Procrastination in probabilistic logic

Benja Fallenstein

Background: Reflection in probabilistic logic Christiano et al. [**?**] have proposed a way of working around Tarski's undefinability of truth by assigning *probabilities* (rather than truth values) to logical statements, and requiring not that their system knows these probability *exactly*, but only that it knows them up to an infinitesimal error. Their system can be characterized as follows. Let $\mathcal{L}$ be the language of set theory extended by a single predicate symbol P, and $\mathcal{S}$ the Stone space of complete theories in this language (we take "complete" to imply "consistent"). Write $\Delta(\mathcal{S})$ for the set of probability distributions over $\mathcal{S}$ endowed with the Borel-σ-algebra. For $\mathbb{P} \in \Delta(\mathcal{S})$ and S a theory in the language $\mathcal{L}$, in the language $\mathcal{L}$, write $\mathbb{P}[S] := \mathbb{P}(\{T : S \subseteq T\})$ for the probability of the set of all complete theories extending S; if φ is a sentence in $\mathcal{L}$, we write $\mathbb{P}[\varphi] := \mathbb{P}[\{\varphi\}]$ and call this *the probability of* φ. Christiano et al.'s proof can be used to show that there is a $\mathbb{P} \in \Delta(\mathcal{S})$ such that (i) $\mathbb{P}[\text{ZFC}] = 1$; (ii) $\mathbb{P}[P \in \Delta(\mathcal{S})]$; and (iii) the following *reflection principle*: For all sentences φ in $\mathcal{L}$ and all $a, b \in \mathbb{Q}$, if $a < \mathbb{P}[\varphi] < b$, then $\mathbb{P}[a < P[\varphi] < b] = 1$.

Thus, for any arbitrarily small open interval I around the true value of $\mathbb{P}[\varphi]$, the value of $P[\varphi]$ almost surely lies inside I. However, this does not mean that $\mathbb{P}[\varphi] = a$ implies $\mathbb{P}[P[\varphi] = a] = 1$ for all rational a: It is possible for a complete theory T to contain the statement $|P[\varphi] - a| < \delta$ for all rational $\delta > 0$, and also the statement $|P[\varphi] - a| > 0$. In this case, we say that $|P[\varphi] - a| > 0$ is a (nonzero) *infinitesimal*, and it follows that T can have only nonstandard models. Christiano et al. show that in their system, the set of all complete theories that have standard models is a null set: $\mathbb{P}$-almost all complete theories have only nonstandard models.

Background: The procrastination paradox This reliance on nonstandard numbers may seem suspicious, but it is not immediately obvious whether or not it causes problems in applications of the system. In this report, I show an analog in Christiano et al.'s system of a result about nonstandard theories of arithmetic called the *procrastination paradox* [**?**]. Consider an immortal rational agent trying to decide whether to do a certain necessary task today, or postpone it till tomorrow; we imagine that it is very important that this task get done *eventually*, but it does not matter *when* the task is done. In the original form of the procrastination paradox, we imagine that our agent will do the task today *unless it can be proven in a certain formal system that it will do*

the task at a later time, in which case our agent decides that for the moment, it is alright to procrastinate.

Informally, the problem arises if the agent trusts the reasoning of future versions of itself too much; then, it reasons that tomorrow, it will either press the button or show that the button gets pressed at a later time, implying that in either case the button gets pressed, and it is therefore not necessary to press the button today. The agent reasons the same way at all future times, meaning that the button never actually gets pressed.

