Testability and Ockham's Razor:

How Formal and Statistical Learning Theory Converge in the New Riddle of Induction

Daniel Steel

Department of Philosophy

503 S. Kedzie Hall

Michigan State University

East Lansing, MI 48823-1032

Email: steel@msu.edu

Abstract

Nelson Goodman's new riddle of induction forcefully illustrates a basic challenge that must be confronted by any adequate theory of inductive inference: provide some basis for choosing among alternative hypotheses that fit past data but make divergent predictions. One response to this challenge is to appeal to some version of Ockham's razor, according to which simpler hypotheses should be preferred over more complex ones. Statistical learning theory takes this approach by showing how a concept similar to Popper's notion of degrees of testability is linked to minimizing expected predictive error. In contrast, formal learning theory explains Ockham's razor by reference to the goal of efficient convergence to the truth, where efficiency is understood as minimizing the maximum number of retractions of conjecture or "mind changes." In this essay, I show that, despite their differences, statistical and formal learning theory yield precisely the same result for a class of inductive problems that I call *strongly VC ordered*, of which Goodman's riddle is just one example.

1. Introduction

Nelson Goodman's widely discussed new riddle of induction[1] is based on a simple device for generating alternative hypotheses that fit past data perfectly well but which have very different implications for the future. In Goodman's most famous example, observations of green emeralds fit the hypothesis that all emeralds are green but also the hypothesis that all emeralds are grue, where an object is grue just in case it is green and observed before some future date *t* or blue and not observed before *t*. The new riddle of induction illustrates a challenge that must be confronted by any adequate theory of inductive inference: provide some principled grounds for selecting among alternative hypotheses that fit the data equally well but which make divergent forecasts. One response to this challenge is appeal to some version of Ockham's razor, according to which simpler hypotheses should be preferred over more complex ones. I examine two very different approaches to inductive inference—statistical and formal learning theory—that provide precise accounts of Ockham's razor, and I show that they give identical solutions in a broad range of inductive problems that includes Goodman's riddle.

Statistical learning theory recommends that, in selecting a hypothesis, a balance be struck between fit with past data and something known as Vapnik-Chervonenkis (VC) dimension (Vapnik 2000). The aim of statistical learning theory is to select hypotheses in such a way as to minimize the expected error of predictions about the next batch of observations. From this perspective, the preference for simpler hypotheses is explained by a preference for lower VC dimension, which is advantageous insofar as it makes one less prone to over-fit random noise in the data. The aim of accurate predictions should be

[1] See Goodman (1946, 1954). There are two edited volumes of articles dedicated to Goodman's riddle (Stalker 1994, Elgin 1997).

carefully distinguished from the goal of discovering the true hypothesis. For example, the desire to minimize expected predictive error can lead to selecting a simple hypothesis even when it is known that the truth is more complex (Vapnik 2000, 116). In contrast, formal learning theory explains Ockham's razor as a means for converging to the truth as efficiently as possible, where efficiency is understood as minimizing the maximum number of "mind changes" or retractions of conjectures (cf. Kelly 2004, 2007a, 2007b). Since the central aim of formal learning theory is to learn the true hypothesis while avoiding unnecessary sidetracks and U-turns, it would not recommend choosing a hypothesis known to be false over one that might be true. In this essay, I show that, despite their manifest differences, statistical and formal learning theory surprisingly yield precisely the same result in a class of inductive problems that I call *strongly VC ordered*, of which Goodman's riddle is just one example.

In a book on implications of statistical learning theory for philosophical issues, Gilbert Harman and Sanjeev Kulkarni suggest that a preference for lower VC dimension might explain the intuitive judgment that one should conjecture green before grue in Goodman's riddle (2007, 65-69). As they point out, however, the standard procedure in statistical learning theory is to presume that the same probability distribution generates both past and future data. But this is an example of exactly the sort of uniformity principle that Goodman's riddle aims to show is uninformative unless some predicates are privileged over others. "Is it the probability of *green* or the probability of *grue* that remains constant?" Goodman would have asked. Aside from the special case in which the chance of green is always .5, if the probability of green is constant then probability of grue is not, and vice versa. Thus, an approach that presumes that past and future data are generated by the same probability distribution does not seem a promising line of response to Goodman's riddle.

Yet the concept of VC dimension can be directly applied to Goodman's riddle without any assumptions about identically distributed outcomes, or indeed any mention of probability whatever. I explain how this is so, and then ask whether there is some reason to prefer sets of hypotheses with lower VC dimension in the context of Goodman's riddle. The conceptual similarity between VC dimension and Karl Popper's (1959) notion of testability is very suggestive at this point. Popper claimed that testable hypotheses are to be preferred because they are a more efficient means for advancing scientific knowledge. In this essay, I show that VC dimension and efficiency in the sense of minimizing the maximum number of retractions are closely linked in a generalized version of Goodman's riddle. In particular, if hypotheses are grouped according to the number of color switches that they predict, then the VC dimension of a set of hypotheses equals the number of color switches predicted by the hypotheses in that set. In this context, Ockham's razor recommends that, among those hypotheses consistent with the data, one always conjecture the hypothesis from the set with the lowest VC dimension. For example, if all emeralds observed so far are green, then Ockham's razor conjectures that all are green. In the generalized version of Goodman's riddle, Ockham's razor minimizes the maximum number of mind changes and any logically reliable method incurs one additional mind change to the maximum possible whenever it conjectures a hypothesis not from the set with lowest VC dimension consistent with the data. Moreover, I prove that the generalized Goodman's riddle is only one example of a broader class of cases. The connection between VC dimension and minimizing the maximum number of mind changes can be extended to a class of inductive problems that are what I call *strongly VC ordered*. The convergence of statistical and formal learning theory is a surprising and potentially important clue to the nature of inductive inference.

2. Goodman's Riddle and VC Dimension

Goodman originally presented his riddle of induction by means of the following example:

> Suppose we had drawn a marble from a certain bowl on each of the ninety-nine days up to and including VE day, and each marble drawn was red. We would expect that the marble drawn on the following day would also be red. So far all is well. … But increase of credibility, projection, "confirmation" in any intuitive sense, does not occur in the case of every predicate under similar circumstances. Let "S" be the predicate "is drawn by VE day and is red, or is drawn later and is non-red." (1946, 383)

Although this example is less well-known than the case of green versus grue emeralds, it does a better job of highlighting the fundamental problem. In the emerald example, our background knowledge may include information that makes it very unlikely that all emeralds are grue.[2] For example, if all emeralds are grue, then the process by which they are dug out of the earth must somehow selectively extract only green emeralds before *t* and only blue ones after that date. Yet as Peter Godfrey-Smith (2003) points out, our knowledge of the actual processes by which emeralds are excavated might render such a possibility highly unlikely. In Goodman's example of drawing balls from a bowl, however, no information is provided about the process by which balls are selected. For

[2] Indeed, as some authors have noted (cf. Thomson 1966), given that an emerald is defined in mineralogy as a green beryl, there is a strong case for judging "all emeralds are green" to be an analytic truth.

all we know, the sampling process might be biased towards selecting red balls before VE day and blue ones after that. Thus, in this case we must directly confront the question of whether there is any basis for preferences among alternatives that fit the data equally well in the absence of prior information that makes some more probable than others.

