A new perspective on the arithmetical completeness of GL [*]

Paula Henk

Abstract

Solovay's proof of the arithmetical completeness of the provability logic GL proceeds by simulating a finite Kripke model inside the theory of Peano Arithmetic (PA). In this article, a new perspective on the proof of GL's arithmetical completeness will be given. Instead of simulating a Kripke structure inside the theory of PA, it will be embedded into an arithmetically defined Kripke structure. We will examine the relation of strong interpretability, which will turn out to have exactly the suitable properties for assuming the role of the accessibility relation in a Kripke structure whose domain consists of models of PA.

Given any finite Kripke model for GL, we can then find a bisimilar model whose nodes are certain nonstandard models of PA. The arithmetical completeness of GL is an immediate consequence of this result. In order to define the bisimulation, however, and to prove its existence, the most crucial and ingenious ingredients of Solovay's original proof are needed. The main result of the current work is thus not so much a new proof as a new perspective on an already known proof.

1 Introduction: PA and GL

Peano arithmetic (PA) is a first-order theory in the language $\mathcal{L} = \{0, S, +, \cdot\}$, where 0 is a constant, S a unary function symbol, and $+$, $\cdot$ binary function symbols[1]. PA contains six axioms and one axiom schema, the original purpose of which was to capture the structure of the natural numbers. However, PA fails to achieve this goal, as it is not even ω-categorical: also (countable) structures *other* than the natural numbers satisfy the axioms of PA. These so-called nonstandard models will play an important role in the current article.

The natural numbers, where the non-logical constants are interpreted in the obvious way, are of course among the models of PA - in fact they form an initial segment of any model of PA. This justifies the usual interpretation of PA as proving statements about natural numbers. Since the latter can be used to code (via gödelnumbering) syntactic objects of PA, like e.g. proofs, PA is able to

[*] This is a report on an individual project supervised by Albert Visser and Dick de Jongh.

[1] Throughout this article it will be easier to assume that we are working with a relational language. See p.4 in [3] for a treatment of function symbols in a relational language.

engage in the self-reflective activity of "speaking about" (i.e. proving sentences about) provability in itself. In particular, there is a predicate $\mathrm{Bew}_{\mathrm{PA}}(x)$ of PA which expresses the provability relation in PA. As a consequence of this, we have, for all sentences φ and ψ of $\mathcal{L}$:

1. if $\mathrm{PA} \vdash \varphi$, then $\mathrm{PA} \vdash \mathrm{Bew}_{\mathrm{PA}}(\overline{\ulcorner\varphi\urcorner})$, where $\ulcorner\varphi\urcorner$ stands for the code of φ, and $\overline{\ulcorner\varphi\urcorner}$ for the numeral[2] of this code.
2. $\mathrm{PA} \vdash \mathrm{Bew}_{\mathrm{PA}}(\overline{\ulcorner\varphi \to \psi\urcorner}) \to \big(\mathrm{Bew}_{\mathrm{PA}}(\overline{\ulcorner\varphi\urcorner}) \to \mathrm{Bew}_{\mathrm{PA}}(\overline{\ulcorner\psi\urcorner})\big)$
3. $\mathrm{PA} \vdash \mathrm{Bew}_{\mathrm{PA}}(\overline{\ulcorner\varphi\urcorner}) \to \mathrm{Bew}_{\mathrm{PA}}\left(\overline{\ulcorner\mathrm{Bew}_{\mathrm{PA}}(\overline{\ulcorner\varphi\urcorner})\urcorner}\right)$

These self-reflective capacities of PA, the so-called Hilbert-Bernays-Löb derivability conditions, are in fact crucial for proving the Second Incompleteness Theorem. According to the second, PA "knows" that modus ponens is among its rules of inference. According to the third, PA "knows" that it has the self-reflective capacity stated in 1.

But the self-reflective capacities of PA also have a limit. For example, we might expect PA to "know" that if it can prove a sentence φ, then φ holds, i.e. $\mathrm{PA} \vdash \mathrm{Bew}_{\mathrm{PA}}(\overline{\ulcorner\varphi\urcorner}) \to \varphi$. However, according to Löb's Theorem, PA proves this statement only for sentences which it already can prove. Intuitively, this means that PA does not engage in counterfactual speculations about its ability to prove theorems. If PA does not know that a sentence is true (i.e. $\mathrm{PA} \nvdash \varphi$), then it also will not claim that the sentence *would* be true in case it would be provable (i.e. then also $\mathrm{PA} \nvdash \mathrm{Bew}_{\mathrm{PA}}(\overline{\ulcorner\varphi\urcorner}) \to \varphi$). It is therefore a reasonable question to ask whether there is a way to characterise provable sentences in PA that involve the provability predicate.

