

Inductive Simplicity and the Matrix*

Gilbert Harman
Department of Philosophy
Princeton University
harman@princeton.edu

Sanjeev Kulkarni
Department of Electrical
Princeton University
kulkarni@princeton.edu Engineering

June 27, 2003

Abstract

In certain statistical learning problems, a policy of choosing simpler rules that account fairly well for data is likely to have less error on new cases than a policy of choosing complex rules that have less error on the data. The relevant kind of simplicity is not to be measured in terms of the number of parameters needed to specify a given member of a class of rules but might be measured in terms of the VC dimension of such a class. The rationale for using simplicity so measured can be extended to allow simplicity to decide among empirically equivalent hypotheses. The extended rationale provides reasons of simplicity to reject certain sorts of philosophical skepticism.

Introduction

An inductive inference infers a conclusion in order to account for data, where data plus background assumptions do not guarantee the truth of the conclusion. We are here concerned with the special case in which the inductive conclusion is a generalization or other sort of general hypothesis or rule that might explain the data. We are not concerned with “transductive inference” directly to a classification of a new example without finding a general rule (Vapnik, 2000, p. 293).

It is characteristic of inductive generalizations that, even when the data are consistent with each other, competing generalizations or general hypotheses or rules would account equally well for the data. An adequate theory of inductive generalization must specify a criterion for selecting preferred hypotheses from among the competing possibilities. To specify

such a criterion is to address what Goodman (1965) calls *the new riddle of induction*.

One form of the new riddle of induction is the *curve fitting problem*: given a number of data-points for a function, what curve should be inferred? It is sometimes suggested that one should infer the *simplest* curve that fits the data, or, more precisely, that inductive generalization must balance considerations of data-coverage against considerations of simplicity.

If so, we need an account of the sort of simplicity that plays or could legitimately play such a role in inductive reasoning.

Simplicity

Distinguish two goals for an account of simplicity in this sense. One is to account for the sorts of simplicity people *actually use* in inductive generalization. Another is to say what sorts of simplicity it is *reasonable* to appeal to in inductive generalization.

To use simplicity in this way in inductive reasoning is not to assume the world is simple. The issue is comparative simplicity. Induction favors a simpler hypothesis over a less simpler hypothesis that fits the data equally well. Given enough data, that preference can lead to the acceptance of very unsimple hypotheses.

Various conceptions of the relevant sort of simplicity have been proposed.

1. A simpler theory is a theory that minimizes entities, or kinds of things, or unexplained events.
2. A simpler theory is a theory that is simpler to state, with a relatively minimal description length.

*For the Annual Meeting of the Cognitive Science Society Meeting in Boston, 2003

3. A simpler theory is a theory that is simpler or easier to use for practical or theoretical purposes.
4. A simpler theory is one that is chosen from a class of hypotheses with relatively few parameters.
5. A simpler theory is one that is chosen from a class of hypotheses with a smaller VC-dimension, where VC-dimension is a measure of the “richness” of the class. (Roughly speaking, the richer a given class of hypotheses is, the harder it is for there to be data that is not captured by one of the hypotheses in the class. See, e.g., Hastie et al., 2001, pp. 210-13.)

Distinguish explicit and implicit appeals to simplicity in science. Explicit appeals to simplicity are often controversial (Sober, 1988, 1990), whereas implicit reliance on simplicity is commonsensical. Given as data that all jonars examined so far have been green and the absence of any special considerations, anyone would take seriously the hypothesis that all jonars are green and no sane person would take seriously the hypothesis that all jonars are grue, that is: either green if examined before noon tomorrow, or blue if not so examined (Goodman, 1965).

Someone might explicitly argue for one rather than another hypothesis on the grounds that the first appeals to fewer entities or kinds of entity, or allows for fewer unexplained events than the other. Such explicit appeals to simplicity are used to distinguish among hypotheses that are being taken seriously. Implicit appeals to simplicity occur when hypotheses that are simpler in certain respects are just not taken seriously. We are here concerned with the latter sort of relative simplicity, which distinguishes hypotheses taken seriously from hypotheses not taken seriously.