In what follows, an inductive problem $\mathcal{P}$ will be defined by a set of possible sequences of data, a set of alternative hypotheses, and the data so far. For example, the outcomes in Goodman's example of drawing balls from an urn can be represented by sequences of 1s and 0s, where 1 corresponds to red and 0 to blue. Let Γ denote the set of all infinite sequences of 1s and 0s. Thus, Γ represents the set of all possible sequences of data in Goodman's ball drawing example, were it to be continued indefinitely. One data sequence, which we may call the *actual sequence*, will in fact occur, and the aim is to discover the hypothesis that is true of it. The *data so far* consist of a finite, initial segment of the actual sequence. In Goodman's riddle, the data so far are ninety-nine red balls. Thus, whatever the actual sequence is in Goodman's riddle, it must begin with ninety-nine 1s. Let d denote the data so far. Further observations extend the data so far by concatenating additional segments of 1s or 0s to the end of d. Let $d*m$ denote the concatenation of d and an m-place segment of 1s and 0s. For example, $d*5$ could be the ninety-nine 1s with the five-place segment 11001 tacked on to the end. The set of possible data sequences in an inductive problem need not always be Γ. For instance, if our background knowledge rules out some sequences, it could be a proper subset of Γ. Consider an inductive problem $\mathcal{P}$ whose set of possible data sequences is Ω. Then an *extension of the data* in $\mathcal{P}$ is defined as follows: the data so far are an extension of the data, and if e is an extension of the data and $e*m$ is an initial segment of a sequence in Ω, then $e*m$ is also an extension of the data. Nothing else is an extension of the data. Notice that this characterizes the set of extensions of the data that are *possible*. The actual extension of the data is one of many extensions. An *extension of length m* is the concatenation of an extension of the data e and an $m \geq 0$ long segment of data units. Finally, the *n^{th} extension of e* will refer to the observation occurring n places after the last observation in e.

The problem in Goodman's riddle is to provide some basis for choosing among alternative hypotheses that perfectly fit the data so far but which make diverging predictions about what the extension of the data will be. One such hypothesis is that all of the balls are red. Goodman suggests another alternative according to which the balls drawn before VE day are red and all those drawn afterwards are blue. Presumably, Goodman does not intend that there is anything special about VE day, since otherwise his riddle could be definitively resolved simply by waiting until that day to see if the color switch occurred. Clearly, Goodman's point is that *no matter how many red balls are observed*, there will always be an alternative that fits the data yet predicts that the next ball is blue. Thus, it is natural to interpret Goodman's riddle as including a set of infinitely many alternative hypotheses of the form: all are rue$_n$, where an object is rue$_n$ if it occurs before the n^{th} extension of the data so far and is red, or does not occur before the n^{th} extension of the data so far and is blue. For any $n \geq 1$, the hypothesis that all of the balls are rue$_n$ fits the data so far in Goodman's riddle. For ease of expression, I will use "all are rue" as a shorthand for "there is an n such that all are rue$_n$." To ensure that the set of alternative hypotheses is exhaustive, I include an alternative that says "none of the above," which in this case means neither all red nor all rue. Neither all red nor all rue corresponds to the set of all and only those data sequences that contain a segment in which a 0 is followed by a 1. To see this, note that if all of the balls are red, then the actual sequence is an endless sequence of 1s, whereas if all are rue, the actual sequence consists of a finite initial segment of 1s followed by and endless series of 0s. In neither of these two cases can a 0 occur before a 1. Goodman's riddle, therefore, can be construed as the inductive problem in which the set data sequences is Γ, the set of hypotheses is {all are red; all are rue$_1$; all are rue$_2$; all are rue$_3$; …; neither all red nor all rue}, and the data so far are a string of ninety-nine red balls.

Is there any reason for favoring the hypothesis that all of the balls are red over every rue alternative, for example, the alternative that all of the balls are rue$_{100}$? One idea is that among hypotheses that fit the data equally well, those that posses desirable characteristics such as simplicity or testability should be preferred. The concept of VC dimension developed in statistical learning theory, which bears some important similarities to Popper's notion of degrees of testability, is a precisely defined concept that is applicable in this case. VC dimension is defined by means of the concept of *shattering*. For cases such as Goodman's riddle in which measurement error is negligible, shattering can be defined in the following way. A set of hypotheses S *shatters* an extension of the data of length *m* if and only if S is consistent with every extension of the data of length *m*. A set of hypotheses S is *consistent with the data* just in case there is a hypothesis in S that is consistent with those data. For example, in Goodman's riddle, there are two possibilities for the next observation: it could be 1 or it could be 0. Thus, {all are red} does not shatter an extension of the data of length one, since all red is not consistent with the next datum being a 0. In contrast, {*n*: all are rue$_n$} does shatter an extension of length one, because there will be a hypothesis in this set consistent with the next observation whether it is 1 or 0. However, {n: all are rue$_n$} does not shatter an extension of length two, since all are rue is not consistent with the next two observations being 01.

The VC dimension of a set of hypotheses S given the data so far, then, is the length of the longest extension of the data that is shattered by S. In what follows, "the VC dimension of S given the data so far" will be abbreviated to "the VC dimension of S." In Goodman's ball drawing example, the VC dimension of {all are red} is zero, while the VC dimension of {n: all are rue$_n$} is 1. Suppose that we considered hypotheses that allowed more than one color switch, for example, red-blue-red, red-blue-red-blue, red-blue-red-blue-red, and so on. The set of red-blue-red hypotheses has a VC dimension of 2. Whether the next two observations are 11, 10, 01, or 00, we can find a hypothesis in the red-blue-red set consistent with the data, but there is no red-blue-red hypothesis consistent with 010. By similar reasoning, the VC dimension of the set of red-blue-red-blue hypotheses is 3; the VC dimension of the set of red-blue-red-blue-red hypotheses is 4, and so on. In general, if we group hypotheses in Goodman's riddle according to the number of color switches, then the VC dimension of any set equals the number color switches predicted by the hypotheses it contains.

One interesting feature of VC dimension is its similarity to Popper's notion of degrees of testability. The lower the VC dimension of a set of hypotheses, the smaller the number of observations capable of refuting the claim that the true hypothesis is a member of that set. For example, a single observation can prove that not all are red, two observations can show that not all are rue, three observations can refute the claim that the true hypothesis is in the red-blue-red set, and so on. Indeed, VC dimension bears a striking resemblance to one of Popper's proposals about how to make the concept of degrees of testability more exact, namely, his notion of the *dimension* of a theory (1959, 97-98). Roughly, if the minimum number of basic observation sentences capable of refuting a theory t is $(d + 1)$, then the dimension of t is d. However, Popper speaks of the dimension of a theory, whereas VC dimension is a property of sets of hypotheses. For example, the VC dimension of $\{n$: all are rue$_n\}$ is 1, but the same is not true of the individual hypothesis that all are rue$_{100}$. But if we read Popper's "theory" as "set of hypotheses," then VC and Popper dimension are equivalent.[3]

A basic result of statistical learning theory is that the probability of fit with future data can, under some fairly general circumstances, be shown to involve a tradeoff between fit with past data and lower VC dimension. However, this result assumes that the probability distribution generating the data is independent and identically distributed. Independent means that the outcome of one observation makes no difference to the probabilities of earlier or later outcomes. Identically distributed data means that the probabilities of the possible outcomes are the same for each observation, past, present, or future. For example, a biased coin might have a .75 probability of coming up heads each time it is flipped. Identical distribution is an example of the type of uniformity of nature assumption that Hume famously called into question. In contrast to Hume, Goodman pointed out that even if we grant that nature is uniform, that assumption is uninformative