As it turns out, the self-reflective capacities of PA can be neatly analysed by using modal logic. In order to do that, the operator $\Box$ is interpreted as "it is provable in PA that ..." - as opposed to the more traditional "it is necessary that ...". More specifically, the analysis requires us to establish a correspondence between modal sentences and sentences of $\mathcal{L}$. For this purpose, we introduce the notion of a realisation - an assignment of sentences of PA to propositional letters in the modal language. Given a realisation $*$, a sentence φ of the modal language is translated into a sentence φ^* of PA according to the following recursive definition:

1. $p^* = \varphi_p^*$
2. $\bot^* = \bot$
3. $(\varphi \to \psi)^* = (\varphi^* \to \psi^*)$
4. $\Box\varphi^* = \mathrm{Bew}_{\mathrm{PA}}\left(\overline{\ulcorner\varphi^*\urcorner}\right)$

[2] For each natural number n, its numeral $\overline{n}$ is the closed term of $\mathcal{L}$ denoting it, i.e. 0, preceeded by n applications of S. E.g. the numeral of 5 is $SSSSS0$

where φ_p^* denotes the $\mathcal{L}$-sentence assigned to the sentence letter p by the realisation $*$.

There is in fact a modal system that is perfectly suited for characterising formal provability in PA. GL (named after Gödel and M.H. Löb) is obtained by adding to K the Löb axiom $\Box(\Box\varphi \rightarrow \varphi) \rightarrow \Box\varphi$; this system is sound and complete with respect to frames that are finite transitive and irreflexive trees (equivalently, transitive and converse well-founded frames).

Solovay's proofs of the arithmetical soundness and completeness of the modal system GL show that GL is *the* logic of formal provability in PA in the following sense:

- If GL $\vdash \varphi$, then PA $\vdash \varphi^*$ for all realisations $*$ (Arithmetical soundness).
- If PA $\vdash \varphi^*$ for all realisations $*$, then GL $\vdash \varphi$ (Arithmetical completeness).

GL's arithmetical completeness and soundness thus means that the translations of GL's theorems are exactly the sentences that are always provable in PA, i.e. regardless of which sentences are assigned to the propositional letters.

In section 3, we will give a brief outline of Solovay's proof of the arithmetical completeness of GL. By making use of certain sentences constructed in this proof, we are able to show the existence of a bisimulation between a finite Kripke structure for GL, and the Big Model $\mathfrak{N}$ whose domain consists of certain nonstandard models of PA. This result yields an alternative proof for the arithmetical completeness of GL. Section 2 contains the preliminary work that is needed to construct the Big Model $\mathfrak{N}$ in the first place. In particular, we establish the existence of a relation that holds between models of PA, and is suitable for assuming the role of the accessibility relation in $\mathfrak{N}$.

2 Constructing the Big Model

This section introduces the relation of strong interpretability on the class of models of PA. We will establish that this relation has certain properties which entitles us to view it as an accessibility relation, in the sense of Kripke structures, between models of PA.

2.1 Interpretability

The notion of interpretability allows us to compare first-order theories. Intuitively, if a theory T interprets a theory S then T is at least as strong as S.

Definition Let S and T be first-order theories in the languages $\mathcal{L}_S$ and $\mathcal{L}_T$ respectively. A function I from the formulas of $\mathcal{L}_S$ to formulas of $\mathcal{L}_T$ is an interpretation of S and T if

1. for each predicate symbol[3] P of S, there is a formula ψ_P of T, and for all variables x and y, $I\left(P\left(x,y\right)\right) = \psi_P\left(x,y\right)$. [4]
2. I commutes with the propositional connectives: for any formulas φ and ψ of S,
   (a) $I\left(\neg\varphi\right) = \neg I\left(\varphi\right)$
   (b) $I\left(\varphi \rightarrow \psi\right) = I\left(\varphi\right) \rightarrow I\left(\psi\right)$
3. I commutes with the quantifiers in a relativized sense: there is a formula $\delta(x)$ of T containing one free variable, such that for any variable y and any formula φ of S, $I(\exists y\varphi) = \exists y\left(\delta(y) \wedge I(\varphi)\right)$. Furthermore, $T \vdash \exists x\delta\left(x\right)$.
4. For all axioms φ of S, $T \vdash I\left(\varphi\right)$.

