Mind Changes and Testability:

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 approach to this challenge is to balance fit with the data against some desirable property such as testability or simplicity. Statistical learning theory takes this approach by emphasizing something known as Vapnik-Chervonenkis (VC) dimension, a concept similar to Popper's notion of degrees of testability. In contrast, formal learning theory explains how choices among alternatives that fit the data equally well can be guided by the goal of enhancing 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 basic 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 approach to this challenge is to propose that inductive inference should balance fit with the data against some desirable property such as testability or simplicity. Vapnik-Chervonenkis (VC) dimension (Vapnik 2000), a fundamental concept in statistical learning theory, is a precisely defined property of this kind that is applicable to Goodman's riddle. An alternative approach, pursued in the formal learning literature, is that choices among hypotheses that fit past data equally well should be based on considerations about fast and efficient convergence to truth. Kevin Kelly argues that this approach is superior to accounts, such as statistical learning theory, that attempt to balance with the data against simplicity, testability, or some other allegedly desirable characteristic (2007, forthcoming). In this essay, I show that, despite their apparent differences, statistical and formal learning theory 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.

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

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 notion of testability is very suggestive in this regard. Popper (1999 [1959]) claimed that testable hypotheses are to be preferred not because they are more likely to be true, but because they are a more efficient means for advancing scientific knowledge. This insight leads directly to a second approach to Goodman's riddle, from a genre known as formal learning theory (cf. Kelly 1996, Martin and Osherson 1998), which emphasizes efficient

convergence to the truth. For the present purposes, efficiency is understood as choosing a method that converges to the truth while minimizing the maximum number of "mind changes," that is, retractions of conjectures. In a version of Goodman's riddle in which there is a grue hypothesis for each future date, efficiency in this sense requires conjecturing all green before any of the grue alternatives (Schulte 1999a, 1999b).

I extend these results to a more general version of Goodman's riddle in which the number of possible color switches can be as high as one likes, and I demonstrate a close connection between efficiency and VC dimension in that context. In particular, if hypotheses are grouped according to the number of color switches, 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, the *natural projection rule* 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 the natural projection rule conjectures that all are green. In the generalized version of Goodman's riddle, the natural projection rule 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*.

Finally, I consider the relevance of these results to curve fitting with polynomial equations. Although curve fitting with polynomials resembles the generalized Goodman's riddle in being a VC ordered inductive problem, there are at least two methodologically significant differences between the two cases. The first is that, if polynomials are ordered by degree, then curve fitting is VC ordered but not *strongly* VC ordered. Secondly, while concerns about measurement error are negligible in Goodman's riddle, they are unavoidable in the case of curve fitting. Because of these differences, the results alluded to in the foregoing paragraph do not extend to curve fitting with polynomials. The connection between VC dimension and efficient convergence in curve fitting examples, therefore, remains an open question.

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 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 partition of 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

[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.

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. Consider an inductive problem $\mathcal{P}$ whose set of data sequences is Γ. 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 Γ, 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 1s or 0s. 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 partition of hypotheses is {all are red; for each $n \geq 1$, a hypothesis that all are rue$_n$; 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 bears some important similarities to Popper's notion of degrees of testability and, unlike Popper's notion, is both precisely defined and usefully applicable in a wide range of cases. 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.

There are several appealing features of VC dimension. First, the VC dimension of a set of hypotheses is independent of the language that one uses to state the hypotheses or the data. Since VC dimension is determined solely by logical entailment relations, it cannot be altered by logically equivalent reformulations of sentences. In addition, VC dimension is drawn from a more general theory of statistical inference that is applicable to common scientific inference problems such as curve fitting. Furthermore, there is a straightforward connection between VC dimension and 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. However, VC dimension is not identical to Popper's concept of degrees of testability, since VC dimension is a property of sets rather than individual 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$_{100}$.

