment connecting $\mathbf{j}$ and $\mathbf{o}$. For retraction measure ρ_{dis}, the construction is the same except that $\mathbf{v}$ is the intersection of $\overline{\mathbf{cm}}$ with $\overline{\mathbf{pj}}$. Note that when $\mathbf{p} = \mathbf{i}$, the Ockham zone for ρ_{ret} is the entire half-side $\overline{\mathbf{jm}'}$, whereas measure ρ_{dis} allows only for movement directly to the corner $\mathbf{j}$, as is already required in the disjunctive theory choice setting described in section 10. Thus, the strong version of Ockham's razor is tied to the plausible aim of preserving as much content as possible. In practice, however, an open-minded Bayesian never puts *full* credence in the currently simplest theory

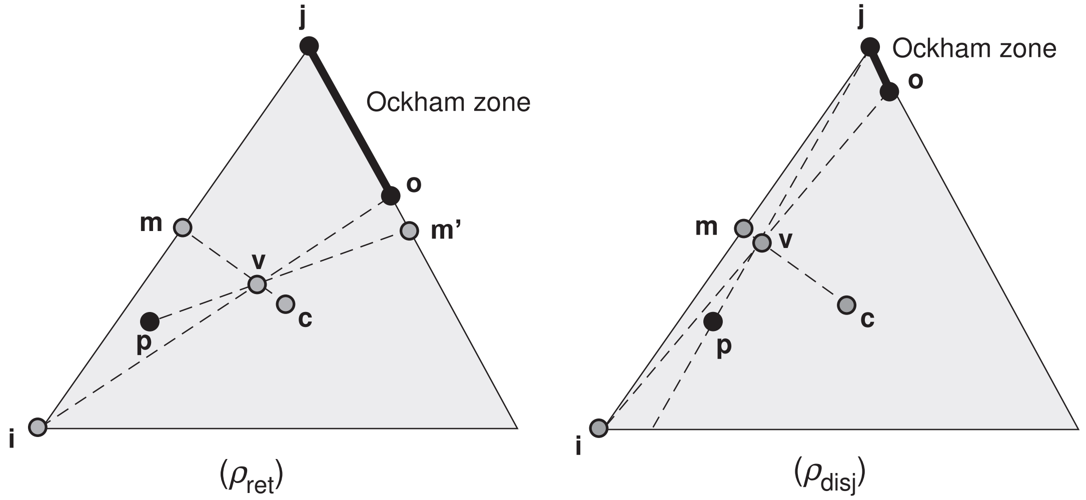

Figure 20: Two versions of Ockham's Bayesian razor

and in that case the Ockham zone for ρ_{ret} allows some leeway but is still more stringent than the zone for ρ_{ret}.

The Gettier-like counterexample presented in section 10 can also arise in 4 dimensions or more for Bayesian agents when the no short path assumption fails (just embed the example into the upper faces of the 4-dimensional unit cube and project it down onto the 3-dimensional simplex contained in that cube). The potential responses reviewed in section 10 apply here as well.

12 Conclusion

This study reviewed the major justifications of Ockham's razor in philosophy, statistics, and machine learning, and found that they fail to explain, in a non-circular manner, how Ockham's razor is more conducive to finding true theories than alternative methods would be. The failure of standard approaches to connect simplicity with theoretical truth was traced to the concepts of truth-conduciveness underlying the respective arguments. Reliable indication of the truth is too strong to establish without (a) trading empirical truth for accurate prediction or (b) begging the question by means of a prior bias against complex possibilities. Convergence in the limit is too weak to single out simplicity as the right bias to have in the short run. An intermediate concept of truth-conduciveness is effective pursuit of the truth, where effectiveness is measured in terms of such costs as total retractions and errors prior to convergence. Then one can prove, without circularity or substituting predictive accuracy for theoretical truth, that Ockham's razor is the best possible strategy for finding true theories.

That result, called the Ockham efficiency theorem, can be extended to problems with branching paths of simplicity, to problems in which defeated theories are not refuted, to random strategies and, except in some interesting, Gettier-like cases, to Bayesian degrees of belief and to strategies that produce disjunctions of theories. The ultimate goal, which has not yet been reached, is to extend the Ockham efficiency argument to statistical inference.

13 Acknowledgements

The results on random strategies were obtained in close collaboration with Conor Mayo-Wilson, who also provided detailed comments on the draft. The definitions concerning Bayesian retractions were arrived at in close collaboration with Hanti Lin, who also formulated and proved many of the theorems. Unusually detailed comments were provided by the editor Prasanta Bandyopadhyay and by the anonymous reviewer. They were greatly appreciated.