Given that $S \vdash \varphi$, and T interprets S via I, it follows that $T \vdash I\left(\varphi\right)$, i.e. T proves the translations of all theorems of S.

An important application of the notion of interpretability are proofs of relative consistency. If T interprets S, then T's consistency implies S's consistency. If $S \vdash \bot$, then $T \vdash I\left(\bot\right)$, and hence, because I commutes with the propositional connectives, also $T \vdash \bot$.

The notion of interpretation can also be viewed from a semantic perspective. Given a model $\mathcal{N}$ of a theory T, an interpretation I of S in T allows the construction of a model $\mathcal{M}$ of S inside $\mathcal{N}$. The domain of $\mathcal{M}$ will be the subset $\mathrm{dom}\left(\mathcal{M}\right) := \{m\epsilon\mathrm{dom}(\mathcal{N}) \mid \mathcal{N} \vDash \delta(m)\}$. Since $T \vdash \exists x\delta\left(x\right)$, $\mathrm{dom}\left(\mathcal{M}\right) \neq \emptyset$. Given a formula $\varphi\left(x_1, \ldots, x_n\right)$ of S, and elements $a_1, \ldots, a_n\epsilon\mathrm{dom}\left(\mathcal{M}\right)$ we define:

$$\mathcal{M} \vDash \varphi\left[a_1, \ldots, a_n\right] :\Leftrightarrow \mathcal{N} \vDash I\left(\varphi\right)\left[a_1, \ldots, a_n\right]$$

where $I\left(\varphi\right)\left(x_1, \ldots, x_n\right)$ is the formula of T assigned to $\varphi\left(x_1, \ldots, x_n\right)$ by I. If $\mathcal{M}$ is constructed in such a way from $\mathcal{N}$, we say that $\mathcal{M}$ is an *internal model* of $\mathcal{N}$, or internal to $\mathcal{N}$.

[3] We will treat the case that P is a binary relation symbol. The other cases are similar.

[4] For the sake of simplicity, we will require the equality symbol to be interpreted as the equality symbol, although in the general case this is not necessary.

Requirements for an accessibility relation

In this article, the notion of internal models will be used to shed new light on the proof of the arithmetical completeness of the modal logic GL. Given some modifications, the relation between a model and its internal models will turn out to have exactly the right properties for assuming the role of the accessibility relation in a Kripke structure whose domain consists of models of PA.

The box of modal logic is translated into the language of PA as the provability predicate $\mathrm{Bew}_{\mathrm{PA}}$. In Kripke models, we have the following truth definition for modal formulas that have the form $\Box\varphi$:

$$\mathfrak{M}, i \vDash \Box\varphi \iff \text{for all } j \text{ s.t. } iRj\,, \mathfrak{M}, j \vDash \varphi$$

In order to construct a Kripke model whose domain consists of models of PA, we must thus find a relation $\preccurlyeq$ on models of PA s.t.

$$\mathcal{N} \vDash \mathrm{Bew}_{\mathrm{PA}}(\overline{\ulcorner\varphi\urcorner}) \iff \text{for all } \mathcal{M} \text{ s.t. } \mathcal{M} \preccurlyeq \mathcal{N}, \mathcal{M} \vDash \varphi$$

Since the relation of interpretability is reflexive (we can simply take the identity function for I), each model is internal to itself. The above requirement would thus impose that for any model $\mathcal{N}$ of PA and any sentence φ, $\mathcal{N} \vDash \mathrm{Bew}_{\mathrm{PA}}(\overline{\ulcorner\varphi\urcorner})$ implies $\mathcal{N} \vDash \varphi$. However, consider the Gödel sentence G which is a fixed point of $\neg\mathrm{Bew}_{\mathrm{PA}}(x)$:

$$\mathrm{PA} \vdash G \leftrightarrow \neg\mathrm{Bew}_{\mathrm{PA}}\left(\overline{\ulcorner G\urcorner}\right).$$

Since G is independent of PA, we have a model $\mathcal{M}$ of PA s.t. $\mathcal{M} \vDash \mathrm{Bew}_{\mathrm{PA}}\left(\overline{\ulcorner G\urcorner}\right)$ and $\mathcal{M} \vDash \neg G$, violating the above requirement. Hence the relation we are looking for must be irreflexive. In the next section, we strengthen the notion of interpretability so as to satisfy the above requirement for truth definition in Kripke structures.