[3] Cornfield et al. (2005) suggest that VC dimension differs from Popper's concept of dimension in that the concept of VC dimension, unlike Popper's notion, is designed for cases in which the data consists of inexact measurements of real valued variables, such as weight or height. This strikes me not as a difference in the concepts of dimension, but rather a difference in what is regarded as a possible extension of the data in a given inductive problem. If the data consists of inexact measurements of two real valued variables, then each datum consists of pairs of intervals of real numbers, not points that may be perfectly co-linear. Clearly, the concept of shattering should be defined with respect to possible extensions of the data only.

without a privileged mode of description. In the example of drawing balls from a bowl, identical distribution could be taken to mean that the chance of red is the same for every observation, or that the chance of rue_{100} is constant, or that the chance of rue_{250} is. To assume that identically distributed data means that the probability of red is constant is to beg precisely the question being posed in Goodman's riddle.[4] But as the discussion above showed, assumptions about probability distributions are not needed to apply the concept of VC dimension to Goodman's riddle. Moreover, the conceptual similarity between VC dimension and Popper's notion of degrees of testability are very suggestive. One of the central themes of Popper's philosophy of science is that scientists should prefer highly testable hypotheses because they further scientific progress by generating precise predictions that put the hypothesis at greater risk of refutation. For Popper, the reason for preferring testable hypotheses had nothing to do with probability and everything to do with enhancing the efficiency of scientific inquiry. Thus, one could approach Goodman's riddle by specifying a pertinent sense of efficiency and demonstrating that a preference for lower VC dimension enhances efficiency in that sense. That is what the next section does.

3. Logical Reliability and Efficiency in Goodman's Riddle

The approach to Goodman's riddle found in the literature on formal learning theory shows that, when there is a grue hypothesis for each future date, efficient convergence to truth favors conjecturing all green before any of the grue alternatives (Schulte 1999b,

[4] Harman and Kulkarni suggest an alternative formulation of Goodman's riddle in which the grue hypotheses assert that all objects of particular masses are green while those of other masses are blue (2007, pp. 66). But in this version of the riddle it is presumed that the data are identically distributed with regard to green and mass, rather than with regard to grue and some "Goodmanized" version of mass. So, this still begs the question raised by Goodman's riddle.

1999a). In this section, I explain how this result can be reformulated in terms of VC dimension and extended to a more general case in which the number of possible color switches can be as high as one likes.

Consider a version of Goodman's riddle in which universal generalizations predicting an arbitrary, though finite, number of color switches are considered. As explained in the previous section, it is possible to group universal generalizations in this extended version of Goodman's riddle according to the number of color switches: zero for all are red; one for all are rue; two for red-blue-red, and so on. Finally, to ensure that the all of the possibilities have been covered, we need a hypothesis that says "none of the above." This gives us the following:

S_0: {all are red}

S_1: {n: all are rue$_n$}

S_2: {n, m: all are red-blue$_n$-red$_{n+m}$}

S_3: {n, m, l: all are red-blue$_n$-red$_{n+m}$-blue$_{n+m+l}$}

⋮

S_k

N = {there are more than k color switches}

The subscripts attached to "blue" and "red" indicate the observation at which the switch to that color occurs. For example, all are red-blue$_n$-red$_{n+m}$ says that the first blue ball occurs at the n^{th} observation while the next red ball occurs at the n+m^{th} observation, after which point only red balls are observed. Finally, the last set is labeled N for "none of the above," and contains only one member asserting that there are more than k color switches and hence that all of the other hypotheses are false. This set is included to ensure that the hypotheses are collectively exhaustive as well as mutually exclusive. If the hypotheses in Goodman's riddle are grouped in this way, then the VC dimension of set S_i equals *i*. Thus, the VC dimension of S_0 equals 0, the VC dimension of S_1 equals 1, and so on. The VC dimension of N is infinite, because it is consistent with any finite segment of observations.

I will call the inductive problem that considers this extended set of alternatives the *generalized Goodman's riddle*. More specifically, in the generalized Goodman's riddle, Γ (the set of all infinite sequences of 1s and 0s) is the set of possible data sequences, the partition of hypotheses is $\{S_0, \ldots, S_k, N\}$, and the data so far consist of an unbroken segment of 1s. What one might call the Humean inductive problem is the special case of the generalized Goodman's riddle in which S_0 = {all are red} and N = {not all are red}. What Oliver Schulte (1999a) calls the infinitely iterated new riddle of induction is the special case in which S_0 = {all are red}, S_1 = {*n*: all are rue$_n$}, and N = {neither all red nor all rue}.

A few additional concepts are needed at this point. An *inductive method* is a procedure for indicating hypotheses from a set of alternatives on the basis of a finite segment of data. I will say that a method *uniquely indicates* a hypothesis h if it indicates h and no other hypothesis. For the present purposes, only two modes of indication will be considered: concluding and conjecturing. To conclude is to indicate while issuing a definitive pronouncement that the indicated hypothesis is true, and hence that no further data are needed to answer the question. In contrast, conjecturing is a more tentative mode of indication not involving any such definitive pronouncement. A distinctive feature of inductive inference problems is that there is no method assured of correctly concluding the true hypothesis. For example, no matter how many red balls are drawn from the bowl, it possible that the next will be blue. Nevertheless, the method may be assured of permanently indicating the true hypothesis eventually even if it never concludes that the truth has been discovered. The notion of logical reliability is a way of making this idea precise. A method is *logically reliable* with regard to a set of data sequences and a partition of hypotheses if and only if, for any data sequence under consideration there is an *n* such that the method uniquely indicates the hypothesis true of that sequence by the n^{th} observation and does not change its indication after that. In other words, a logically reliable method is assured of eventually settling on the true hypothesis, but it might not issue any definitive pronouncement or sign when this happens.

In the context of Goodman's riddle, Ockham's razor would naturally be interpreted as asserting that we should not postulate more color switches than needed to account for the data. That idea can be formulated in terms of VC dimension as follows:

> Ockham's razor: Conjecture the hypothesis consistent with the data from the S_i with lowest VC dimension; if no S_i contains a hypothesis consistent with the data, conclude that the true hypothesis is in N.

Thus, Ockham's razor will conjecture all are red so long as all of the balls observed so far are red. It will conjecture that all are rue_n if the first blue ball was observed at *n* and no further red balls have been observed. In sum, in the generalized Goodman's riddle, if *i* color switches have been observed so far, Ockham's razor finds the hypothesis in S_i consistent with the data and conjectures that it is true. If the number of color switches observed in the data is greater than *k*, Ockham's razor concludes that the "none of the above" hypothesis in N is true.

It is easy to show that Ockham's razor is logically reliable in the generalized Goodman's riddle. First, notice that if S_i is the set with lowest VC dimension consistent with the data, then there is exactly one hypothesis in S_i consistent with the data. If S_i is the set with lowest VC dimension consistent with the data, then exactly *i* color switches have been observed in the data so far. Since S_i contains all possible hypotheses that predict exactly *i* color switches, it must contain at least one hypothesis consistent with the data. But since no two hypotheses predict the same color switches and no hypothesis in S_i predicts more than *i* color switches, there can be no more than one hypothesis is S_i consistent with the data. Suppose, then, that, for some *i*, h in S_i is the true hypothesis. Then *i* color switches will eventually be observed in the data, at which point S_i will be the set with lowest VC dimension consistent with the data, and h the sole member of S_i consistent with the data. Then Ockham's razor conjectures h, and since h is true, never changes its conjecture after that. On the other hand, suppose that no S_i contains the true hypothesis. Then $k + 1$ color switches will eventually be observed in the data, at which point Ockham's razor correctly concludes N.

