Minimum message length as a truth-conducive simplicity measure

Stephen Petersen
steve@stevepetersen.net

Department of Philosophy
Niagara University

Formal Epistemology Workshop
2 June 2007

> "How easy it is for the gullible to clutch at the hope that *somehow*, deep in the cabalistic mysteries of information, computation, Gödel, Escher, and Bach, there *really is* an occult connection between simplicity and reality that will direct us unswervingly to the Truth; that prior probabilities constructed to favor computationally compressible strings really are *informative* and that learning can be *defined* as data-compression. After all, aren't these things constructed out of "information"? I know whereof I speak—I have met these glassy-eyed wretches (full professors, even) and they are beyond salvation."
>
> —*Kevin Kelly*

Introduction

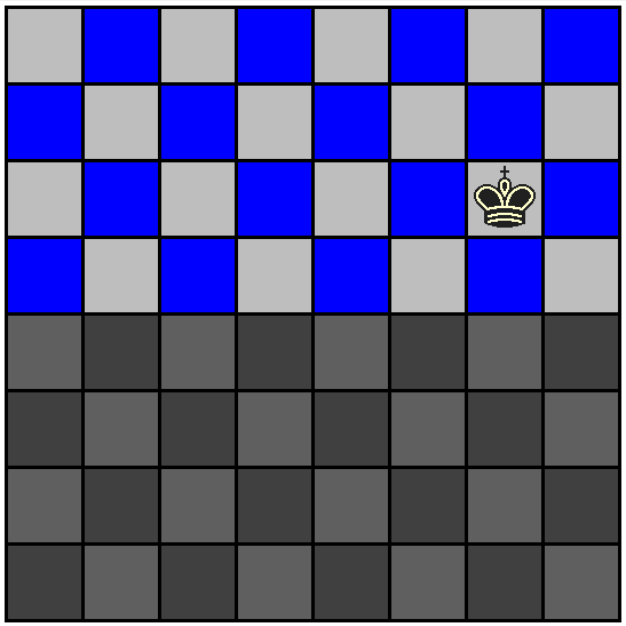
- The problem area
  - philosophy of science
  - epistemology
  - cognitive science
- Prospects for the *Minimum Message Length* (MML) algorithm
- A purported truth-conducive simplicity measure

Outline

Shannon information: the basic idea

- Uncertainty about data source $\approx$ variety of possible results
- More possibilities $\approx$ more uncertainty
- *Entropy* as uncertainty-measure
- *Information* as reduction in entropy (uncertainty)

Shannon information: an example



- How many *yes/no questions* needed to find the king?
- 6 divisions of possibility space in half
- $(\frac{1}{2})^6 = \frac{1}{64}$
- $\log_{\frac{1}{2}} \frac{1}{64} = \log_2 64 = 6$
- Measure in *bits*
- Efficient binary encoding uses 6 digits

Shannon information: information and entropy

- Risky approach: *e.g.*, "is it on square f7?"
- If yes, we gain 6 bits of *information*
- If no, we gain only about .023 bits
- The ideal strategy maximizes information gained with either answer

Shannon information: subjectivity

- Suppose odds were $\frac{1}{2}$ the king is on its original square
- Clever first question: "is it on its home square?"
- Now *on average* 3.99 yes/no questions needed
- The entropy now 3.99 bits
- New digital encoding wise, too
- Lower entropy reflects receiver's lower uncertainty about source
- A *subjective* measure

Shannon information: definitions

- Shannon information: $h(x) = \log_2 \frac{1}{\mathrm{P}(x)}$
- Shannon message length: $\mathrm{Len}_\mathrm{s}(x) = h(x)$
- Shannon entropy: $H(X) = \sum_x \mathrm{P}(x) h(x)$

Kolmogorov complexity in one slide

- Universal Turing Machine U, target string s
- "Description" $d_U(s)$: the *shortest* input to U that returns s
- Kolmogorov complexity $K_U(s) = |d_U(s)|$
- Detectable patterns (repeated sequences, the digits of π in binary ...) $\rightarrow$ shorter programs
- Interesting measure of randomness
- $\text{Len}_s(s)$ and $K(s)$ closely related

MML: The basic idea

- Bayes: most probable hypothesis given the data
- Shannon: more probable claims allow shorter messages
- MML: find most probable hypothesis by shortest message
- $\approx$ find truth *via* simplicity

MML is truth-conducive: the "proof"