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. Thus, if two sets of alternatives fit the data equally well, one should select a hypothesis from a set with 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 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. In other words, since Goodman's riddle indicates an ambiguity in any principle about the uniformity of nature, approaches that presume such principles are not a good answer the riddle he posed.[3]

But VC dimension can be applied directly to Goodman's riddle without assumptions about identically distributed outcomes or any reference to probability at all. Moreover, the conceptual similarity between VC dimension and Popper's notion of degrees of testability suggests an approach to Goodman's riddle. One of the central themes of Popper's philosophy of science is that science does not aim for highly probable theories, but instead for highly testable ones. Since highly testable theories are more easily refuted than less testable ones, it is very doubtful that testability is, in general, an indicator of high probability. Nevertheless, Popper held that highly testable hypotheses are to be preferred because they further scientific progress by generating precise

[3] 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 exactly the question raised by Goodman's riddle.

predictions that put the hypothesis at 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, 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

$N = (S_0 \cup \ldots \cup S_k)^c$

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, $(S_0 \cup \ldots \cup S_k)^c$ is the complement of $S_0 \cup \ldots \cup S_k$, and contains only one member, namely, a statement asserting that neither S_0 nor … nor S_k contains the true hypothesis. I label this set N for "none of the above." 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.

The following, then, is a logically reliable rule for conjecturing hypotheses in the generalized Goodman's riddle:

Natural Projection Rule:[4] 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 N.

Thus, the natural projection rule 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, the natural projection rule 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*, the natural projection rule concludes that the true hypothesis is not to be found in $S_0 \cup \ldots \cup S_k$.

It is easy to show that the natural projection rule 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 the natural projection rule conjectures h, and since h is true, never changes its conjecture after that. On the other hand, suppose that no

[4] I borrow the label "natural projection rule" from Schulte (1999a).

S_i contains the true hypothesis. Then $k + 1$ color switches will eventually be observed in the data, at which point the natural projection rule correctly concludes N.

Any logically reliable method will share three important features with the natural projection rule. 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 the natural projection rule 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 the natural projection rule and furthermore each divergence from the natural projection rule adds one additional mind change to the maximum still possible. First, in the generalized Goodman's riddle the maximum number of mind changes of the natural projection rule 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 the natural projection rule, 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 the natural projection rule 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 the natural projection rule 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 the natural projection rule.

In short, if efficiency is understood in terms of minimizing the maximum number of mind changes, then the natural projection rule 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 the natural projection

rule 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 the natural projection rule 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 and partition of hypotheses Π. 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 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, $N = (C_0 \cup \ldots \cup$

$C_k)^c$. At most, N contains a single statement asserting that no hypothesis in any of the C_i's is true. If the true hypothesis must be in $C_0 \cup \ldots \cup C_k$, then N is the empty set. When N is the empty set, proposition 1 and the generalization of it below hold as before except that $k + 1$ becomes k in (a) and (b). The definition of VC order can also apply in cases in which there is no upper bound on k.

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 version of Goodman's riddle in which the set of alternatives is $[S_0 = \{\text{all are red}\}, S_1 = \{n\text{: all are rue}_n\}, N = \{\text{neither all red nor all rue}\}]$. 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 are inductive problems that are VC ordered but not strongly VC ordered, as is illustrated by the example of curve fitting with polynomials in section 5.

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 $\{C_0, \ldots, C_k, N\}$ and whose set of possible data sequences is Γ. Then the natural projection rule 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 $\{C_0, \ldots, C_k, N\}$ 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 the natural projection rule 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 the natural projection rule. [5]

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 the natural projection rule 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. The corollary is important, because one might suppose that speakers of different languages would be inclined to group hypotheses in different ways. Thus, if distinct strong VC orderings of a single inductive problem could make a difference to which hypothesis the natural projection rule conjectured, then it would seem that Goodman's riddle had not been answered after all. But fortunately, the corollary rules out this possibility.[6]

5. Curve Fitting

[5] Kelly (2004, 2007, forthcoming) provides an account of Ockham's razor that rests on a similar result: each postulation of an entity or effect beyond those required by the data increases the maximum number of further mind changes by one. However, Kelly does not link these results to VC dimension.

[6] For more on efficient convergence and language invariance in Goodman's riddle, see (Chart 2000; Schulte 2000).

