Computable probability distributions which converge on believing true Π_1 sentences will disbelieve true Π_2 sentences

Will Sawin, Abram Demski

July 2013

Abstract

It might seem reasonable that after seeing unboundedly many examples of a true Π_1 statement that a rational agent ought to be able to become increasingly confident, converging toward probability 1, that this statement is true. However, we have proven that this plus some plausible coherence properties, necessarily implies arbitrarily low limiting probabilities assigned to some short true Π_2 statements.

1 Introduction

Christiano et al. [2013] have investigated probability distributions over logical statements. Model theory translates between collections of axioms and statements, and models within which those statements are true or false. It has been established that, starting from a probability measure $\mathcal{P}(\mu)$ on models, we can go to a probability measure $\mathbb{P}(\ulcorner\phi\urcorner)$ on logical formulas ϕ which are true or false in those models. This distribution $\mathbb{P}$ will obey desirable coherence constraints such as that if $A \to B$ is a tautology then $\mathbb{P}(\ulcorner A\urcorner) \leq \mathbb{P}(\ulcorner B\urcorner)$. Conversely, starting with a probability distribution on formulas which obeys the same coherence constraints, we can go to a probability measure on models by playing a martingale with formulas selected with their conditional probability given previous formulas, until a complete theory is obtained.

As a concrete example, suppose that, inspired by Solomonoff induction, we would like formulas or axioms $F \in \mathcal{F}$ with length $|F|$ in some alphabet $\mathcal{A}$ to

have probability at least equal to $2^{-|F|}$. Then we can obtain a corresponding probability distribution over formulas, and hence over models, by beginning with the empty formula and iteratively adjoining well-formed formulas randomly generated from $\mathcal{A}$, discarding formulas which yield inconsistencies. Demski [2012] proposes a prior $\mathbb{P}_L$ along these lines and suggests a sampling procedure to converge to $\mathbb{P}$ in the limit.

Algorithm 1 gives this computable procedure, slightly adapted from the published version. In this algorithm, gens() is a function which generates a random sentence according to the probability distribution over formulas, and $\text{con}(\sigma, t)$ is a function which returns true if there is not a proof of contradiction of length t or less from the set σ of sentences, and false otherwise.

Algorithm 1: Sampling procedure searching proofs up to length t

```
Function SAMPLE(t, φ):
    S ⟵ []
    loop
        push(gens(), S)            /* Add a random sentence to S */
        y ⟵ con(S ∪ {φ}, t)
        n ⟵ con(S ∪ {¬φ}, t)
        if ¬y ∧ ¬n then the random sentence was inconsistent
            pop(S)
        if y ∧ ¬n then φ holds in this sample
            return true
        if ¬y ∧ n then ¬φ holds in this sample
            return false
```

Roughly, as we increase t, Algorithm 1 spends longer and longer checking for contradictions and so converges on the true probability distribution in the limit.

By adjoining e.g. the axioms of first-order arithmetic ($\mathcal{PA}$) to the initial empty set S of sentences we can obtain a probability distribution $\mathbb{P}_{\mathcal{PA}}$ which assumes $\mathcal{PA}$, and hence obtain probabilities over logical sentences about first-order arithmetic. Probabilities of $\mathcal{PA}$ theorems will be 1, but statements F not proven or disproven by first-order arithmetic will have probability at least $2^{-|F|}$ since they have at least this probability of being added immediately.

The limit of SAMPLE(t, ϕ) starting with $\mathcal{PA}$ will assign positive probability to Σ_1 statements S which are false in the standard natural numbers

$\mathbb{N}$, or fail to converge to 1 for Π_1 statements true in $\mathbb{N}$, even in the limit of seeing unboundedly many negative (resp. positive) examples of S and no positive (resp. negative) examples of S. This is because SAMPLE has a fixed positive probability of adding S at the start of its exploration. This corresponds to assigning positive probability to nonstandard models in which the Σ_1 statement S is true.

It may seem sensible at first that a rational agent ought to be able to converge toward limiting probabilities of 1 for true Π_1 statements (probabilities of 0 for false Σ_1 statements) after seeing unboundedly many confirmations, via scientific induction and probabilistic reasoning. For example, Hutter et al. [2013] demands this (via the "Gaifman condition") in his logical prior. However, Π_1 convergence is not without downsides, so great as to argue strongly against Π_1 convergence as a desirable property for rational agents. (The alternative being that rational agents should behave similarly to the original *sample* distribution, and not perform scientific induction of unprovable Π_1 statements with probabilistic confidence approaching 1 even in the limit of an infinite number of confirming examples.)