2.2 Strong interpretability

Roughly speaking, a strong interpretation of S in T is an interpretation I where T has a truth predicate for the internally constructed models of S. More precisely, there is a formula γ of T s.t. for all sentences φ of S, $T \vdash I(\varphi) \leftrightarrow \gamma\left(\overline{\ulcorner\varphi\urcorner}\right)$. We will refer to this condition as the minimal requirement for a truth predicate. The truth predicate that comes with a strong interpretation, however, has to satisfy more than just the minimal requirement. Apart from the above, we require T to prove various properties about γ as a truth predicate. For example, we want T to prove that $\varphi \wedge \psi$ is true in a (internally constructed) model of S if and only if both φ and ψ are true in that model, and that a formula of the form $\exists x\varphi$ is true in a (internally constructed) model of S if and only if the formula has a witnessing element in that model.

Preliminaries

Since we are working with a truth predicate, which can only be applied to (numerals of codes of) sentences, not to formulas and assignments, the last requirement poses a technical problem. In order to say that existential formulas have witnesses, T has to be able to express that an existential formula is in the extension of γ if and only if some instance of it is in the extension of γ. Since only sentences can be in the extension of γ, we need a *sentence* expressing the fact that $\exists x\varphi$ has a witness. Note that this is not a problem if S is a Henkin theory, i.e. if for any any existential sentence $\exists x\varphi$ in S, also some sentence of the form $\varphi(c)$ is in S, where c is a constant symbol. Then T can say that $\exists x\varphi$ has a witness by proving the statement $\gamma\left(\overline{\ulcorner\exists x\varphi\urcorner}\right)\leftrightarrow\gamma\left(\overline{\ulcorner\varphi\left(c\right)\urcorner}\right)$.

In order to deal with this problem, we consider the function κ which numerates a countable set $\{c_n\}_{n\in\mathbb{N}}$ of new constant symbols inside T:

$$\ulcorner c_n\urcorner = m \iff \mathrm{T}\vdash\kappa\left(\overline{n}\right)=\overline{m}$$

We assume that the coding of these constant symbols extends the standard coding, thereby ensuring that the element $\kappa[a]$ always codes a constant, and nothing else. We also assume κ to adequately represent the fact that the coding is injective, i.e.

$$T\vdash\forall xy\left(x\neq y\rightarrow\kappa\left(x\right)\neq\kappa\left(y\right)\right).$$

In particular, κ is an injective function from elements in the extension of δ to codes of constants, and so we can use it to assign names to the elements of an internally constructed model of S. If a is such an element (i.e. if $\mathcal{N}\vDash\delta\left[a\right]$ where $\mathcal{N}$ is a model of T), we let c_a, the a'th new constant, be its name according to T, and so $\kappa\left[a\right]$ is the code of the constant that names a.

The definition of a strong interpretation makes use of formulas representing certain primitive recursive functions and relations inside PA, which will be introduced in the following. Let $\sigma\left(x\right)$ be the formula numerating the axioms of S in T, i.e.

$$\varphi\in S\iff T\vdash\sigma\left(\overline{\ulcorner\varphi\urcorner}\right)$$

The formulas $\nu\left(x\right)$ and $\nu_F\left(x\right)$ numerate the sentences of S and formulas of S containing one free variable, respectively, inside T, e.g.

$$n\text{ is the code of an }S\text{-sentence}\iff\mathrm{T}\vdash\nu\left(\overline{n}\right)$$

Similarly, $\mathrm{Var}\left(x\right)$ numerates variables inside T.

Let $\mathrm{Sbs}\left(x,y,z\right)$ be the formula numerating the substitution function inside PA, i.e[5]

$$\left[t/x\right]\varphi\text{ is }\psi\iff\mathrm{T}\vdash\mathrm{Sbs}\left(\overline{\ulcorner t\urcorner},\overline{\ulcorner x\urcorner},\overline{\ulcorner\varphi\urcorner}\right)=\overline{\ulcorner\psi\urcorner}$$

The symbol "$*$" represents the concatenation function, thus $x*\overline{\ulcorner\rightarrow\urcorner}*y$ is the numeral of the code of an implication whose antecedent is the formula whose code is denoted by the numeral x, and the consequent the formula whose code is denoted by the numeral y.

