SCIENTIFIC INDUCTION IN PROBABILISTIC MATHEMATICS (BRIEF TECHNICAL NOTE)

JEREMY HAHN

This document is part of a collection of quick writeups of results from the December 2013 MIRI research workshop, written during or directly after the workshop. It describes work done by Paul Christiano, Jeremy Hahn, and others.

1. Priors over Mathematical Statements

Imagine a mathematician reasoning in the language $\mathcal{L}$ of first order logic, equipped with infinite stocks of constant, function, and relation symbols. The mathematician has finite memory, and keeps in its head only a subset S of the well-formed sentences in $\mathcal{L}$. For simplicity's sake, we will assume that S contains $\top$ (so our mathematician can reason about truth), and that S contains a sentence ϕ if and only if it also contains its negation $\neg\phi$. We wish to describe a function $\mathbb{P} : S \to [0, 1]$ that describes the mathematician's beliefs about the statements in S. The mathematician's beliefs should be **logically coherent**, meaning that

- $\mathbb{P}(\top) = 1$
- For $\phi, \psi \in S$, if $\phi \wedge \psi$ and $\phi \wedge \neg\psi$ are in S then $\mathbb{P}(\phi) = \mathbb{P}(\phi \wedge \psi) + \mathbb{P}(\phi \wedge \neg\psi)$.
- If ϕ and ψ are equivalent in first-order logic, and the proof can be written entirely with statements in S, then $\mathbb{P}(\phi) = \mathbb{P}(\psi)$.

Our mathematician believes a statement $\phi \in S$ to be true if $\mathbb{P}(\phi) = 1$. For example, our mathematician may believe with absolute certainity in the axioms of Peano Arithmetic. Coherence will then imply that simple consequences of those axioms are also be believed to be true, but the mathematician may be unsure of the truth of any statement whose proof requires wandering outside of S. Certainly, the mathematician may be unsure of any statement which is independent of the axioms of Peano Arithmetic.

The function $\mathbb{P}$ can be interpreted as the mathematician's **prior**. As new axioms are assumed, beliefs should change by updating $\mathbb{P}$ according to Bayes' rule. There are many possible coherent $\mathbb{P}$ that could serve as potential priors. A primary goal of the workshop was to determine the best possible prior $\mathbb{P}$, or at least to identify properties beyond coherence we might expect from a well-behaved prior. Very roughly, here are four properties we settled upon:

(1) The prior should be a computable function of S, and it should be possible to quickly approximate the result of conditioning on any new axiom.
(2) The prior should have good behavior as S varies. In particular, as $S \to \mathcal{L}$ through the addition of more and more sentences, $\mathbb{P}_S$ should tend towards a coherent limit independent of the order in which sentences were added.
(3) If $\phi \in S$ is some sentence of length ℓ, and logical coherence does not force $\mathbb{P}(\phi) = 0$, then $\mathbb{P}(\phi)$ should be at least $2^{-\ell}$. This condition ensures that the prior remains reasonably eclectic; we should avoid assigning negligibly small probabilities to any statement which might turn out to be true. From now on, we shall denote $2^{-\text{length}(\phi)}$ by $\mu(\phi)$.
(4) If one conditions on a statistical statement, such as that 90% of a sequence of statements $\phi(0), \phi(1), ..., \phi(N)$ are true, then $\mathbb{P}(\phi(c))$ should be approximately 0.9 for any specific $0 \leq c \leq N$ unless logical coherence demands otherwise. This last condition is both subtle and difficult to formalize; pinning it down formed the body of disscussion at the workshop.

2. Past Work on Logical Priors and the Failure of Property (4)

Abram Demski [REF] and Marcus Hutter [REF] have both proposed coherent priors $\mathbb{P}$. From our point of view, Demski's proposal is closest to satisfactory. Though he phrases his proposal only in the case $S = \mathcal{L}$, it is clear how to adapt his proposal to arbitrary S. In any case, the proposal is to follow the following process:

- Choose a random sentence from S, with the probability that ϕ is chosen proportional to $\mu(\phi) = 2^{-\text{length}(\phi)}$.
- Repeat the previous step indefinitely, with the caveat that no sentence should be chosen that contradicts the previously chosen sentences or logically follows from the previously chosen sentences. Halt if there are no longer any sentences to choose from.