Any logically reliable method will share three important features with Ockham's razor. First, for each universal generalization consistent with the data so far, there must be a finite number of further observations that suffices for it to be conjectured. Any method that did not have this feature would not be logically reliable, because there would be a hypothesis that it would never permanently indicate even if it were true. Secondly, if none of the universal generalizations are consistent with the data, it must permanently indicate "none of the above." Finally, any logically reliable method must eventually indicate *only one* of the hypotheses consistent with the data. To see this, consider a method that always conjectured *all* of the hypotheses consistent with the data. This method satisfies the above two requirements, but it is not logically reliable. For example, if all the balls are red, then there will always be infinitely many rue alternatives consistent with the data. Thus, a method that conjectures all hypotheses consistent with the data will never settle on a single alternative in this case and hence is not guaranteed of eventually uniquely indicating the true hypothesis.

However, there are many methods that satisfy these conditions. In particular, logical reliability is consistent with conjecturing a hypothesis not from the set with lowest VC dimension consistent with the data. For example, one could begin by conjecturing that all are rue_{100}, and revert to Ockham's razor if the first 100 balls are all red. Although logical reliability is consistent conjecturing hypotheses in various orders, the order in which hypotheses are conjectured matters to the efficiency with which the method discovers the truth. One mark of an efficient method is that avoids unnecessary vacillations in which hypothesis is conjectured—an efficient method is one that, as it were, takes the most direct path to its destination (cf. Kelly 2004). Of course, which path is the most direct depends on which hypothesis is actually true: the quickest way to the truth is to believe the true hypothesis. However, this advice is obviously not very helpful if one does not know which alternative is true. But even when the truth is not known, methods may be shown to differ with regard to the *maximum* number of retractions of conjecture, or "mind changes," they can undergo.

In the generalized Goodman's riddle, no method has a lower maximum number of mind changes than Ockham's razor and furthermore each divergence from Ockham's razor adds one additional mind change to the maximum still possible. First, in the generalized Goodman's riddle the maximum number of mind changes of Ockham's razor is $k + 1$. That number of mind changes occurs when at least $k + 1$ color switches occur in the data. Secondly, no alternative method that is logically reliable has a smaller maximum number of mind changes. That is because, for any logically reliable method and any hypothesis h consistent with the data, there must be a finite number of further observations that suffice to make the method conjecture h. Thus, just as Ockham's razor, any logically reliable method can be forced by the data to conjecture at least one hypothesis from each S_i, and then to conclude N. Finally, whenever a logically reliable method conjectures a hypothesis not drawn from the set with lowest VC dimension, it increases its maximum number of further mind changes by one. For consider any point at which the method conjectures a hypothesis that is not drawn from the set with lowest VC dimension consistent from the data. Then the method conjectures the occurrence of a color switch in addition to those already observed. But if the method is logically reliable, it must, given a sufficiently large number of further observations without any additional color switches, eventually change its conjecture to the hypothesis from the set with lowest VC dimension consistent with the data. Yet further data may refute this last conjecture and force the method to conjecture again from S_i. Thus, while in the worst case Ockham's razor makes exactly one conjecture from each remaining S_i, a logically reliable method that skips ahead and conjectures a hypothesis from a S_i with VC dimension higher than necessary can be made to conjecture at least twice from S_i and at least once from the others. It may be helpful to encapsulate the results of this reasoning in the following proposition:

Proposition 1: In the generalized Goodman's riddle:

(a) The maximum number of mind changes of Ockham's razor is $k + 1$,

(b) There is no logically reliable method for which the maximum number of mind changes is strictly less than $k + 1$, and

(b) Whenever a logically reliable method conjectures a hypothesis *not* drawn from the set with lowest VC dimension given the data, its maximum number of further mind changes from that point is at least one greater than that of Ockham's razor.

In short, if efficiency is understood in terms of minimizing the maximum number of mind changes, then Ockham's razor is the most efficient route to the truth in the generalized Goodman's riddle.

4. Generalizing the Result

One might wonder whether the result in the foregoing section is an instance of a more general relationship between VC dimension and efficient convergence or whether it is merely one peculiar and isolated example. In this section, I show that proposition 1 is true of any inductive problem that is what I call *strongly VC ordered* and whose set of data sequences is Γ (the set of all infinite strings of 1s and 0s). Perhaps the best way to approach this result is by considering one more variation on the new riddle of induction.

Consider the following version of Goodman's riddle adapted from Israel Scheffler (1963). Let us say that an object is grue$_n$ just in case it is the n^{th} data point and is blue or is not the n^{th} data point and is green. Thus, the hypothesis that all emeralds are grue$_n$ would assert that every emerald is green except for the n^{th} one, which is blue. The set $\{n$: all are grue$_n\}$ obviously shatters an extension of the data of length one, since there is a hypothesis in the set consistent with the next emerald being either green or blue. However, {*n*: all are grue$_n$} does not shatter a data segment of length two, because it cannot accommodate two blue emeralds in succession. By similar reasoning, the set of hypotheses predicting two exceptions shatters a data segment of length two but not of length three. In general, if hypotheses are grouped by the number of exceptions, then VC dimension of a set of hypotheses equals the number of exceptions those hypotheses predict. For 0 through k, let E_i be the set of all and only those hypotheses that predict i exceptions. For example, {all are green} is E_0; {*n*: all are grue$_n$} is E_1, and so on. Finally, let $N = (E_0 \cup \ldots \cup E_k)^c$. Now consider an inductive problem with the set of data sequences Γ and the partition $\{E_0, \ldots, E_k; N\}$, which we may call *Scheffler's riddle*. The reasoning that demonstrated proposition 1 for the generalized Goodman's problem also applies here as well. The maximum number of mind changes for Ockham's razor in Scheffler's riddle is $k + 1$. This occurs when $k + 1$ or more exceptions are observed. And by the same reasoning as that given for the case of the generalized Goodman's riddle, any logically reliable method can be forced by the data to conjecture at least once from each E_i and then to conclude that N. Hence, no logically reliable method has a maximum number of mind changes strictly less than $k + 1$ in Scheffler's riddle. Moreover, whenever a logically reliable method conjectures a hypothesis not drawn from the set with lowest VC dimension consistent with the data, it increases its maximum number of mind changes from that point by one. Consider a point at which this logically reliable rule conjectures a hypothesis from a set with VC dimension higher than necessary. Thus, the method conjectures additional exceptions beyond those already observed in the data. But since it is logically reliable, it will have to eventually switch to conjecturing that there will be no more exceptions if a sufficiently long sequence of data without further exceptions is observed. At this point, the alternative method is in the same position as Ockham's razor while having made at least one more mind change.

Scheffler's riddle and the generalized Goodman's riddle clearly share a common structure. First, both are *VC ordered*, where that expression is defined as follows.

