PARAMETRIC BOUNDED VERSION OF LÖB'S THEOREM

PATRICK STEVENS

https://www.patrickstevens.co.uk/misc/ParametricBoundedLoeb2016/ParametricBoundedLoeb2016.pdf

1. Introduction

I was recently made aware of a preprint[1] of a paper which proves a bounded version of Löb's Theorem.

Theorem 1.1 (Parametric Bounded Löb)**.** *If $\Box A$ is the operator "there exists a proof of A in Peano arithmetic" and $\Box_k A$ is the operator "there exists a proof of A in k or fewer lines in Peano arithmetic", then for every formula p of one free variable in the language of PA, and every computable $f : \mathbb{N} \to \mathbb{N}$ which grows sufficiently fast, it is true that*

$$(\exists \hat{k})[(\vdash [\forall k][\Box_{f(k)} p(k) \to p(k)]) \Rightarrow (\vdash [\forall k > \hat{k}][p(k)])]$$

(Colour is used only to emphasise logical chunks of the formula.)

The paper gives plenty of motivation about why this result should be interesting and useful: section 6 of the paper, for instance, is an application to the one-shot Prisoner's Dilemma played between agents who have access to each other's source code. However, I believe that while the theorem may be true and the proof may be correct, its application may not be as straightforward as the paper suggests.

2. Background

Theorem 2.1 (Löb's Theorem)**.** *Suppose $\Box\ulcorner A\urcorner$ denotes "the formula A with Gödel number $\ulcorner A\urcorner$ is provable". If*

$$\text{PA} \vdash (\Box\ulcorner P\urcorner \to P)$$

then

$$\text{PA} \vdash P$$

Löb's Theorem is at heart a statement about the incompatibility of the interpretation of the box as "provable" with the intuitively plausible deduction rule that $\Box\ulcorner P\urcorner \to P$. ("If we have a proof of P, then we can deduce P!") The Critch paper has an example in Section 1.4 where P is the Riemann hypothesis.

Date: 24th July 2016.

3. Problem with the paper

Suppose $\mathcal{M}$ is a model of Peano arithmetic, in which our agent is working. It is a fact of first-order logic (through the Löwenheim-Skolem theorem) that there is no first-order way of distinguishing any particular model of PA. Therefore the model of PA could be non-standard; this is not something a first-order reasoning agent could determine.

If the agent is working with a non-standard model of PA, then all the theorems of the Critch paper may well go through. However, they become substantially less useful, as follows.

Let us write M for the underlying class (or set) of the model $\mathcal{M}$ of PA. Then the statement

$$(\exists \hat{k})[(\vdash [\forall k][\Box_{f(k)} p(k) \to p(k)]) \Rightarrow (\vdash [\forall k > \hat{k}][p(k)])]$$

when relativised to the model $\mathcal{M}$ becomes

$$(\exists \hat{k} \in M)[(\vdash [\forall k \in M][\Box^{\mathcal{M}}_{f(k)} p(k) \to p(k)]) \Rightarrow (\vdash [\forall k \in M^{>\hat{k}}][p(k)])]$$

where $\Box^{\mathcal{M}}_{f(k)}$ is now shorthand for "there is a proof-object P in M such that P encodes a M-proof of $p(k)$ which is fewer than $f(k)$ lines long".

Notice that the quantifiers have been restricted to M; in particular, $\hat{k}$ might be a non-standard natural number. Likewise, the "there is a proof" predicate is now "there is an object which M unpacks into a proof"; but such objects may be non-standard naturals themselves, and unpack into non-standard proofs (which $\mathcal{M}$ still believes are proofs, because it doesn't know the difference between "standard" and "non-standard").

3.1. **Aside: non-standard proof objects.**

What is a non-standard proof object? Let's imagine we have some specific statements a_i for each natural i such that $a_i \to a_{i+1}$ for each i, and such that a_0 is an axiom of PA. I'm using a_i only for shorthand; the reader should imagine I had some specific statements and specific proofs of $a_i \to a_{i+1}$.

Consider the following proof of a_2:

(1) a_0 (axiom)
(2) a_1 (by writing out the proof of $a_0 \to a_1$ above this line)
(3) a_2 (by writing out the proof of $a_1 \to a_2$ above this line)

If we take a simple Gödel numbering scheme, namely "take the number to be an ASCII string in base 256", it's easy to see that this proof has a Gödel number. After all, we're imagining that I have specific proofs of $a_i \to a_{i+1}$, so I could just write them in. Then you're reading this document which was originally encoded as ASCII, so the Gödel numbering scheme must have worked.

Similarly, there is a Gödel number corresponding to the following:

(1) a_0 (axiom)
(2) a_1 (by writing out the proof of $a_0 \to a_1$ above this line)
(3) ...
(4) a_k (by writing out the proof of $a_{k-1} \to a_k$ above this line)

Now, suppose we're working in a non-standard model, and fix non-standard K. Then there is a (probably non-standard) natural L corresponding to the following proof:

(1) a_0 (axiom)
(2) a_1 (by writing out the proof of $a_0 \to a_1$ above this line)

(3) ...
(4) a_K (by writing out the proof of $a_{K-1} \to a_K$ above this line)

Now, this is not a "proof" in our intuitive sense of the word, because from our perspective it's infinitely long. However, the model still thinks this is a proof, and that it's coded by the (non-standard) natural L.

3.2. **Implication for PBL.** So the model $\mathcal{M}$ believes there is a natural $\hat{k}$ such that ...But if that natural is non-standard (and remember that this is not something the model can determine without breaking into second-order logic!) then PBL doesn't really help us. It simply tells us that all sufficiently-large non-standard naturals have a certain property; but that doesn't necessarily mean any standard naturals have that property. And the application to the Prisoners' Dilemma in Critch's paper requires a standard finite $\hat{k}$.

If we, constructing the agent Fairbot, could somehow guarantee that it would be working within the standard model of PA, then all would be well. However, we can't do that within first-order logic. It could be the case that when constructing Fairbot, the only sufficiently-large naturals turn out to be non-standard. When we eventually come to run $\text{Fairbot}_k(\text{Fairbot}_k)$, it could therefore be that it will take nonstandardly-many proof steps to discover the "(coooperate, cooperate)" outcome. In practice, therefore, the agents would not find that outcome: we can only run them for standardly-many steps, and all non-standard naturals look infinite to us.

4. Acknowledgements

My thanks are due to Miëtek Bak (who persuaded me that there might be a problem with the article) and to John Aspden (who very capably forced Miëtek to clarify his objection until I finally understood it). As ever, any mistakes in this article are due only to me.

References

[1] Andrew Critch, *Parametric Bounded Löb's Theorem and Robust Cooperation of Bounded Agents*, http://arxiv.org/abs/1602.04184v4