Simplicity and the Matrix

Philosophers sometimes discuss whether you have any reason to reject certain parasitic skeptical hypotheses, for example, the hypothesis that you are a brain in a vat programmed in such a way as to have the sorts of experiences of someone in (what you take to be) the ordinary world, an issue illustrated in the movie, *The Matrix* (Wachowski and Wachowski, 1999). Do you have any reason to think you are not in the Matrix?

The philosophical brain-in-a-vat hypothesis, *BIV*, is parasitic on your ordinary hypothesis *OH* because *BIV* says that your experience has been and will be what would be expected if *OH* were true. So, in order to use *BIV* to predict or explain experience, you

are first to figure out what *OH* predicts or explains and use that explanation enclosed in a wrapper: it’s as if *OH*, *OH* predicts or explains *E*, so that’s why *E*. So, there is a sense in which *BIV* is less simple than *OH*.

Does the relative simplicity of *OH* as compared with *BIV* make *OH* more reasonable to accept than *BIV*? The answer depends on what reasons there ever are, if any, to accept simpler rather than more complex hypotheses. We here sketch a well-known result in Statistical Learning Theory. In certain cases, a policy of hypothesis selection that favors simpler hypotheses over less simple hypotheses reduces predictive error in new cases.

Although that result seems to apply only to uses of simplicity to distinguish hypotheses that make different predictions about new cases, we will suggest that it may also be applicable to distinguishing empirically equivalent hypotheses like *OH* and *BIV*.

Policies

Reichenbach (1938, 1949, 1956) suggests that the problem of justifying induction is the problem of justifying certain inductive policies. There is no way to prove that the conclusion of a particular inductive inference must be true or even probable. That is what distinguishes induction from deduction. But it is possible to prove important results about inductive policies. Reichenbach’s particular approach leads to “formal learning theory” concerned with “learning in the limit” (Putnam, 1963; Solomonoff, 1964; Gold, 1967; Blum and Blum, 1975; Jain et al., 1999; Kelley, 1996; Kulkarni and Tse, 1994; Kulkarni and Zeitouni, 1991, 1995; Schulte, 2002).

More statistical approaches to pattern recognition and statistical learning are also concerned with proving results about policies (Vapnik, 2000; Hastie, et al., 2001).

Simplicity and Data

A good inductive reasoner does not always just accept the simplest rule that fits the data perfectly, because the situation may be ineliminably stochastic and in any event the data will be noisy because of measurement errors, etc.

For example, consider a classification problem in which we must choose a decision rule from a certain class of rules, a rule that will divide new items into two groups, yes and no, based on their observed features, where we assume there is an unknown probability distribution determining what items we see,

what features they have, and whether they are yes examples or no examples. We have some data with correct classifications (training data).

The best decision rule will be the one with the lowest expected error (assuming that errors of one sort—false positives—are no worse than others of the other sort—false negatives.) We want to find a rule whose expected error rate is as close as possible to the best decision rule.

In this context, it is important to avoid “overfitting” the data. Given a choice between a more complex rule that perfectly fits the data and a simpler rule that gives a good approximation, the simpler rule is likely to have less error with new cases. (Forster and Sober, 1994, forthcoming.)

Of course, data cannot be ignored. The point is that good inductive practice balances simplicity of rule against error in the data.

Ordering Issues

How are rules to be ordered in terms of simplicity? Distinguish two sorts of simplicity ordering. Rules can be well-ordered, for example, by length of formulation (with rules of the same length ordered alphabetically). Or rules can be placed in classes which are well-ordered, e.g. in terms of the number of parameters needed to pick out a particular member from a given class or in terms of the VC-dimension of the class (a measure of the richness of the class of rules in terms of ability to find a rule that fits the data).