Formally, the original procrastination paradox shows that if a recursive sequence T_n of first-order theories extending Peano Arithmetic satisfies $T_n \vdash \Box_{T_{n+1}}(\varphi) \to \varphi$ for every $n \in \mathbb{N}$ and every sentence φ, where $\Box_{T_{n+1}}(\varphi)$ is the statement that φ is provable in T_{n+1}, then all T_n have only nonstandard models. The proof uses the diagonal lemma to define a predicate $\varphi(n) :\leftrightarrow \neg\Box_{T_n}\big(\exists k > n\colon \varphi(k)\big)$, where we interpret $\varphi(n)$ to mean that the agent presses the button at time n; in other words, the agent presses the button iff it cannot find an argument that in T_n that the button will get pressed at some later time.[1] Then, T_n proves that $\varphi(n+1) \vee \neg\varphi(n+1)$, and that $\neg\varphi(n+1)$ is equivalent to $\Box_{T_{n+1}}\big(\exists k > n+1\colon \varphi(k)\big)$, which (according to T_n's special axiom) implies $\exists k > n+1\colon \varphi(k)$; hence, T_n proves $\varphi(n+1) \vee \exists k > n+1\colon \varphi(k)$, which immediately yields $T_n \vdash \exists k > n\colon \varphi(k)$. In other words, $\mathbb{N} \vDash \forall n\colon \Box_{T_n}\big(\exists k > n\colon \varphi(k)\big)$. But this is equivalent to $\mathbb{N} \vDash \forall n\colon \neg\varphi(n)$; in other words, all T_n are unsound on the standard model. However, the T_n can nevertheless be *consistent* (see [**?**]), implying that they have *nonstandard* models $\mathcal{M}_n$ which satisfy $\mathcal{M}_n \vDash \exists k > n\colon \varphi(k)$; thus, these models contain *nonstandard* "times" k at which the button does in fact get pressed.

Procrastination in probabilistic logic We now formulate an analog of this argument in Christiano et al.'s system. To do so, we fix a definable sequence of rational numbers $\varepsilon_n > 0$ such that $\varepsilon_n \to 0$ (for example, we may set $\varepsilon_n := 2^{-n}$), and consider an agent that will press the button at time n if the probability that it will get pressed at some later time is less than $1 - \varepsilon_n$. Formally, we use the diagonal lemma to define a predicate $\varphi(n)$ such that ZFC $\vdash \varphi(n) \leftrightarrow P[\exists k > n\colon \varphi(k)] < 1 - \varepsilon_n$.

The following theorem can be interpreted as saying that Christiano et al.'s system believes for certain that it will press the button eventually.

Theorem 1. *For all $n \in \mathbb{N}$, $\mathbb{P}[\exists k > n\colon \varphi(k)] = 1$.*

Proof. Suppose not. Then there is an $n_0 \in \mathbb{N}$ such that $\mathbb{P}[\exists k > n_0\colon \varphi(k)]$ is less than one, and therefore less than $1 - \varepsilon_n$ for some $n > n_0$. Since $\exists k > n_0\colon \varphi(k)$ is implied by $\exists k > n\colon \varphi(k)$, the former statement must be at least as probable as the latter one, i.e., $1 - \varepsilon_n > \mathbb{P}[\exists k > n_0\colon \varphi(k)] \geq \mathbb{P}[\exists k > n\colon \varphi(k)]$, and hence by the reflection principle

$$\mathbb{P}[\varphi(n)] \;=\; \mathbb{P}[P[\exists k > n\colon \varphi(k)] < 1 - \varepsilon_n] \;=\; 1.$$

But this implies $\mathbb{P}[\exists k > n_0\colon \varphi(k)] = 1$, contradicting our assumption that this probability is less than one. □

[1] Since provability is only semidecidable, this means that our agent is not computable; we could consider computable agents at the expense of some additional formalism, but will stick to the simple version here. (Note to self: Actually work out a computable version in detail one of these days.)

The following result can be interpreted as saying that even though the system is certain that it will eventually press the button, it does not in fact do so.

Theorem 2. *For all* $n \in \mathbb{N}$, $\mathbb{P}[\varphi(n)] = 0$.

Proof. By the previous theorem, $\mathbb{P}[\exists k > n\colon \varphi(k)] = 1 > 1 - \epsilon_n$, and hence

$$\mathbb{P}[\varphi(n)] \;=\; \mathbb{P}[P[\exists k > n\colon \varphi(k)] < 1 - \epsilon_n] \;=\; 0.$$

□

These two results are not contradictory: They merely imply that $\mathbb{P}$ is supported on complete theories whose (nonstandard) models believe that the button gets pressed *at some nonstandard time*, just as in the procrastination paradox for logical theories.