5-1-2002

Why is Selecting the Simplest Hypothesis (Consistent with Data) a Good Idea? A Simple Explanation

Vladik Kreinovich
University of Texas at El Paso, vladik@utep.edu

Luc Longpre
University of Texas at El Paso, longpre@utep.edu

Scott Ferson

Lev Ginzburg

Comments:
UTEP-CS-02-19.
Published in Bulletin of the *European Association for Theoretical Computer Science (EATCS)*, 2002, Vol. 77, pp. 191-194.

Why Is Selecting the Simplest Hypothesis (Consistent with Data) a Good Idea? A Simple Explanation

Vladik Kreinovich[1], Luc Longpré[1],
Scott Ferson[2], and Lev Ginzburg[2]

[1]Department of Computer Science
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
{vladik,longpre}@cs.utep.edu

[2]Applied Biomathematics
100 North Country Road
Setauket, NY 11733, USA
{scott,lev}@ramas.com

"Everything should be made as simple as possible, but not simpler."
A. Einstein, *Autobiographical Notes* [3]

When there are several different hypotheses that can explain an observed phenomenon, which of these hypotheses should we choose?

Occam, a well known 13th century philosopher, was the first to formulate a natural idea: choose the simplest of these hypotheses. This principle has been successfully used in many areas of science.

Occam's principle works very well: why?

In many cases, this Occam's principle is in good accordance with common sense. For example, when we observe that the amount N_t of the radioactive material decays exponentially with time t – i.e., that for some constant α, we have $N_t = \exp(-\alpha \cdot t)$ for $t = 1, 2, \ldots, T$ – then it is natural to select the hypothesis that $N_t = \exp(-\alpha \cdot t)$ for all t, although from the purely mathematical viewpoint, a hypothesis that, say, $N_t = \exp(-\alpha \cdot t)$ for $t \leq T$ and $N_t = \cos(\ln(t))$ for $t > T$ would be also equally good in explaining all the data.

Occam's principle is deeper than such common sense examples. Not only it works well in physics, but, e.g., the authors of a recent survey [2] on expert estimates noted, to their surprise, that the simplest models explaining the original expert estimates turned out to be the best in fitting the following data as well.

How can we explain this unexpected efficiency of Occam's principle?

Existing explanations of Occam's principle: what's good, what's bad, and what we are planning to do. Occam's principle has a nice insightful explanation within Algorithmic Information Theory – i.e., theory of Kolmogorov complexity; see, e.g., [1, 6].

The connection between Occam's principle and Kolmogorov complexity dates back to pioneer papers [7, 8] by R. J. Solomonoff – one of the three founders (with A. N. Kolmogorov and G. J. Chaitin) of this area. This explanation, however, is somewhat technical and requires technical-level understanding of Kolmogorov complexity.

For *probabilistic* physical theories, the technicality of the existing explanations is probably unavoidable. Indeed, before we proceed with any such explanation, we need to first formalize what it means for a probabilistic theory to be consistent with the observations (i.e., whether the observations are random relative to the probability measure predicted by the theory). The necessity for such formalization is what started Kolmogorov complexity in the first place.

The main objective of this paper is to give a simple explanation for Occam's principle for *deterministic* physical theories, an explanation that would be more accessible to researchers outside theory of computing.

Our explanation of Occam's principle: definitions and the main results. Let O and H be two countable sets. Elements of the set O will be called *observations*, elements of the set H will be called *hypotheses*.

For simplicity, we will consider discrete time, with time moments $1, 2, \ldots, n, \ldots$ The set of all possible moments of time will be denoted by $T = \{1, 2, \ldots\}$.

Let $p : H \times T \to O$ be a function called *prediction function*[1]. We say that a hypothesis h predicts observation $p(h, t)$ at time t.

We assume that different hypotheses lead to different predictions, i.e., that if $h \neq h'$, then $p(h, t) \neq p(h', t)$ for some moment of time $t \in T$.