The second approach, in which classes of rules are well-ordered can allow for nondenumerably many hypotheses. For example all linear hypotheses of the form $f(x) = ax + b$ can be included in one of the classes, allowing the parameters a and b to range over all real numbers. And, even if only denumerably many hypotheses are considered, such as those that can be expressed in some favored notation, those hypotheses need not themselves be well-ordered. If the infinitely many linear hypotheses are ordered before the infinitely many quadratic hypotheses, the ordering is not a well-ordering.

Using simplicity of representation as a measure of simplicity is affected by what system of representation is chosen. Any particular hypothesis can of course be given an arbitrarily simple representation. Goodman (1965) observes that we might define the predicate *grue* that applies to something x at a given time t if and only if, either x is examined for color prior to noon tomorrow and is green at t , or x is not examined for color prior to noon tomorrow and is blue at t . This allows the bizarre hypothesis men-

tioned earlier to be represented as “All jonars are grue,” which is as simple as “All jonars are green.” Indeed, any rule can be represented with a single symbol.

Remember, however, that the issue has to do with inductive policies. A policy that measures the simplicity of a rule by the length of its representation, must put restrictions on how rules are to be represented. Otherwise, all rules would count as equally simple and simplicity, so interpreted, could not be used to distinguish among rules.

Trigonometric Hypotheses

It may seem we can simply fix the system of representation ahead of time. But not for all time! Mathematics and science regularly introduce new concepts that allow simple statements of what we come to consider simple rules. A rule that appealed to trigonometric concepts would require an infinite representation in a system that allowed only polynomial representations of functions. We need to allow the relevant system of representation to be able to include useful new notions, such as trigonometric notions (Harman, 1999).

Allowing trigonometric functions creates a problem for the idea that we should order classes rules in terms of the number of parameters needed to specify a particular member of the class. Consider the class of curves of the form, $a \sin(bx) + c$. Three parameters are enough to specify a particular function from this class. But, however many data points there are, it will almost always be possible to find some sine curve of this form that goes through every one of the points. The problem is that the sine curve can have a very high frequency. But in general it is not rational always to prefer a sine function from that class over a polynomial in a class of functions requiring four parameters to be specified. (The difficulty here is that the class of sine curves has infinite VC-dimension.)

These issues are discussed in Statistical Learning Theory, which examines the likely predictive success of procedures for choosing rules of classification, given relatively minimal assumptions. There are many useful results. (Vapnik, 2000; Hastie, et al., 2001).

Empirically Equivalent Rules

Suppose two hypotheses, H and D , are empirically equivalent. For example, where H is some highly regarded scientific hypothesis, let D be the corresponding demonic hypothesis that a powerful god-like demon has arranged that the data you get will be ex-

actly as expected if H were true. Could simplicity as analyzed in Statistical Learning Theory provide a reason to accept H rather than D ?

One might suppose that the answer is “no”, because the kinds of analyses provided by Statistical Learning Theory concern how to minimize predictive error and these hypotheses make exactly the same predictions. Indeed, if we identify the hypotheses with their predictions, they are the same hypothesis.

But it isn’t obvious that hypotheses that make the same predictions should be identified. The way a hypothesis is represented suggest what class of hypotheses it belongs to for purposes of assessing simplicity. Different representations suggest different classes. Even mathematically equivalent hypotheses might be treated differently within Statistical Learning Theory. The class of linear hypotheses, $f(x) = ax + b$, is simpler than the class of quadratic hypotheses, $f(x) = ax^2 + bx + c$, on various measures—number of parameters, VC-dimension, etc. If the first parameter of a quadratic hypothesis is 0, the hypothesis is mathematically equivalent to a linear hypothesis. But its linear representation belongs to a simpler class than the quadratic representation. So for purposes of choice of rule, there is reason to count the linear representation as simpler than the quadratic representation.