> VC Order: Let $\mathcal{P}$ be an inductive problem defined by the set of data sequences Ω, set of hypotheses Π, and data so far *d*. Then $\mathcal{P}$ is *VC ordered by* $\{C_0, \ldots, C_k, N\}$ if and only if $(C_0 \cup \ldots \cup C_k \cup N) = \Pi$, and, for every C_i in $\{C_0, \ldots, C_k\}$, the VC dimension of C_i given *d* equals *i*.

I will say that an inductive problem is *VC ordered* (full stop) if there is some $\{C_0, \ldots, C_k, N\}$ such that the problem is VC ordered by $\{C_0, \ldots, C_k, N\}$. As before, N contains a single statement asserting that no hypothesis in any of the C_i's is true. If our background knowledge entails that the true hypothesis is in $C_0 \cup \ldots \cup C_k$, N may be empty. In this case, proposition 1 and the generalization of it below hold as before except that $k + 1$ becomes k in (a) and (b).

Both the generalized Goodman's and Scheffler's riddle are VC ordered, but their similarities do not end there. They are also what I call *strongly VC ordered.*

> Strong VC Order: An inductive problem $\mathcal{P}$ is *strongly VC ordered by* $\{C_0, \ldots, C_k, N\}$ if and only if $\mathcal{P}$ is VC ordered by $\{C_0, \ldots, C_k, N\}$ and, for every C_i in $\{C_0, \ldots, C_k\}$, any extension *e* of the data in $\mathcal{P}$, and any $m \geq 0$, if C_i is the set with lowest VC dimension consistent with *e*, then C_{i+m} (if it exists) shatters $e{*}m$ but no further extension of *e*.

In other words, so long as the data are consistent with C_i, the VC dimension of C_{i+m} is m greater than that of C_i. To illustrate, consider the infinitely iterated new riddle of induction in which S_0 = {all are red}, S_1 = {n: all are rue$_n$}, and N = {neither all red nor all rue}. Consider an extension e of the data that is consistent with S_0. Thus, e must be an unbroken segment of 1s. Then strong VC order requires that S_0 shatter a further extension of e of only zero length and that S_1 shatter a further extension of only length one. Both of these conditions obtain. First, S_0 is consistent with e and hence is consistent with a further extension of length zero, but S_0 does not shatter an extension of length one because it is not consistent with the next datum being a blue ball. Secondly, S_1 shatters any further extension of e of length one, because S_1 is consistent with the next ball being either red or blue. However, S_1 does not shatter an extension of e of length two or longer, since S_1 would be refuted by a blue ball followed by a red one. Likewise, if S_1 is the set with lowest VC dimension consistent with the data, then it shatters an extension of only length zero, since it would be refuted if the next ball is red.

There may be more than one strong VC ordering of a single inductive problem. For instance, the infinitely iterated new riddle is strongly VC ordered by S_0 = {all are red}, S_1 = {n: all are rue$_n$}, but it is also strongly VC ordered by S_0 = {all are red}, S_1 = {all are red} $\cup$ {n: all are rue$_n$}. However, not every VC ordering of hypotheses in an inductive problem is a *strong* VC ordering. For example, suppose that the hypotheses in the infinitely iterated new riddle where grouped as follows: S_0 = {all are rue$_2$}, S_1 = {all are red} $\cup$ {n: all are rue$_n$}. This is a VC ordering, because {all are rue$_2$} shatters an extension of the data so far of only length zero while {all are red} $\cup$ {n: all are rue$_n$} shatters an extension of only length one. But it is not a strong VC ordering. Notice that if the next two observations are of a red then a blue ball, there is only one hypothesis in {all are red} ∪ {n: all are rue$_n$} consistent with the data, namely, all are rue$_2$. Thus, given this ordering, there is an extension *e* of the data such that S$_0$ is consistent with *e* but S$_1$ does not shatter an extension of *e* of length one.

Given the concept of strong VC order, we can generalize the results from the previous section.

Proposition 2: Let be $\mathcal{P}$ be an inductive problem that is strongly VC ordered by {C$_0$, …, C$_k$, N} and whose set of possible data sequences is Γ. Then Ockham's razor is logically reliable in $\mathcal{P}$.

The proof of proposition 2 is provided in the appendix. This result is helpful, since it shows that it is not necessary to include a qualification about the existence of a logically reliable method when generalizing proposition 1 to strongly VC ordered inductive problems whose set of data sequences is Γ. We can now proceed to the generalized version of proposition 1.

Proposition 3: Let be $\mathcal{P}$ be an inductive problem that is strongly VC ordered by {C$_0$, …, C$_k$, N} and whose set of possible data sequences is Γ. Then the following is true of $\mathcal{P}$:

(a) The maximum number of mind changes of Ockham's razor is $k + 1$,

(b) There is no logically reliable method for which the maximum number of mind changes is strictly less than $k + 1$, and

(c) Whenever a logically reliable method conjectures a hypothesis *not* drawn from the C$_i$ with lowest VC dimension given the data, its maximum number of further mind changes from that point is at least one greater than that of Ockham's razor.

The proof of proposition 3 is given in the appendix. The generalized Goodman's riddle, therefore, is only one example of a larger class of cases in there is a tight connection between VC dimension minimizing the maximum number of mind changes. Moreover, proposition 3(c) entails that an inductive problem cannot have distinct strong VC orderings that make a difference to what Ockham's razor conjectures. For otherwise there would be two methods for making conjectures from the data such that each had a strictly lower maximum number of further mind changes than the other, which is absurd. This corollary is given a more exact formulation and proof in the appendix.

5. Language Invariance

One of the original purposes of Goodman's riddle was to demonstrate that some theories of inductive inference—particularly, the "instance" model of confirmation (Hempel 1965)—were not language invariant. In other words, distinct results would be generated from the same inductive problem depending on which language was used to describe the hypotheses and observations. Language invariance seems like a very desirable feature for a theory of inductive inference to have: the appropriate inference in an inductive problem should not be affected, for example, by whether it is stated in English or French or Chinese. Yet the concept of VC dimension is language dependent insofar as it presupposes, without explanation, a natural or preferred way of carving the data up into units. In this section, I address this problem by proposing that the correct coding of the data is one that results in what I call a *maximal* strong VC ordering.

Although the VC dimension of a set cannot be altered by faithful translations of the hypotheses it contains, VC dimension does depend on what counts as an individual data unit. For example, in Goodman's riddle it is natural to assume that one ball is one datum, where 1 indicates a red ball and 0 a blue one. But we could take the basic of unit of data to be pairs of balls, coding 1 for two reds, 2 for red then blue, 3 for blue then red, and 4 for two blues. If an even number of balls has been observed, any description of the data in the original coding can be exactly translated into this alternative coding. Yet which coding we choose will make a difference to VC dimension in Goodman's riddle. Given the one-at-a-time coding the VC dimension of $\{n$: all are rue$_n\}$ is one, since it is consistent with the next observation being 1 or 0 but not with the next two observations being 01. But given the two-at-a-time coding the VC dimension of $\{n$: all are rue$_n\}$ is zero, since this set would be refuted if the next observation is 3 (i.e. blue then red). Thus, if we used the two-at-a-time coding, then the only strong VC ordering in Goodman's riddle would be one in which all $S_0 = \{$all are red$\} \cup \{n$: all are rue$_n\}$.[5] Notice that the two-at-a-time coding in Goodman's riddle is ruled out in the results in section 4 by the assumption that the possible data sequences were represented by strings of 1s and 0s. However, there appears to be no reason why the data must be coded solely in terms of 1s and 0s. Consequently, unless some reason is given to prefer the one-at-a-time to the two-at-a-time coding, the results in section 4 would seem to crucially depend on an arbitrary limitation of scope to inductive problems in which data are coded solely by 1s and 0s.

