

Simplicity tracks truth because compression tracks probability

August 26, 2008*

Observational evidence is never enough to determine a unique scientific theory. We are therefore forever destined to pick a privileged theory (from among those compatible with the data) on criteria that are not straightforwardly evidential. The paragon of such theoretical virtues, of course, is simplicity. The simpler theory will provide at least certain pragmatic charms, but we generally find ourselves more compelled by it than that—not just to use it as the handiest of the theories, but also to *believe* it as the truest. This compulsion of ours is tricky to justify, though, for it is difficult to see how simplicity could be any kind of truth-indicator.

For a while now it has been fashionable to take a statistical approach to this problem, outlining a formal notion of simplicity that has pleasing theoretical properties. Just for some examples, there's the “Akaike information criterion” approach, the similar “minimum message length” and “minimum description length” approaches, and the “Vapnik-Chervonenkis dimensionality” approach.¹ It should be both intriguing and suspicious that there are so many such programs. I here propose

*This is a *draft* (version: *compression-truth.tex,v2.10*), and should not be circulated or cited without permission. Thanks especially to [several people of learning and distinction].

¹See, for example, Forster and Sober (1994), Wallace (2005), Grünwald (2007), and Harman and Kulkarni (2006), respectively. I should add here that I do not include the formal learning-theoretic approach of Kelly (1996) in this “statistical” category—nor would he. Kelly argues that simplicity gets us to the truth faster, as measured by the number of times we can be forced to change theories. His program is more ambitious than my argument, and (I think) compatible with it.

some explanation for this phenomenon: at the heart of these statistical justifications is something correct—something both more simple and more modest than previously proposed, but something that nonetheless gives genuine alethic support for picking the simpler theory. That common core is in the nature of *data compression*, and its inherent tie to the accurate tracking of probabilities.²

1 Compression, simplicity, and probability

As finite human creatures, we cannot easily store all the information we receive through our sensory nerves. Even if we could—say, with the help of a USB interface and some external hard drives—it would probably be of little use to us, since we would have trouble accessing and sorting through that vast repository of raw information in time sufficient to help us act. We thus resort to shortcuts and heuristics; we abstract from the wash of sensory information to important patterns that help us cope. Some of this abstraction is hardwired (roughly speaking), as for example with border and depth detection in vision; other abstractions we learn, as for example telling a Kandinsky from a Klee. The harsh reality of data storage and retrieval forces us to choose the data that’s “important”. This in turn continually forces us, in effect, to theorize. We cannot keep every sense-datum of every raven-like visual experience we’ve ever had, so we hypothesize key points—such as that all ravens are black—and remember those.

In fact if we think of our total experience as a huge set of data points (say, the activation states of all our sensory nerves over time), then we can think of theorizing as an attempt to *compress* that data, presenting the information in a more compact form. Holding data-fit as equal, we could say a theory is simpler to the extent it compresses the data better. When theorizing is understood this way, it is easy to see how finite physical creatures might do it, since we know that data compression can be done by algorithm. Taking simplicity as degree of data compression is thus a tidy, naturalistic construal of inferring to simpler explanations.

²MacKay (2003) emphasizes this tie, and in many ways is the inspiration for this paper.

Furthermore, to the extent this is a good way to measure a theory's simplicity, we have a good argument that simplicity tracks truth. For as Shannon (1948) showed, our ability to compress data depends directly on how well we assess the probability of their occurrences. There are simple heuristics to see why this must be the case. Take some chunk of data—it is easiest to think of it in terms of a computer file of zeroes and ones. No compression program could shrink *every* such file without loss of information, by the pigeonhole principle: there are 2^n different files with n bits, and only $2^n - 2$ different files of fewer bits (with the majority of those only 1 bit shorter). Alternatively, here is a *reductio* argument for the same conclusion: if a compression scheme could take any file and shrink it losslessly, then it could shrink that compressed file too, and so on—so that all files could be compressed to one bit in length by that scheme, without loss of information. Clearly, this is impossible, so there cannot be a compression scheme that compresses every file.³

The lesson is that compression schemes must play the odds—any algorithm that losslessly shrinks some files must *expand* others (or simply fail on them). The trick is to set up the compression scheme so that it compresses the kind of files it sees frequently, and expands kinds of files that come along only very rarely, if ever. But this means compression only works to the extent we have a good idea of what kind of files are most often fed to the compressor. If any of the 2^n different files is just as probable as any other, then compression is impossible. (In the long run, at any rate; the information-theoretic entropy is maximal in such a case, and Shannon showed compression can only go down to the entropy.) If however some types of files are more probable than others, we can exploit these regularities. It is overwhelmingly probable that an English text file will have many more letters like 'e' than like 'q', and many more strings like 'the' than like 'xqq', and compression algorithms thrive on such facts. Compressors specifically designed for English text will do better on such files than generic compressors; generic compression schemes simply “expect” that there are regularities of some kind in the data, and are constructed to find them. (The similarity of such