Goodman's riddle is similar in some respects to the problem of fitting curves to data. It sometimes supposed, therefore, that an answer to Goodman's riddle must also be an answer to curve fitting problems (cf. Schwartz 2005, 376). But although similar, it is far from clear that Goodman's riddle and curve fitting problems are the same in all relevant respects. I discuss two methodologically significant differences. First, if polynomials are ordered by degree, then curve fitting with polynomials is VC ordered but not strongly VC ordered. The second difference has to do observational error. It is not unreasonable to assume that there are no errors in the data in an example wherein one is presented with a series of balls and decides in each instance whether the ball is red or blue. However, observational error cannot be ignored in an example in which the data consist of measurements of real valued variables. A consequence of these differences is that curve fitting is not a case that falls under the purview of the results from section 4.

Suppose that one is concerned to discover a function $y = f(x)$ that will correctly predict the value of y given a value of x. In this case, the data consists of ordered pairs of real numbers specifying values for x and y, that is, points on a Cartesian plane. For the moment, let us make the unrealistic assumption that there is no measurement error. Suppose moreover that it is assumed that the true function is a polynomial. Right away, we can see several differences between this case and the types of examples considered in the previous section. Most obviously, the possible sequences of data cannot be represented by strings of 1s and 0s, but rather sequences of pairs of real numbers. Nevertheless, curve fitting with polynomials is a VC ordered inductive problem. Suppose that the data so far consist of two points. Then the set of straight lines is of VC dimension 0, since it will be refuted by the next observation if it is not collinear with the

first two. Call this set F_0. Next consider the set of the parabolas, call it F_1. Given two data points, $F_0 \cup F_1$ has a VC dimension of 1, since it is assured of containing a hypothesis consistent with the next data point. However, $F_0 \cup F_1$ does not shatter an extension of the data of length 2: it would be refuted if one but not the other of the next two data points is collinear with the first two. In general, when the data so far consists of two points a Cartesian plane, the polynomial functions are VC ordered by $\{F_0, F_0 \cup F_1, F_0 \cup F_1 \cup F_2, \ldots\}$, where for each F_i contains polynomial functions of degree $i + 1$.

However, this is not a strong VC ordering. In a strong VC ordering, if C_i is the set with lowest VC dimension consistent with the data, then C_{i+1} shatters an extension of the data of length one. However, that condition does not obtain in this case. For example, suppose the data consist of three collinear points. Then F_0 is the set with lowest VC dimension consistent with the data, but $F_0 \cup F_1$ does not shatter an extension of the data of length one, since $F_0 \cup F_1$ is consistent with the next data point only if it is collinear with the first three. Strong VC ordering is essential for proposition 3(c). For inductive problems that are VC ordered by not strongly VC ordered, it is not always the case that a reliable method that jumps ahead to conjecturing a hypothesis from a set S_i of VC dimension higher than necessary can be forced to go back to a set with lower VC dimension and then be forced to conjecture from S_i a second time. For example, consider a method that conjectures a parabola when given only two data points. If it is reliable, the method can be forced to conjecture a straight line if given a sufficient number of further data points collinear with the first two. But now the method cannot be forced to conjecture another parabola, because no parabola is consistent with more than two collinear points.

A second important difference between Goodman's riddle and curve fitting is that in curve fitting measurement error cannot be plausibly idealized out of the picture. Measurements of points on a Cartesian plane can only be approximate, a fact which has significant implications for any attempt to extend the results of the foregoing section to curve fitting examples. Propositions 1 through 3 depend on the assumption that observations may decisively refute a hypothesis, something that cannot be presumed when measurements bear only a probabilistic relationship to the thing being measured. Likewise, the definition of shattering provided in section 2—and presumed throughout sections 3 and 4—is tailored for the special case in which measurement error is absent. That definition stated that 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*. This definition of shattering obviously will not do in cases in which there is measurement error. For example, in curve fitting it is typically assumed that errors are normally distributed with a mean of zero. Thus, so long as the variance of the normal distribution is strictly positive, *any* possible observation has a probability greater than zero given *any* hypothesis. Hence, shattering defined in terms of logical consistency rules practically nothing out if there is measurement error.