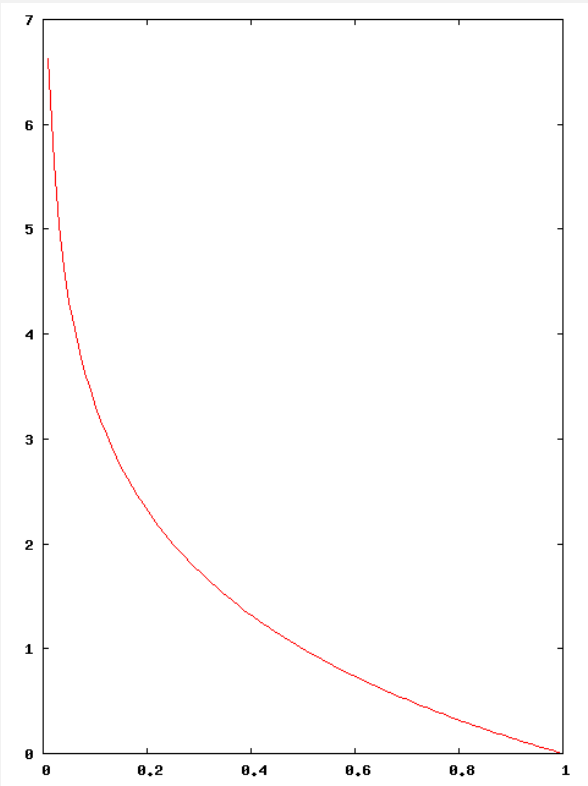


Figure: $\log_2 \frac{1}{\mathrm{P}(x)}$

We have truth-related reason to pick

$$\begin{aligned}
\operatorname*{argmax}_i \mathrm{P}(h_i|e) &= \operatorname*{argmax}_i \frac{\mathrm{P}(h_i, e)}{\mathrm{P}(e)} \\
&= \operatorname*{argmax}_i \mathrm{P}(h_i, e) \\
&= \operatorname*{argmin}_i \log_2 \frac{1}{\mathrm{P}(h_i, e)} \\
&= \operatorname*{argmin}_i \mathrm{Len_s}(h_i, e)
\end{aligned}$$

Simplicity and data-fit

- Epistemic reason to pick shortest statement of hypothesis with data
- Big rabbit, small hat
- Note especially:

$$\begin{aligned}\text{Len}_s(h_i, e) &= -\log_2 \text{P}(h_i, e) \\ &= -\log_2 \text{P}(h_i)\text{P}(e|h_i) \\ &= -\log_2 \text{P}(h_i) + -\log_2 \text{P}(e|h_i) \\ &= \text{Len}_s(h_i) + \text{Len}_s(e|h_i)\end{aligned}$$

- Balance of simplicity and data-fit!

The MML explanation

Table: Wallace's "explanation"

$\text{Len}_s(h)$	$\text{Len}_s(e\|h)$
"assertion"	"details"
curve	data error
pattern	noise
regularities	exceptions
compressor	compressed file

Explaining to a UTM

- The eyes get still glassier . . .
- Assume the Shannon-encoding scheme h is computable
- Program a Universal Turing Machine U to compute it with h_U
- U determines an implicit prior distribution H_U for such programs
- $\text{Len}_s(h|H_U) \approx K_U(h_U) = |h_U|$
- $\text{Len}_s(e|h, H_U) \approx K_{U|h}(e)$