Let $C : H \to N$ be a function that assigns to every hypothesis a natural number. We will call the value $C(h)$ a *complexity* of the hypothesis h. We will require that for every hypothesis h, there are only finitely many hypotheses $h' \in H$ that are simpler than h (i.e., hypotheses for which $C(h') \le C(h)$).

By *data*, we mean a sequence of observations $o_1, \ldots, o_n \in O$. We will say that o_t is the *observation at moment* t.

We say that a hypothesis h is *consistent* with the data $o_1, \ldots, o_n$ if for every t from 1 to n, we have $p(h, t) = o_t$.

We assume that there exists an *actual world history*, i.e., an infinite sequence of observations $o_1, \ldots, o_n, \ldots$ We assume that the class H contains the *correct* hypothesis, i.e., a hypothesis h_0 that is consistent with all the observations (to be more precise, a hypothesis

[1] While in practice, we would expect this function to be computable, in our derivation, we do not make any assumptions about the computability of the functions involved.

for which $p(h_0, t) = o_t$ for all t). Since we assumed that different hypotheses lead to different predictions, there is only one correct hypothesis.

For every time t, the correct hypothesis is consistent with the observations $o_1, \ldots, o_t$. In addition to the correct hypothesis, we may have several other hypothesis consistent with the same observations. Based on these observations, we do not know which of these hypotheses is correct, so we must select one of them. Occam's principle says that at each moment of time t, we select the simplest of all the hypotheses h that are consistent with the observations $o_1, \ldots, o_t$ (i.e., the hypothesis with the smallest possible complexity $C(h)$; if there are several simplest hypotheses, we select one of them arbitrarily).

It turns out that if we follow Occam's principle, then, eventually, we will pick the correct hypothesis. On the other hand, if we consistently pick a non-simplest hypothesis, we may never select the correct hypothesis. Let us formulate these results in more precise form (and prove them):

Theorem 1. *Let $o_1, \ldots, o_t, \ldots$ be an actual world history. If at every moment of time t, among all hypotheses which are consistent with the observations $o_1, \ldots, o_t$, we select the simplest one (or one of the simplest ones), then there exists a moment of time t_0 after which we always select the correct hypothesis.*

Theorem 2. *If for some actual world history, for every t, among all hypothesis which are consistent with the observations $o_1, \ldots, o_t$, we select a hypothesis which is not the simplest, then there exists a time t_0 after which we will never select a correct hypothesis.*

Comment. From the purely mathematical viewpoint, it is possible that for some t, there is only one hypothesis consistent with observations $o_1, \ldots, o_t$. In this case, Theorem 2 is true by default. In practice, however, there are always many hypotheses consistent with given data.

Proof of Theorem 1. Let h_0 be the correct hypothesis. Due to the property of the complexity function, there exist only finitely many hypotheses $h_1, \ldots, h_m$ that are simpler than h_0, i.e., for which $C(h_i) \le C(h_0)$. Since different hypotheses lead to different predictions, for each of these hypotheses h_i, there exists a moment of time t_i for which its prediction is different from the prediction of the correct hypothesis h_0, i.e., for which $p(h_i, t_i) \ne p(h_0, t_i)$.

Let $t_0 \stackrel{\text{def}}{=} \max(t_1, \ldots, t_n)$, and let us show that for every $t \ge t_0$, h_0 is selected. Indeed, the correct hypothesis h_0 is clearly consistent with all the observations $o_1, \ldots, o_t$. Since we select the simplest hypothesis consistent with the observations, we must now show that for any $h \ne h_0$ such that $C(h) \le C(h_0)$, the hypothesis h is not consistent with $o_1, \ldots, o_t$. The only hypotheses h for which $C(h) \le C(h_0)$ are $h_1, \ldots, h_m$. For each of these h_i, we have $p(h_i, t_i) \ne p(h_0, t_i)$. Since h_0 is correct, we have $p(h_0, t_i) = o_{t_i}$, hence $p(h_i, t_i) \ne o_{t_i}$. Due to our choice of t_0, we have $t_i \le t_0 \le t$, hence the hypothesis h_i is not consistent with one of the observations $o_1, \ldots, o_t$ – namely, with the observation o_{t_i}.