The more general definition of shattering, then, is as follows. A set of hypotheses S shatters an extension of the data of length *m* if and only if, for each extension of length *m*, there is a hypothesis in S that perfectly fits that extension. A hypothesis h "perfectly fits" an observation just in case the observed value is the expected or mean value given h. For example, if errors are normally distributed with a mean of zero, then the expected observation given h is just the value that would have to be observed if h were true and

there were no measurement error. In other words, if there is no measurement error, then the data "perfectly fits" a hypothesis exactly when it is consistent with it. Thus, we can see that the definition of shattering given in section 2 is simply a consequence of the more general definition for the special case in which measurement error is absent.

In sum, the generalized Goodman's riddle and curve fitting are similar in that both are VC ordered inductive problems. Nevertheless, they are importantly distinct in that measurement error is negligible in Goodman's riddle but not curve fitting and insofar as the generalized Goodman's riddle is strongly VC ordered while curve fitting with polynomials is not. As a consequence, the results from section 4 do not hold in the case of curve fitting with polynomials. However, that does not rule out the possibility that *analogous* results might obtain. For example, Kelly sketches an account of how the link between simplicity and minimizing retractions can be understood in statistical examples (2004, 497-502).

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, where the noise is represented by an

independent and identically distributed probability distribution. Nevertheless, this paper shows that VC dimension directly maps onto efficient convergence in a class of examples—which I term strongly VC ordered inductive problems—of which Goodman's riddle is just one example. This result naturally raises the question of whether there are further basic similarities between formal and statistical learning theory.

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 $\{C_0, \ldots, C_k, N\}$ and whose set of possible data sequences is Γ. Then for every C_i in $\{C_0, \ldots, C_k\}$, 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, as Γ contains all sequences of 1s and 0s, 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, for any observation in the sequence later than i, h is consistent with exactly one of the possible outcomes for

that observation. For example, consider a hypothesis that allowed that any observation up to and including *i* could be either 0 or 1 but which predicts that every observation after *i* is 0. This hypothesis would be *i*-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 data point *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 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 the natural projection rule 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 also not consistent with the data, then the induction step is complete. 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, the natural projection rule 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, the natural

projection rule 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. By way of contradiction, suppose that there is no extension of e, e^*m_i, such that M conjectures h given e^*m_i. From lemma 3, there is exactly one member of C_i, call it h, consistent with e. And since □ contains all infinite sequences of 1s and 0s, there is a sequence in □ that has e as an initial segment and of which h is true. Yet 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 the natural projection rule is k +1,

(b) For logically reliable method that is logically reliable in $\mathcal{P}$, the maximum number of mind changes is at least 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 the natural projection rule.

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 $d^*m_0^*...^*m_i$, and that, given $d^*m_0^*...^*m_i$, M conjectures a hypothesis in C_i. But since $\mathcal{P}$ is strongly VC ordered, there is a further extension of $d^*m_0^*...^*m_i$ of one additional observation, call it $d^*m_0^*...^*m_i^*1$, such that C_{i+1} is the set with lowest VC dimension consistent with $d^*m_0^*...^*m_i^*1$. From lemma 4, there is an extension of $d^*m_0^*...^*m_i^*1$, call it $d^*m_0^*...^*m_i^*m_{i+1}$, given which M conjectures the hypothesis from C_{i+1} consistent with $d^*m_0^*...^*m_i^*1$. Moreover, the hypothesis from C_{i+1} is not a member of C_i, since otherwise C_i would be consistent with $d^*m_0^*...^*m_i^*1$. 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, the natural projection rule is logically reliable in $\mathcal{P}$. Therefore, from the proof of proposition 1(b), the maximum number of mind changes of the natural projection rule 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 the natural projection rule 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 drawn from C_a but instead from 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, the natural projection rule 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 the natural projection rule. •

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 the *natural projection rule on* $\{C_0, \ldots, C_k, N\}$ refer to the natural projection rule 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 the natural projection rule 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 the natural projection rule 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.

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.

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

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

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

_____ (forthcoming). "Ockham's Razor, Truth, and Information", in *Handbook of the Philosophy of Information*, J. van Behthem and P. Adriaans, eds.

Martin, E. and Osherson, D. (1998). *Elements of Scientific Inquiry*, MIT Press, Cambridge, MA.

Popper, K. (1999 [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.