Applications of the concept of VC dimension, then, must rely on some implicit criterion for deciding what an acceptable coding of the data is. I propose that the

[5] I thank Greg Novack for this example. The problem of language variance with regard to units of data appears to have also been recognized by Popper in his discussion of theory dimension (1959, 283-284).

criterion is that the coding should be chosen so as to generate a *maximal* strong VC ordering. Suppose the inductive problem $\mathcal{P}$ is strongly VC ordered by $\{C_0, \ldots, C_k, N\}$. I will call this strong VC ordering *maximal* just in case there is no strong VC ordering of $\mathcal{P}$ by $\{S_0, \ldots, S_j, N\}$ such that $j > k$. In other words, a maximal strong VC ordering divides the hypotheses into as many sets as any strong VC ordering could. It is easy to see that the two-at-a-time coding in Goodman's riddle results in a *non*-maximal VC ordering, since it lumps sets of hypotheses together that would be separate in the one-at-a-time coding. A non-maximal VC ordering, then, is deficient because it obscures relevant distinctions among hypotheses that are inherent in the structure of the inductive problem.

Let us consider a further example to illustrate the concept of a maximal strong VC ordering. Suppose that the inductive problem is to learn how many instances of the triplet 101 occur in the data sequence, where that number must be finite. Hypotheses in this case specify both the number of occurrences of the triplet 101 and the places in the sequence at which they occur. For example, "no 101" would be the hypothesis that the triplet 101 never occurs. Likewise, "one 101_5" would be the hypothesis that the triplet 101 occurs exactly once and does so at the fifth observation. In this case, a maximal strong VC ordering would group the hypotheses like so: $S_0 = \{\text{no } 101\}$, $S_1 = \{n\text{: one } 101_n\}$, and so on. One way to do this would be to code the data in the following way: code any occurrence of the triplet (101) as 0 and any 1 or 0 that is not part of a 101 triplet as 1. Thus, the segment 11000101 would be coded as 111110. This strategy would also work for cases in which the "raw data" contain numerals besides 1 and 0. For example, suppose that any integer from 0 to 9 may occur in the data sequence and the inductive problem is to discover the number of occurrences of segments of the form (n, 0_1, …, 0_n,

n), where $n \neq 0$.[6] In this case, a maximal strong VC ordering results from coding any segment of the form (n, 0_1, …, 0_n, n) as 0, and any numeral that is not a part of such a segment as 1. Indeed, a similar coding works for Goodman's riddle. Call every red-then-blue pair and every blue-then-red pair a *color switch*. Then a maximal strong VC ordering of Goodman's riddle would result from coding each color switch as 0, and each ball that is not part of a color switch as 1.

Notice that the codings described in the previous paragraph result in an inductive problem that is identical to Scheffler's riddle discussed in section 4. Scheffler's riddle is a very simple version of what Kevin Kelly calls the "effect accounting problem" (2007a, 564), which is the type of inductive problem for which Ockham's razor is pertinent. The challenge in an effect accounting problem is to learn the number of effects (assumed to be finite) that will occur in a sequence of data when there is no limit on how late in the sequence an effect can occur. In Kelly's proposal (2004, 2007a, 2007b), an effect is a segment of data that reduces the number of retractions that a logically reliable method can be forced to make. For example, if n is the unknown number of color switches in Goodman's riddle, then any logically reliable method can be forced to make $n - 1$ additional mind changes after having observed one color switch. As Kelly explains, more scientifically interesting cases, such as fitting a polynomial equation to a scatter of inexact data, can also be construed as effect accounting problems. Kelly's approach is language invariant because effects are defined solely in terms of the number of forcible additional retractions for a logically reliable method and not in terms of any syntactical feature, such as the number of basic statements in a conjunction. This concept of an

[6] I thank Kevin Kelly for this example.

effect is linked to VC dimension in that one way to generate a maximal strong VC ordering in an effect accounting problem is to code effects as 0 and non-effects as 1.

There are two main points to take away from this discussion of effects and maximal strong VC orderings. First, it shows that restricting the results in section 4 to cases in which the data are represented as sequences of 1s and 0s is in fact far less of a limitation than might have originally appeared. That is because many distinct types of effect accounting problems can be recoded so as to be equivalent to Scheffler's riddle. Secondly, I conjecture that maximal strong VC orderings of an effect accounting problem in general track effects as defined by Kelly and thereby avoid the problem of language variance considered in this section. [7]

6. Conclusion

This essay has demonstrated an interesting yet hitherto unnoticed conceptual link between two very different approaches to inductive inference, namely, formal and statistical learning theory. These two theories are based on very different concepts and aims developed in response to distinct types of examples. Formal learning theory is built around the concepts of logical reliability and efficient convergence, and it typically addresses examples involving indefinitely long sequences of data that are assumed to be error-free. In contrast, statistical learning theory relies on the concept of VC-dimension and primarily treats cases involving noisy data, and its primary aim is to reduce the short term risk of predictive error. Nevertheless, this paper shows that VC dimension directly maps onto efficient convergence in a class of examples—which I term strongly VC

[7] There are, however, further issues relating to efficient convergence and language invariance in Goodman's riddle that are not discussed here (cf. Chart 2000; Schulte 2000).

ordered inductive problems—of which Goodman's riddle is just one example. These results suggest that further inquiry into the connections between VC dimension, efficient convergence, and minimization of expected predictive error may be a fruitful source of insights regarding the foundations of inductive inference.

Appendix

The following lemmas and definitions will be used for the proofs of propositions 2 and 3.

Lemma 1: Let be $\mathcal{P}$ be an inductive problem that is strongly VC ordered by {C_0, …, C_k, N} and whose set of possible data sequences is Γ. Then for every C_i in {C_0, …, C_k}, there is an extension e_i of the data of length *i* such that C_i is the set with lowest VC dimension consistent with e_i.

Proof: The proof proceeds by induction on *i*. The base case is trivial: since the VC dimension of C_0 is 0, C_0 is the set with lowest VC dimension given the data so far. Hence, the data so far is the extension e_0 of length 0 such that C_0 is the set with lowest VC dimension consistent with e_0. For the induction step, suppose that there is an extension e_i of the data of length *i* such that C_i is the set with lowest VC dimension consistent with e_i. Then since $\mathcal{P}$ is strongly VC ordered, C_i shatters an extension of e_i of only zero length, while C_{i+1} shatters an extension of e_i of length one. Thus, since $\mathcal{P}$ is strongly VC ordered by {C_0, …, C_k, N}, there is an extension of e_i of length one, call it e_{i+1}, in $\mathcal{P}$ such that C_{i+1} is the set with lowest VC dimension consistent with e_{i+1}. •

Definition: I will say that a hypothesis h is *i-specific* if and only if h is consistent with exactly one of the possible outcomes for any observation later than *i*. For example, consider a hypothesis that allowed that any datum up to and including the 10th could be either 0 or 1 but which predicts that every subsequent datum is 0. This hypothesis would be 10-specific.

Lemma 2: Let be $\mathcal{P}$ be an inductive problem that is strongly VC ordered by $\{C_0, \ldots, C_k, N\}$ and whose set of possible data sequences is Γ. Then for every C_i in $\{C_0, \ldots, C_k\}$, every hypothesis in C_i is i-specific.