The theorem is proven.

Proof of Theorem 2. In the proof of Theorem 1, we have shown that there exists a moment t_0 such that for all consequent moments of time $t \ge t_0$, the correct hypotheses is the simplest hypothesis among all hypotheses which are consistent with the observations $o_1, \ldots, o_t$. Since, by assumption, we always select a hypothesis that is not the simplest, this means that for all such moments of time $t \geq t_0$, we are not selecting the correct hypothesis. The theorem is proven.

Comment. Our proofs are clearly applicable to Kolmogorov complexity $C(h)$: indeed, the Kolmogorov complexity is, by definition, the shortest length of a program computing h. There are only finitely many shorter programs and therefore, only finitely many hypotheses of smaller Kolmogorov complexity.

It is worth mentioning that in the above proofs, the function $C(h)$ does not have to be Kolmogorov complexity, it can be any function – provided that for every hypothesis h, there are only finitely many hypotheses $h' \in H$ with smaller values of $C(h')$ (i.e., for which $C(h') \leq C(h)$).

Thus, our results cover not only the natural idea of selecting the simplest hypothesis, but – arguably – similarly natural philosophical ideas produced by physicists, such as (Einstein's favorite) selecting the most *beautiful* hypothesis (in Einstein's words, "The pursuit of truth and beauty"). Here, $C(h)$ is the degree to which h is not esthetically pleasing (for attempts to formalize this notion in Kolmogorov complexity-related terms, see, e.g., [4, 5]).

Acknowledgments. This work was supported in part by NASA under cooperative agreement NCC5-209 and grant NCC 2-1232, by NSF grants CDA-9522207, ERA-0112968 and 9710940 Mexico/Conacyt, by the Air Force Office of Scientific Research grant F49620-00-1-0365, and by Small Business Innovation Research grant 9R44CA81741 to Applied Biomathematics from the National Cancer Institute (NCI), a component of the National Institutes of Health (NIH).

References

[1] C. Calude, *Information and randomness: An algorithmic perspective*, Springer-Verlag, Berlin, 1994.

[2] R. T. Clemen and R. L. Winkler, "Combining probability distributions from experts in risk analysis", *Risk Analysis*, 1999, Vol. 19, No. 2, pp. 187–203.

[3] A. Einstein, *Autobiographical Notes*; edited by P. A. Schlipp, Open Court Publ., Chicago, 1991.

[4] M. Koshelev, "Towards The Use of Aesthetics in Decision Making: Kolmogorov Complexity Formalizes Birkhoff's Idea", *Bulletin of the European Association for Theoretical Computer Science (EATCS)*, 1998, Vol. 66, pp. 166–170.

[5] V. Kreinovich, L. Longpré, and M. Koshelev, "Kolmogorov complexity, statistical regularization of inverse problems, and Birkhoff's formalization of beauty", In: A. Mohamad-Djafari (ed.), *Bayesian Inference for Inverse Problems, Proceedings of the SPIE/International Society for Optical Engineering*, Vol. 3459, San Diego, CA, 1998, pp. 159–170.

[6] M. Li and P. Vitányi, *An introduction to Kolmogorov complexity and its applications*, Springer-Verlag, N.Y., 1997.

[7] R. J. Solomonoff, *A preliminary report on a general theory of inductive inference*, Technical Report ZTB-138, Zator Company, Cambridge, MA, November 1960.

[8] R. J. Solomonoff, "A formal theory of inductive inference, Parts 1 and 2", *Information and Control*, 1964, Vol. 7, pp. 1–22, 224–254.