We will demonstrate that convergence on Π_1 truths implies bad behavior with respect to assigning probabilities to statements in Π_2. In particular, Π_1 convergence along with basic coherence properties implies that some true statements in Π_2 will be assigned probability zero in the limit. (Contrast to the limiting distribution of SAMPLE where short true Π_2 statements S will never have probability falling below $2^{-|S|}$.)

2 Main result

Let $P : \mathbb{N}\times\mathbb{N} \to [0,1]$ be a function, where $P(t, \ulcorner\phi\urcorner)$ represents the probability that an agent assigns to the statement ϕ at time t (where $\ulcorner\phi\urcorner$ is the Gödel number of ϕ). So that an agent can actually determine the probabilities it assigns, we require that each bit of $P(t, \ulcorner\phi\urcorner)$ is a computable function of t and $\ulcorner\phi\urcorner$.

Theorem 1. *$P(t, \ulcorner\phi\urcorner)$ cannot satisfy all three of the following properties:*

1. **Coherence:** *If $\phi \implies \neg\psi$, then*

$$\limsup_{t\to\infty} P(t, \ulcorner\phi\urcorner) + P(t, \ulcorner\psi\urcorner) \leq 1.$$

2. **Scientific induction:** *If ϕ is a true Π_1 sentence, then*

$$\lim_{t\to\infty} P(t, \ulcorner\phi\urcorner) = 1.$$

3. Π_2 **open mindedness:** *If ϕ is a true Π_2 statement, then*

$$\liminf_{t\to\infty} P(t, \ulcorner\phi\urcorner) > 0.$$

Proof. First note that conditions 1 and 2 imply that if ϕ is a false Π_2 sentence, then $\lim_{t\to\infty} P(t, \ulcorner\phi\urcorner) = 0$. This is simply because any false Π_2 sentence must have some counterexample which is a true Π_1 sentence. If ψ is this counterexample, then $\lim_{t\to\infty} P(t, \ulcorner\psi\urcorner) = 1$, and $\liminf_{t\to\infty} P(t, \ulcorner\phi\urcorner) + P(t, \ulcorner\psi\urcorner) \leq 1$, so $\lim_{t\to\infty} P(t, \ulcorner\phi\urcorner) = 0$.

So if there were a function $P(t, \ulcorner\phi\urcorner)$ satisfying all these conditions, then each Π_2 statement ϕ would be equivalent to the statement $\liminf_{t\to\infty} P(t, \ulcorner\phi\urcorner) > 0$. This is in fact a Σ_2 statement: $\exists n, b : \forall t \geq n : P(t, \ulcorner\phi\urcorner) \geq 1/2^b$. So every Π_2 statement is equivalent to a Σ_2 statement and is thus a Δ_2 statement, but the inclusion $\Delta_2 \subset \Pi_2$ is known to be strict, so this is a contradiction. □

3 Interpretation

A system with Π_1 convergence assigns probabilities approaching 1 to Π_1 statements for which it has seen sufficiently many examples. A Π_2 statement is composed of infinitely many Σ_1 statements, each of which is a negated Π_1 statement. By the time we have observed the truth of the first trillion Σ_1 statements, at the end of sufficient time, there may be some further Σ_1 statement for the trillion-and-first case, where the corresponding scientific induction on its negated Π_1 counterpart tells us to be extremely confident that no example will ever be found. By the time the trillion-and-first positive example of the Σ_1 statement has been found, we are even more confident of the untruth of the trillion-and-second statement. Thus attempting to create a probability distribution which performs scientific induction on Π_1 statements, converging to probability 1 for the true versions of such statements, can create zero limiting probabilities assigned to true Π_2 statements.

By our main result, this effect is inevitable.

References

Paul Christiano, Eliezer Yudkowsky, Marcello Herreshoff, and Mihaly Barasz. Definability of truth in probabilistic logic. `http://intelligence.org/files/DefinabilityTruthDraft.pdf`, 2013.

Abram Demski. Logical prior probability. In *Artificial General Intelligence*, pages 50–59. Springer, 2012.

Marcus Hutter, John W. Lloyd, Kee Siong Ng, and William T. B. Uther. Probabilities on sentences in an expressive logic. *Journal of Applied Logic*, 11(4):386–420, 2013.