Proof: By way of contradiction, suppose that C_i contains a hypothesis h that is not i-specific. That is, for some observation n greater than i, h is consistent with the observation at n being either 1 or 0. By lemma 1, there is an extension e_i of length i in $\mathcal{P}$ such that C_i is the set with lowest VC dimension consistent with e_i. And since Γ includes all sequences of 1s and 0s, there is an extension of e_i to $n - 1$, call it e_{n-1}, such that C_i is the set with lowest VC dimension consistent with e_{n-1}. Now, since h is in C_i and h is consistent with n being either 1 or 0, C shatters an extension of e_{n-1} of length one. But since $\mathcal{P}$ is strongly VC ordered, C_i does not shatter an extension of e_{n-1} of length one, which is a contradiction. •

Definition: Let h_1 and h_2 be two hypotheses consistent with an extension e of the data. I will say that h_1 and h_2 are *distinct* given e if and only if there is some future observation o such that h_1 and h_2 predict distinct values for o. Thus, I will say that there are at least two hypotheses in C_i consistent with e if and only if there is a pair h_1 and h_2 in C_i distinct given e. The intuitive idea here is that since hypotheses considered here only make claims about the sequence of observations, hypotheses consistent with the data so far are genuinely distinct only if they disagree about future observations.

Lemma 3: Let be $\mathcal{P}$ be an inductive problem that is strongly VC ordered by $\{C_0, \ldots, C_k, N\}$ and whose set of possible data sequences is Γ. Then for every C_i in $\{C_0, \ldots, C_k\}$and any extension *e* of the data in $\mathcal{P}$, if C_i is the set of lowest VC dimension consistent with *e*, then there is exactly one hypothesis in C_i consistent with *e*.

Proof: Let C_a be the set with lowest VC dimension consistent with *e*. Since C_a is consistent with *e*, there is at least one hypothesis in C_a consistent with *e*. By way of contradiction, suppose that there are at least two hypotheses in C_a consistent with *e*. For convenience, label these two hypotheses h_1 and h_2. Since h_1 and h_2 are distinct and the set of data sequences is Γ, there must be a first observation in the sequence, call it *o*, such that one entails that *o* is 1 while the other entails that *o* is 0. Since both h_1 and h_2 are consistent with *e*, *o* is later than the last observation in *e*. Consider, then, a further extension of *e*, *e*m*, such that *e*m* is consistent with h_1 and h_2, and the last observation in *e*m* immediately precedes *o*. Since Γ contains all possible infinite strings of 1s and 0s, there is an *e*m* in $\mathcal{P}$ that satisfies this condition. But now C_a shatters a further extension of *e*m* of length one, since it contains a hypothesis consistent with *o* being 1 and another consistent with *o* being 0. But this contradicts the hypothesis that $\mathcal{P}$ is strongly VC ordered, which entails that if C_a is the set with lowest VC dimension consistent with *e*m*, then C_a shatters a further extension of *e*m* of zero length only. •

Proposition 2: Let be $\mathcal{P}$ be an inductive problem that is strongly VC ordered by $\{C_0, \ldots, C_k, N\}$ and whose set of possible data sequences is Γ. Then Ockham's razor is logically reliable in $\mathcal{P}$.

Proof: The proof of proposition 2 is by cases. Suppose that no C_i in $\{C_0, \ldots, C_k\}$ contains the true hypothesis. Then we can show by induction on i that every C_i will eventually be inconsistent with the data. Since $\mathcal{P}$ is VC ordered, C_0 is the set with lowest VC dimension given the data d so far. Thus, by lemma 3, there is exactly one hypothesis, call it h_0, in C_0 consistent with d so far. By lemma 2, h_0 is 0-specific, which means that h_0 always makes a definite prediction (either 1 or 0) about the next observation. If h_0 is not true, then one of these predictions is false, and consequently C_0 is no longer consistent with the data. For the induction step, suppose that C_i is not consistent with the data. If C_{i+1} is consistent with the data, then it is the set with lowest VC dimension consistent with the data. Then by lemma 3, there is exactly one hypothesis, call it h_{i+1}, in C_{i+1} consistent with the data. By lemma 2, h_{i+1} is $(i+1)$-specific. But since the data is not consistent with C_i and C_i shatters an extension of the data of length i, the original data so far d must have been extended by at least $(i + 1)$ observations. Hence, h_{i+1} always makes a definite prediction (either 1 or 0) about each further observation. If C_{i+1} does not contain the true hypothesis, then one of these predictions is false, and consequently C_{i+1} is no longer consistent with the data. Thus, if no hypothesis in $C_0 \cup \ldots \cup C_k$ is true, Ockham's razor will eventually be driven by the data to correctly concluding N. For the second case, suppose that C_i is the set with lowest VC dimension in $\{C_0, \ldots, C_k\}$ that contains the true hypothesis. But then by the reasoning in the first case, every C_j, $j < i$, will eventually be inconsistent with the data. Thus, C_i will eventually be the set with lowest VC dimension consistent with the data. From lemma 3, there will be exactly one hypothesis, call it h_i, consistent with the data in C_i at this point. Thus, Ockham's razor conjectures h_i at this point, and since h_i is true does not change its conjecture thereafter. •

Lemma 4: Let be $\mathcal{P}$ be an inductive problem that is strongly VC ordered by $\{C_0, \ldots, C_k, N\}$ and whose set of possible data sequences is Γ and let the method M be logically reliable in $\mathcal{P}$. Then for every C_i in $\{C_0, \ldots, C_k\}$, if C_i is the set with lowest VC dimension consistent with the data e, then there is an extension of e, e^*m_i, such that, given e^*m_i, M conjectures the hypothesis in C_i consistent with e.

Proof: Suppose that M is logically reliable in $\mathcal{P}$ and that C_i is the set with lowest VC dimension consistent with the data e. From lemma 3, there is exactly one member of C_i, call it h, consistent with e. By way of contradiction, suppose that there is no extension of e, e^*m_i, such that M conjectures h given e^*m_i. Yet since Γ contains all infinite sequences of 1s and 0s, there is a sequence in Γ that has e as an initial segment and of which h is true. And because there is no extension of e, e^*m_i, such that M conjectures h given e^*m_i, M will never conjecture h even when h is true, which contradicts the hypothesis that M is logically reliable. •

Proposition 3: Let be $\mathcal{P}$ be an inductive problem that is strongly VC ordered by $\{C_0, \ldots, C_k, N\}$ and whose set of possible data sequences is Γ. Then the following is true of $\mathcal{P}$:

(a) The maximum number of mind changes of Ockham's razor is k +1,

(b) There is no logically reliable method for which the maximum number of mind changes is strictly less than $k + 1$, and

(c) Whenever a logically reliable method conjectures a hypothesis *not* drawn from the set with lowest VC dimension given the data, its maximum number of further mind changes from that point is at least one greater than that of Ockham's razor.