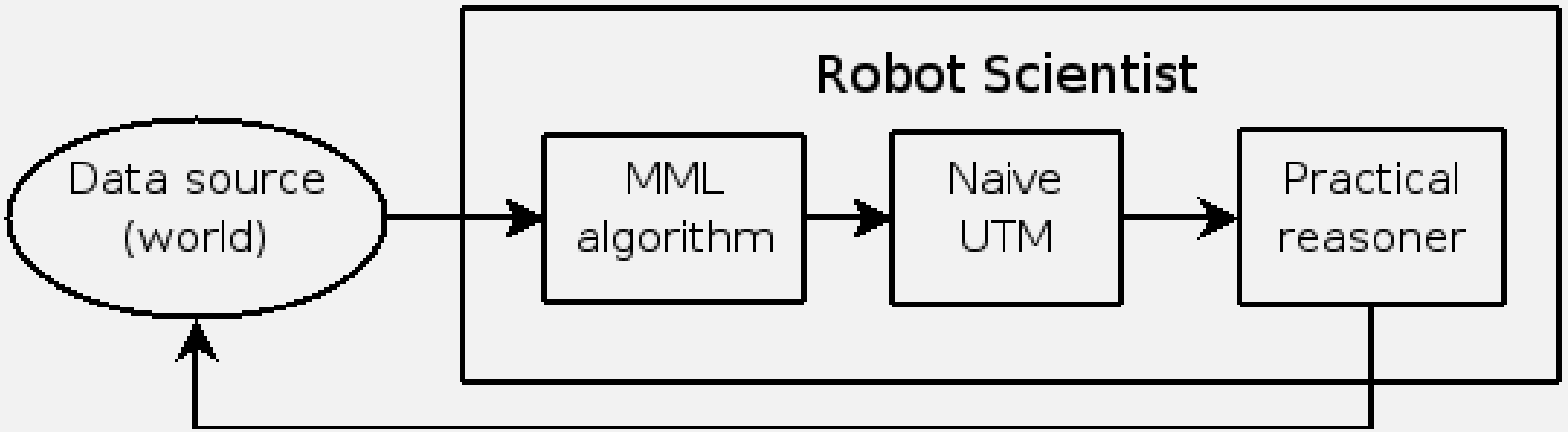
UTMs as uninformative Bayesian priors

- Explanation lengths will depend on choice of UTM
- But this difference is bounded above by a constant
- "The" UTM (up to a constant) as a *maximally uninformative prior*
- A step toward a *non-subjective* information?

Algorithmic science

- Wallace suggests we insist $U|h$ also be a UTM
- An "educated" UTM
  - Its implicit prior favors h
  - Can be further educated with *any* (computable) hypothesis
- "Normal" and "revolutionary" algorithmic science

The robot scientist



How is MML any better than Bayesian inference?

You might reasonably ask:

> "Why do we need MML? Why don't we just pick the hypothesis with the highest Bayesian posterior probability?"

UTMs as universal priors

- The "up to constant" part is rarely negligible
- Suppose UTM_2 emulates UTM_1 with just a 272-bit program
- $\text{P}(h|UTM_2) = 2^{-272} \cdot \text{P}(h|UTM_1)$

The notion of "information"

- Kevin Kelly is (rightly) concerned about abuses of "information"
- Ambiguity: Shannon-information *vs.* learning theory information
- Compare the ambiguity in a word's "representing" a thing, and a *thinker's* "representing" a thing
- The former is "derived", the latter "original"
- I suspect this is the *same* ambiguity

The null hypothesis

- One might get the impression that $\text{Len}_s(h) + \text{Len}_s(e|h)$ should be shorter than $\text{Len}_s(e)$
- Wallace encourages this with his translation to data compression
- Also with his discussion of the null hypothesis:

  > "We will require the length of the explanation message which states and uses an acceptable theory to be shorter than any message restating the data using only the implications of prior premises."

- But this is *never* the case!

Data compression is impossible?

- Suppose for contradiction

$$\begin{aligned} \text{Len}_s(h) + \text{Len}_s(e|h) &< \text{Len}_s(e) \\ -\log \text{P}(h) + -\log \text{P}(e|h) &< -\log \text{P}(e) \\ -\log \text{P}(h)\text{P}(e|h) &< -\log \text{P}(e) \\ -\log \text{P}(h, e) &< -\log \text{P}(e) \\ \text{P}(h, e) &> \text{P}(e) \end{aligned}$$

- Straightforwardly contradicts probability axioms
- There's some important disanalogy with data compression

Too much information

- Elsewhere Wallace acknowledges this
- The explanation will exceed encoded data length by $\text{Len}_s(h|e)$
- Still elsewhere, Wallace takes this as a *virtue* of MML
- $\text{Len}_s(e, h) > \text{Len}_s(e)$ because MML is *abductive*
- Why send this extra information, given the cost?

The advantage of MML over Bayes

- For *discrete* $\mathscr{H}$, MML = Bayesian inference
- MML has the advantage for continuous-valued $\mathscr{H}$
- Bayes: a posterior density, sensitive to parametrization of priors
- MML finds the right discretization $\mathscr{H}*$
- And the right $h \in \mathscr{H}*$

A classic example

Infer coin bias from a series of flips.