[5] $\left[t/x\right]\varphi$ stands for the formula that results when the term t is substituted for all free occurrences of x in φ

Definition of strong interpretability

Let S and T be first-order theories containing PA, and the formulas $\kappa(x)$, $\sigma(x)$, $\nu(x)$, $\nu_F(x)$, $\mathrm{Sbs}(x)$, $\mathrm{Var}(x)$, as above.

Definition We call an interpretation I of S in T **strong**, and write $S \prec T$ if there is a formula γ of T containing one free variable s.t.

1. for all relation symbols[6] P of S,

$$T \vdash \forall xy \left(\delta\left(x\right) \wedge \delta\left(y\right) \rightarrow \left(I\left(P\left(x,y\right)\right) \leftrightarrow \gamma\left(\overline{\ulcorner P \urcorner} * \kappa\left(x\right) * \kappa\left(y\right)\right)\right)\right)$$

2. γ commutes with the propositional connectives:

   (a) $T \vdash \forall x \forall y \left(\nu\left(x\right) \wedge \nu\left(y\right) \rightarrow \left(\gamma\left(x * \overline{\ulcorner \rightarrow \urcorner} * y\right) \leftrightarrow \left(\gamma(x) \rightarrow \gamma\left(y\right)\right)\right)\right)$

   (b) $T \vdash \forall x \left(\nu\left(x\right) \rightarrow \left(\gamma\left(\overline{\ulcorner \neg \urcorner} * x\right) \leftrightarrow \neg\gamma(x)\right)\right)$

3. γ commutes with the quantifiers (in the relativised sense):

$$T \vdash \forall x \forall y \left(\nu_F\left(x\right) \wedge \mathrm{Var}\left(y\right) \rightarrow \left(\gamma\left(\overline{\ulcorner \exists \urcorner} * y * x\right) \leftrightarrow \exists z \left(\delta\left(z\right) \wedge \gamma\left(\mathrm{Sbs}\left(\kappa\left(z\right), y, x\right)\right)\right)\right)\right)$$

4. $T \vdash \forall x \left(\sigma(x) \rightarrow \gamma(x)\right)$

If a model $\mathcal{M}$ of S is given by a strong interpretation (as described in section 2.1) inside a model $\mathcal{N}$ of T, we say that $\mathcal{M}$ is a t-internal ('t' stands for truth) model of $\mathcal{N}$ or t-internal to $\mathcal{N}$, and write $\mathcal{M} \preccurlyeq \mathcal{N}$.

Note that in clause 3., the variable z plays a double role. If $\mathcal{N}$ is a model of T, and $\mathcal{N} \vDash \delta\left[a\right]$ for some $a \in \mathrm{dom}\left(\mathcal{N}\right)$, then a will be an element of the internal model $\mathcal{M}$ of S constructed inside $\mathcal{N}$. The first occurrence of z thus refers to an element in the domain of a model of S. The second instance of z (as an argument of κ), however, is concerned with $\kappa\left(a\right)$ as denoting the code of the constant c_a which functions as a name for a in that model. Clause 3. says that the sentence $\exists x\varphi$ is true in a model of S ($\gamma\left(\overline{\ulcorner \exists x\varphi \urcorner}\right)$) if and only if there is some element in the domain of S ($\delta\left(z\right)$), denoted by the constant c_z, s.t. $\varphi\left[c_z\right]$ is true in the model ($\gamma\left(\overline{\ulcorner \varphi\left[c_z\right] \urcorner}\right)$).

Similarly, in clause 1. the variable x plays different roles. On the left, we are interested in an element of an internally constructed model that is assigned to x by some assignment - this is how variables are usually interpreted. In the scope of γ, however, the element assigned to x interests us insofar as its image under κ codes a constant (naming the element that concerns us on the left side).

[6] We will treat the case where P is a binary relation symbol. The other cases are similar.

Internal quantifiers

In this section, we will explain why the internal quantifiers occurring in the definition of strong interpretability constitute a non-trivial requirement.

Note that clause 4. in the above definition requires T to prove the *formal statement* that all axioms of S are true. If T interprets S, then T containing a truth predicate which satisfies the minimal requirement already allows T to prove the *informal* version of this statement, i.e. that for all sentences φ of S, $T \vdash \gamma(\overline{\ulcorner\varphi\urcorner})$. To see that, note that if φ is a sentence of S, then by the definition of an interpretation, we have $T \vdash I(\varphi)$, and hence also $T \vdash \gamma(\overline{\ulcorner\varphi\urcorner})$ by the minimal requirement.