Demski then declares $\mathbb{P}(\phi)$ to be the probability that ϕ is labeled true by the above random process. Demski's proposed prior manifestly satisfies properties (2) and (3) from the previous section. Though it is feasible to compute the $\mathbb{P}$ arising from Demski's scheme, it would seem difficult to do so quickly; it is also particularly unclear how one might rapidly perform approximate Baysian updates. More subtly, however, Demski's scheme does not seem to satisfy the desired property (4), a point first raised by Paul Christiano.

An example may serve to clarify the point. Let $\phi(x)$ denote a generic function symbol, so that $\phi(0), \phi(1), \phi(2), ...$ are independent atomic propositions logically unconstrained by any axioms. If we use a notion of length such that $\mu(\psi) = \mu(\neg\psi)$ for every $\psi \in S$, then for each n Demski's process is as likely to choose $\phi(n)$ as $\neg\phi(n)$. In other words, in Demski's prior $\mathbb{P}(\phi(n)) = 1/2$ for each n. Suppose, however, that we condition on the statement that exactly 90% of $\phi(1), \phi(2), ..., \phi(10^{100})$ are true. In this case, Demski's scheme will flip fair coins when assigning truth values to about the first $2 \cdot 10^{99}$ $\phi(i)$ it encounters, at which point it will notice that it must declare the remaining $8 \cdot 10^{99}$ $\phi(i)$ true. If Demski's scheme were encountering its first $2 \cdot 10^{99}$ $\phi(i)$ uniformly amongst $\phi(0), ..., \phi(10^{100})$, it would properly assign each $\phi(i)$ a probability of 0.9. Unfortunately, Demski's scheme does not uniformly encounter the $\phi(i)$, but rather is much more likely to encounter the $\phi(i)$ with larger μ first. In particular, even after conditioning on our statistical statement, $\mathbb{P}(\phi(0)) \approx 0.5$. Perhaps even more strangely, one could consider a number such as

$$c = \lfloor 7^{7^{7^7}} \cdot \pi \rfloor \pmod{10^{100}}.$$

Since c admits a relatively short description compared to a random element of $0, 1, ..., 10^{100}$, we can say that $\phi(c) \approx 0.5$. In other words, one can say that Demski's scheme is extremely confident that short statements are counterexamples to broad statistical statistical trends, even in the case that those simple statements are logically unrelated to anything else in the system. This seems like undesired behavior.

3. Some ideas for solutions

At the workshop, after pinning down the above example of undesired behaiviour we turned to other proposals for priors. None of the ideas presented are in a polished enough form to be considered a complete proposal, but we are very happy with the progress that was made. I will list a few of the most promising ideas below:

- Paul Christiano proposed that, from the perspective of computability, it might be best to describe a prior $\mathbb{P}$ as maximizing an entropy-like function. For example, a naive implementation of this idea would set the prior to be the coherent $\mathbb{P}$ maximizing
$$\sum_{\phi \in S} \mu(\phi)\, \mathbb{P}(\phi) \log(\mathbb{P}(\phi)).$$
- The most promising idea for solving (4) is to condition on sentences not of the form '90% of the $\phi(i)$ are true,' but rather of the form 'A random $\phi(i)$ chosen according to μ is true with probability 90%.' The difficulty with this approach is that μ is not accessible within the language itself. It seems possible, by adding new symbols to the language and adding to the definition of coherence, to essentially internalize μ.

- A final idea, which we currently consider less promising than the previous one, is to take advantage of compression. If $\phi(0), ..., \phi(N)$ is anything other that a sequence of independent propositions each with probability $1/2$, then it is possible to compress the sequence of $\phi(i)$ into a sequence $\psi(i)$ such that knowing all the $\psi(i)$ determines all the $\phi(i)$ but the $\psi(i)$ have shorter total length. This suggests that, if one chose groups of sentences at once, or if one chose sentences not only according to length but also with regard to mathematical expressivity, then the problem might work itself out automatically.