Table: Potential explanation lenghts (in *nits*) for 20 heads in 100

$h \in \mathscr{H}*$	$\lvert\mathscr{H}*\rvert$	$\text{Len}_s(h)$	$\text{Len}_s(e\lvert h)$	Total length
$\emptyset$	0	0	51.900	51.900
.20	100	4.615	50.040	54.655
.205	99	3.922	50.045	53.970
.50	1	0	69.315	69.315
.25	10	1.907	50.161	52.068

Bayes, MML, and Kolmogorov

- MML's real strength over Bayesian inference
- MML also gives approximations for Kolmogorov complexity
- A handy bridge between the two
- Neither MML nor Kolmogorov complexity are computable

The simplest UTM

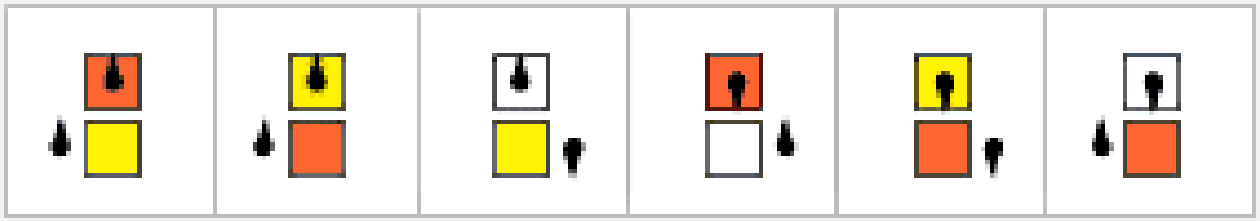
> "...the complexity of a TM is monotone increasing with its number of states. An overly simple TM, say one with only one or two states, cannot be universal. There must be, and is, a simplest UTM, or a small set of simplest UTMs. Adoption of such a UTM as receiver can reasonably be regarded as expressing no expectation about the theory or estimate to be inferred, save that it will be computable ... Adoption of any UTM (or TM) with more states seems necessary to assume something extra, i.e., to adopt a "less ignorant" prior. We therefore suggest that the only prior expressing total ignorance is that implied by a simplest UTM."

"The" simplest UTM

- Will still depend on the way of specifying UTMs
- Wallace says the "simplest" in any such should do
- Is that right?
- Would an educated "simplest" UTM use "green" and not "grue"?

The simplest UTM?

- Wallace may be wrong that a UTM "with only one or two states cannot be universal"
- Wolfram thinks this "two state, three color" one might be:



- $25,000 prize! (`http://wolframprize.org`)

Naturalized intentionality

- "Original information" as "original intentionality"
- *Naturalized theories of intentionality*

Millikan-Dretske

When it's a function of a creature to have an internal element covary with aspects of external circumstances, that element can be representational if such covariance is supposed to help satisfy a *need* of the creature.

Robot scientist representing

- Millikan: the "consumer-side" makes some symbol representational
- Our robot scientist could gain "original information" *via* MML
- (Assuming that inferring hypotheses meant to covary with a data source can help serve its needs)
- The priors in these needs?

Solomonoff's specter

- Again, why abduction?
- Could carry forward the entire posterior distribution over $\mathscr{H}$
- Solomonoff's "algorithmic probability" predictive strategy
- Wallace: Solomonoff will predict better, but

  > "MML attempts to mimic the discovery of natural laws, whereas Solomonoff wants to predict what will happen next with no explicit concern for understanding why."

- Solomonoff is deductive, MML ampliative

Why abduction?

- Again, why abduction?
- Especially given *cost* in message length and predictive power?
- Marcus Hutter's "universal AI" based on Solomonoff
- Hutter: artificial intelligence $\approx$ data compression
- €50,000 prize! (`http://prize.hutter1.net`)

Abduction and AI

- Hunch: AI as the problem of *lossy* compression
- Glean the important parts, toss the rest
- "Important" parts agent-relative
- Compare consumer-side intentionality
- Deductive methods remain badly intractable—maybe not coincidentally
- MML might help with these (related?) problems
- Abduction as "mere heuristics" of intelligent creatures

Summary

- MML is not
  - A source of non-subjective information
  - A clear, clean solution to the problem of the priors
  - A miraculous oracle of truth from simplicity
- MML is
  - A decent algorithm for Bayesian inference over continuous hypothesis spaces
  - A possible step in gaining "real" information
  - A nice bridge of Bayes and Kolmogorov
  - Part of an intriguing model for scientific progress
  - Some truth-related vindication of data compression (a form of simplicity)

Thanks

Thanks to:

- Dennis Whitcomb, Rutgers
- Shahed Sharif, Duke
- Ken Regan, SUNY Buffalo
- Alex Bertland, Niagara University
- Niagara University
- Branden Fitelson and FEW
- You