³These arguments apply only to compression that loses no information—you can do “lossy” compression as far as you like, *if* you're not picky about how much is lost (as the “lzip” joke for computer geeks points out). To the extent you want to keep information, though, these arguments will apply.

algorithms to general-purpose theorizing has led the likes of Baum (2004) and Hutter (2005) to claim that the problem of artificial intelligence simply is the problem of compression.)

2 Simplicity and truth

We are now in a position to spell out the alethic justification of inferring to simpler theories.

2.1 The immodest part

If abductive theories are understood as compressions of raw data, and if the simplicity of such theories is rightly understood as the degree of compression, then there is a straightforward argument that simplicity tracks truth:

1. Degree of compression varies with correctly assessing the probabilities of the data.
2. Therefore, a better-compressing theory is evidence that we are tracking probabilities.
3. We have truth-related reason to pick the theory that better tracks probabilities.
4. Therefore, we have truth-related reason to pick the better-compressing theory.
5. A theory with a higher degree of data compression is a simpler theory.
6. Therefore, we have truth-related reason to pick the simpler theory.

To put premise (2) in Bayesian terms, the higher likelihood of accuracy given good compression indicates a higher probability of accuracy given good compression.⁴ Good compression is evidence that we are correctly tracking probabilities in the data.

⁴For good compression C and high accuracy A , Shannon showed in effect that $P(C|A) > P(C)$, which means $P(A|C) > P(A)$.

2.2 The modest part

This argument comes with an important qualification. Though it gives truth-based reason to prefer simpler theories, and though many since Harman (1965) feel that induction is simply inference to the simplest explanatory theory, this argument does *not* provide a solution to the problem of induction. Good compression of observed data is evidence that we have found actual regularities *in those observed data*. The probabilities at issue here are straightforward frequencies, and these frequencies may not reflect the underlying probability distribution from which the data are sampled. Compare a string of one hundred flips of heads from a coin that is in fact fair. “All heads” would be a good compression that accurately summarizes an important pattern in the observed data, despite the fact that (in the long run) it would not be a good compression of data to come. In this sense, then, the argument is modest; it does nothing to say why we should predict that future data will match our current theories.

Still, compression is at least evidence of accurate pattern assessment in the observations, and that is no trivial task. Consider the challenge of the Hutter Prize, which rewards incrementally greater compression of a 100MB chunk of Wikipedia.⁵ To win some of the money requires eking out ever more elaborate patterns in the data. It’s one thing to notice the frequency of ‘u’ after ‘q’, for example, and quite another to notice the frequency of ‘knowledge’ some time after ‘epistemology’, or that patterns in the data make ‘cat’ probable after the ‘the most familiar felid is the’. Pattern recognition is difficult, in other words, and compression is a good indication that we are catching on to real patterns—in the observed data, at least. This fact should go some way toward reassuring us that our ubiquitous practice of inferring to simpler theories is not merely pragmatic. On the picture here, pragmatic concerns do force us to theorize in the first place, since we cannot keep and process all incoming data. But given the necessity of theorizing, preference for simpler theories (where data fit is held equal) has probabilistic truth-based justification.

⁵<http://prize.hutter1.net>. See Mahoney (2006) for an overview of the rationale for artificial intelligence work.

Aside from such reassurance, I hope this argument will also stimulate interest in the relation between data compression and intelligence. I am not convinced that these are in any important way the *same*, myself, but the relation is fascinating nonetheless.

References

Baum, E. B. (2004). *What is Thought?* MIT Press.

Forster, M. R. 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.

Grünwald, P. D. (2007). *The Minimum Description Length Principle*. MIT Press.

Harman, G. (1965). The inference to the best explanation. *The Philosophical Review*, 74(1):88–95.

Harman, G. and Kulkarni, S. (2006). *Reliable Reasoning: Induction and Statistical Learning Theory*. MIT Press.

Hutter, M. (2005). *Universal Artificial Intelligence*. Springer.

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

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2004 edition.

Mahoney, M. (2006). Rationale for a large text compression benchmark. Online; last accessed August 14, 2008 from <http://cs.fit.edu/~mmahoney/compression/rationale.html>.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656.

Wallace, C. S. (2005). *Statistical and Inductive Inference by Minimum Message Length*. Springer-Verlag.