Similarly, although H and D yield the same predictions, they are not contained in the same hypothesis classes. H falls into a class of hypotheses with a better simplicity ranking than D , perhaps because the class containing H has a lower VC-dimension than the class containing D . The relevant class containing D might contain any hypothesis of the form, “The data will be exactly as expected as if ϕ were true,” where ϕ ranges over all possible scientific hypothesis. Since ϕ has infinite VC-dimension, so does this class containing D . From this perspective, there is reason to prefer H over D even though they are empirically equivalent.

So, we may have reason to think that we are not just living in the Matrix!

References

1. Blum, L., and Blum, M., (1975). “Toward a Mathematical Theory of Inductive Inference,” *Information and Control* 28: 125-55.
2. Forster, M. and Sober, E., (1994). “How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions.” *British Journal for the Philosophy of Science* 45: 1-36.

3. Forster, M. and Sober, E., (forthcoming). “Why Likelihood?” in M. Taper and S. Lee (ed.), *The Nature of Scientific Evidence*, University of Chicago Press.
4. Gold, E. M., (1967). “Language Identification in the Limit,” *Information and Control* 10: 447-74.
5. Goodman, N., (1965). *Fact, Fiction, and Forecast*, 2nd edition. Indianapolis, Indiana: Bobbs-Merrill.
6. Harman, G., (1999). “Simplicity as a pragmatic criterion for deciding what hypotheses to take seriously,” in G. Harman, *Reasoning, Meaning, and Mind*. Oxford: Oxford University Press.
7. Hastie, T., Tibshirani, R., & Friedman, J., (2001). *The Elements of Statistical Learning*. New York, Springer.
8. Jain, S., Osherson, D., Royer, J., and Sharma, A., (1999). *Systems That Learn*, Second Edition. Cambridge, MA: MIT Press.
9. Kelley, K. T., (1996). *The Logic of Reliable Inquiry*. Oxford: Oxford University Press.
10. Kulkarni, S. R. and Tse D. N. C., (1994). “A Paradigm for Class Identification Problems,”; *IEEE Transactions on Information Theory*, Vol. 40, No. 3, pp. 696-705.
11. Kulkarni, S.R. and Zeitouni, O. (1991). “Can one decide the type of the mean from the empirical measure?” *Statistics & Probability Letters*, Vol. 12, pp. 323-327.
12. Kulkarni, S.R. and Zeitouni, O. (1995) “A General Classification Rule for Probability Measures,” *Annals of Statistics*, Vol. 23, No. 4, pp. 1393-1407.
13. Putnam, H., (1963). “Degree of Confirmation and Inductive Logic.” In *The Philosophy of Rudolf Carnap*, ed. A. Schillp. LaSalle, Indiana: Open Court.
14. Reichenbach, H., (1938). *Experience and Prediction*. Chicago: University of Chicago Press.
15. Reichenbach, H., (1949). *The Theory of Probability*, 2nd edition. Berkeley and Los Angeles: University of California Press.
16. Reichenbach, H., (1956). *The Rise of Scientific Philosophy*. Berkeley: University of California Press.

17. Schulte, O. (2002). "Formal Learning Theory," *The Stanford Encyclopedia of Philosophy* (Spring 2002 Edition).
18. Sober, E., (1988). *Reconstructing the Past*. Cambridge, Massachusetts: MIT Press.
19. Sober, E., (1990). "Let's razor Occam's razor." In D. Knowles, ed., *Explanation and Its Limits*. Cambridge: Cambridge University Press.
20. Solomonoff, R. J., (1964). "A Formal Theory of Inductive Inference," *Information and Control* 7: 1-22, 224-54.
21. Vapnik, V. N., (2000). *The Nature of Statistical Learning Theory*, Second Edition. New York: Springer.
22. Wachowski, A., and Wachowski, L., (1999). *The Matrix*. Warner Brothers.