Proof (b): Consider a method M that is logically reliable in $\mathcal{P}$. The proof proceeds by first showing that, for each C_i, there is an extension of the data given which M conjectures at least once from each C_i. This part of the proof is by induction on *i*. For the base case, consider C_0. Since $\mathcal{P}$ is VC ordered, the VC dimension of C_0 is 0, and hence C_0 is the set with lowest VC dimension consistent with the data so far *d*. By lemma 4, there is an extension of *d*, call it $d*m_0$, given which M conjectures the hypothesis from C_0 consistent with *d*. For the induction step, suppose that C_i is the set with lowest VC dimension consistent with $d*m_0*\ldots*m_i$, and that, given $d*m_0*\ldots*m_i$, M conjectures a hypothesis in C_i. But since $\mathcal{P}$ is strongly VC ordered, there is a further extension of $d*m_0*\ldots*m_i$ of one additional observation, call it $d*m_0*\ldots*m_i*1$, such that C_{i+1} is the set with lowest VC dimension consistent with $d*m_0*\ldots*m_i*1$. From lemma 4, there is an extension of $d*m_0*\ldots*m_i*1$, call it $d*m_0*\ldots*m_i*m_{i+1}$, given which M conjectures the hypothesis from C_{i+1} consistent with $d*m_0*\ldots*m_i*1$. Moreover, the hypothesis from C_{i+1} is not a member of C_i, since otherwise C_i would be consistent with $d*m_0*\ldots*m_i*1$. Thus, there is an extension of the data that will make M conjecture at least once from each C_i, for a total of *k* mind changes. Now consider the extension of the data $d*m_0*\ldots*m_k$ given which M conjectures a hypothesis from C_k. Since $\mathcal{P}$ is strongly VC ordered, there is an extension of one additional observation, call it $d*m_0*\ldots*m_k*1$, such that $d*m_0*\ldots*m_k*1$ is inconsistent with C_k. Since M is logically reliable, it must, given $d*m_0*\ldots*m_k*1$, eventually switch to conjecturing a hypothesis from C_k to indicating N, which is mind change $k + 1$. •

Proof (a): From proposition 2, Ockham's razor is logically reliable in $\mathcal{P}$. Therefore, from the proof of proposition 1(b), the maximum number of mind changes of Ockham's razor is at least $k + 1$. And by lemma 3, if C_i is the set with lowest VC dimension consistent with the data, there is exactly one hypothesis in C_i consistent with the data. Thus, the natural projection conjectures no more than one hypothesis from each C_i, and consequently its maximum number of mind changes of Ockham's razor is no greater than $k + 1$. •

Proof (c): Let M be a method that is logically reliable in $\mathcal{P}$. Suppose that C_a is the set with lowest VC dimension given the extension of the data *e*. Suppose, then, that M conjectures a hypothesis not in C_a but instead in C_{a+n}. But by lemma 4, there is an extension $e*m_a$ such that M switches to conjecturing the hypothesis from C_a consistent with *e*. And by the reasoning in the proof of proposition 3(b), there is, for each C_i, $a < i \leq k$, an extension $e*m_a*\ldots*m_i$ such that M conjectures a hypothesis from C_i. Thus, by conjecturing a hypothesis from C_{a+n} before C_a, M can be made to make at least two conjectures from C_{a+n} and at least one from every other set C_a through C_k. In contrast, Ockham's razor makes exactly one conjecture from C_a through C_k in the worst case. Thus, whenever a logically reliable method conjectures a hypothesis not drawn from the set with lowest VC dimension consistent with the data, its maximum number of further mind changes from that point is at least one greater than that of Ockham's razor. •

Corollary: Let be $\mathcal{P}$ be an inductive problem whose set of possible data sequences is Γ and that is strongly VC ordered by both $\{C_0, \ldots, C_k, N\}$ and $\{D_0, \ldots, D_l, N\}$. Then for any extension *e* of the data, any $i \leq k$, and any $j \leq l$, if C_i and D_j, respectively, are the sets with lowest VC dimension in $\{C_0, \ldots, C_k\}$ and $\{D_0, \ldots, D_l\}$ consistent with *e*, then there is an h such that h is the only hypothesis in C_i consistent with *e* and the only hypothesis in D_j consistent with *e*.

Definition: Let Ockham's razor *on* $\{C_0, \ldots, C_k, N\}$ refer to Ockham's razor used in $\mathcal{P}$ when $\mathcal{P}$ is strongly VC ordered by $\{C_0, \ldots, C_k, N\}$.

Proof: Let *e* be an arbitrary extension of the data such that C_i and D_j, respectively, are the sets with lowest VC dimension in $\{C_0, \ldots, C_k\}$ and $\{D_0, \ldots, D_l\}$ consistent with *e*. From lemma 3, there is exactly one member of C_i, call it h_C, consistent with *e* and exactly one member of D_i, call it h_D, consistent with the data. By way of contradiction, suppose that h_C is distinct from h_D. But then h_D is not the hypothesis from the set with lowest VC dimension in $\{C_0, \ldots, C_k\}$ consistent with *e*. Let max_C be the maximum number of further mind changes for Ockham's razor on $\{C_0, \ldots, C_k, N\}$ when the data has been extended to *e*. Then by proposition 3(c), any method that conjectures h_D given *e* has a maximum number of further mind changes at least one greater than max_C. Likewise, h_C is not the hypothesis from the set with lowest VC dimension in $\{D_0, \ldots, D_l\}$ consistent with *e*. Let max_D be the maximum number of further mind changes for Ockham's razor on $\{D_0, \ldots, D_l, N\}$ when the data has been extended to *e*. Then by proposition 3(c), any method that conjectures h_C given *e* has a maximum number of further mind changes at least one greater than max_D. But then $max_C < max_D$ and $max_C > max_D$, which is a contradiction. •

References

Chart, D. (2000). Schulte and Goodman's Riddle. *British Journal for the Philosophy of Science* **51**: 147-149.

Cornfield, D, B Schölkopf, and V Vapnik. (2005). Popper, Falsification and the VC-dimension. Technical Report no. 145. Max Plank Institute for Biological Cybernetics.

Elgin, C. (ed.) (1997). *The Philosophy of Nelson Goodman: Nelson Goodman's New Riddle of Induction*, Garland Publishing, New York.

Goodman, N. (1946). A Query on Confirmation. *Journal of Philosophy* **43**: 383-385.

_____ (1954). *Fact, Fiction and Forecast*, Harvard University Press, Cambridge, MA.

Godfrey-Smith, P. (2003). Goodman's Problem and Scientific Methodology," *Journal of Philosophy* **100**: 573-590.

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

Hempel, C. (1965). "Studies in the Logic of Confirmation", in *Aspects of Scientific Explanation and Other Essays*, The Free Press, New York, pp. 4-46.

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

_____ (2007a). A New Solution to the Puzzle of Simplicity. *Philosophy of Science* **74**: 561-573.

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

Popper, K. (1959). *The Logic of Scientific Discovery*, Routledge, New York.

Scheffler, I. (1963). *Anatomy of Inquiry*, Knopf, New York.

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

_____ (1999b). Means-Ends Epistemology. *British Journal for the Philosophy of Science* **50**: 1-31.

_____ (2000). What to Believe and What to take Seriously: A Reply to David Chart Concerning the Riddle of Induction. *British Journal for the Philosophy of Science* **51**: 151-153.

Schwartz, R. (2005). A Note on Goodman's Problem. *Journal of Philosophy* 83: 375-379.

Stalker, D. (1994). *Grue! The New Riddle of Induction*, Open Court, Chicago, IL.

Thomson, J. J. (1966). Grue. *Journal of Philosophy* **63**: 289-309.

Vapnik, V. (2000). *The Nature of Statistical Learning Theory*, Springer, New York.