14 References

Akaike, H. (1973) “A new look at the statistical model identification”, IEEE Transactions on Automatic Control 19: 716-723.

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

J. Case and Smith, C. (1983) “Comparison of identification criteria for machien inductive inference”, *Theoretical Computer Science* 25: 193-220.

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

Domingos, P. (1999) “The Role of Occam’s Razor in Knowledge Discovery,” *Data Mining and Knowledge Discovery*, vol. 3: 409-425.

Duda, R., Hart, P. and Stork, D. (2001) *Pattern Classification*, New York: Wiley.

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

Forster, M. (2001) “The New Science of Simplicity”, in *Simplicity, Inference, and Modeling*, A. Zellner, H. Keuzenkamp, and M. McAleer, eds., 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 Space-time Theories*, Princeton: Princeton University Press.

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

Gettier, E. (1963) “Is Justified True Belief Knowledge?”, *Analysis* 23: 121-123.

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

Glymour, C. (2001) “Instrumental Probability”, *Monist* 84: 284-300.

Goldenshluger, A. and Greenschtein, E. (2000) “Asymptotically minimax regret procedures in regression model selection and the magnitude of the dimension penalty”, *Annals of Statistics*, 28: 1620-1637.

Grünewald, P. (2007) *The Minimum Description Length Principle*, Cambridge, M.I.T. Press.

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

Heise, D. (1975) *Causal Analysis*, New York: John Wiley and Sons.

Hjorth, J. (1994) *Computer Intensive Statistical Methods: Validation, Model Selection, and Bootstrap*, London: Chapman and Hall.

Jeffreys, H. (1961) *Theory of Probability*, 3rd ed., London: Oxford University Press.

Joyce, J. (1998) “A Nonpragmatic Vindication of Probabilism”, *Philosophy of Science* 65: 73-81.

Kass, R. and Raftery, A. (1995), “Bayes Factors”, *Journal of the American Statistical Association* 90: 773-795.

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

Kelly, K. (2007)“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. (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 Mayo-Wilson, C. (2009) “Ockham Efficiency Theorem for Random Empirical Methods”, Formal Epistemology Workshop 2009, http://fitelson.org/few/kelly_mayo-w

Kelly, K. and Schulte, O. (1995) “The Computable Testability of Theories with Uncomputable Predictions”, *Erkenntnis* 43: 29-66.

Kitcher, P. (1982) “Explanatory Unification”, *Philosophy of Science* 48:507-531.

Kyburg, H. (1977) “Randomness and the Right Reference Class”, *The Journal of Philosophy*, 74: 501-521.

Kuhn, T. (1957) *The Copernican Revolution*, Cambridge: Harvard University Press.

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

Lehrer, K. (1990) *Theory of Knowledge*, Boulder: Westview Press.

Levi, I. (1974) “On Indeterminate Probabilities”, *Journal of Philosophy* 71: 397-418.

Levi, I. (1983) *The Enterprise of Knowledge: An Essay on Knowledge, Credal Probability, and Chance*, Cambridge: M.I.T. Press.

Li, M. and Vitanyi, P. (1993) *An Introduction to Kolmogorov Complexity and its Applications*, New York: Springer.

Luo W. and Schulte, O. (2006) “Mind change efficient learning”, *Information and Computation* 204:989-1011.

Mallows, C. (1973) “Some comments on Cp”, *Technometrics* 15: 661-675.

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

Meek, C. (1995) “Strong completeness and faithfulness in Bayesian networks,” *Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence, Montreal*, pp. 411-418.

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

Myrvold, W. (2003) "A Bayesian Account of the Virtue of Unification," *Philosophy of Science*:399-423.

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

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

Rissanen, J. (2007) *Information and Complexity in Statistical Modeling*, New York: Springer-Verlag.

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

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

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

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

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

Teller, P. (1976) "Conditionalization, Observation, and Change of Preference", in *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, W. Harper and C. Hooker, eds., Dordrecht: D. Reidel.

U.S. Army (2003) *Rifle Marksmanship M16A1, M16A2/3, M16A4, and M4 Carbine*, FM 3-22.9, Headquarters, Dept. of the Army.

Vapnik, V. (1995) *The Nature of Statistical Learning Theory*, Berlin: Springer.

Verma, T. and Pearl, J. (1991) "Equivalence and Synthesis of Causal Models", *Uncertainty in Artificial Intelligence* 6:220-227.

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

Wolpert D. (1996) "The lack of a prior distinction between learning algorithms," *Neural Computation*, 8: pp. 1341-1390.

Zadeh, L. (1965) "Fuzzy sets", *information and Control* 8: 338-353.