However, this alone does not guarantee the validity of clause 4, i.e. it might still be that $T \nvdash \forall x\,(\sigma(x) \rightarrow \gamma(x))$. To see this, consider an arbitrary model $\mathcal{N}$ of T and $a \in \mathrm{dom}(\mathcal{N})$ s.t. $\mathcal{N} \vDash \sigma[a]$.
If a is a standard element, then by properties of σ, a is the code of some sentence φ of S, whence $T \vdash I(\varphi)$, and thus also $T \vdash \gamma[a]$ (by the minimal requirement). Hence if $\mathcal{N}$ is the standard model, $\mathcal{N} \vDash \forall x\,(\sigma(x) \rightarrow \gamma(x))$.
But T also has nonstandard models, and so a might be a nonstandard element. In that case, a is not the code of any sentence of S, and the above argument does not go through, whence it seems possible that $\mathcal{M} \nvDash \gamma[a]$.
Whereas the external meta-language quantifier in the informal statement ranges over natural numbers (via their being codes of elements of S), the internal quantifier in the formal statement possibly also ranges over nonstandard elements a which are not used to code elements of S and thus fail to give us the desired conclusion that $T \vdash \gamma[a]$ immediately.
In fact, we can give a counterexample illustrating that clause 4. does not follow from the minimal requirement. Suppose that T interprets S via I, and let $\sigma(x)$ be s.t. for all φ,

$$\varphi \in S \iff T \vdash \sigma\left(\overline{\ulcorner\varphi\urcorner}\right).$$

Extend the language of T with a one-place predicate γ, and add to T, for any sentence φ in the language of S, the axiom $I(\varphi) \leftrightarrow \gamma\left(\overline{\ulcorner\varphi\urcorner}\right)$. Let T^{γ} be the resulting theory. Clearly, T^{γ} interprets S, and also contains a truth predicate that satisfies the minimal requirement. We will show that T^{γ} has a model that satisfies $\exists x\,(\sigma(x) \wedge \neg\gamma(x))$, i.e. a model where clause 4. is not satisfied. It suffices to show that any finite subset of $T^{\gamma} \cup \{\exists x\,(\sigma(x) \wedge \neg\gamma(x))\}$ has a model. So let T' be a finite subset of T^{γ}. Then there is some $\varphi \in S$ s.t. $T' \nvdash I(\varphi) \leftrightarrow \gamma\left(\overline{\ulcorner\varphi\urcorner}\right)$ (since S contains PA, it is not finitely axiomatisable).

If $T' \nvdash I(\varphi) \rightarrow \gamma\left(\overline{\ulcorner\varphi\urcorner}\right)$, then $T' \cup \left\{I(\varphi) \wedge \neg\gamma(\overline{\ulcorner\varphi\urcorner})\right\}$ is consistent. Since $T \vdash \sigma\left(\overline{\ulcorner\varphi\urcorner}\right) \Leftrightarrow T \vdash I(\varphi)$ (since we assumed $\varphi \in S$), also $T' \cup \left\{\sigma(\overline{\ulcorner\varphi\urcorner}) \wedge \neg\gamma(\overline{\ulcorner\varphi\urcorner})\right\}$ is consistent and thus it has a model. On the other hand, if $T' \nvdash \gamma\left(\overline{\ulcorner\varphi\urcorner}\right) \rightarrow I(\varphi)$, then by properties of I and γ we have that $T' \nvdash I(\neg\varphi) \rightarrow \gamma\left(\overline{\ulcorner\neg\varphi\urcorner}\right)$ (since I commutes with the propositional connectives, the minimal requirement imposes that also γ commutes with the propositional connectives in the externally quantified sense) and so $T' \cup \left\{\sigma(\overline{\ulcorner\neg\varphi\urcorner}) \wedge \neg\gamma(\overline{\ulcorner\neg\varphi\urcorner})\right\}$ has a model. Hence T^{γ} has a model where clause 4. is not satisfied, so the latter does not follow from the minimal requirement for a truth predicate.

A more natural example illustrating the independence of clause 4. from the minimal requirement for a truth predicate involves a non-standard consistency statement, e.g. one based on Feferman-provability (see Feferman, p. 68). According to Corollary 5.10, there is a formula $\pi(x)$ numerating the axioms of PA in PA s.t. PA $\vdash \mathrm{Con}_\pi$. Using the formalised completeness theorem (see section 2.4), we then obtain a formula γ, and an interpretation I of PA in itself, s.t. γ fulfils the minimal requirement for a truth predicate, together with clauses 1. - 3. of the definition of a strong interpretation. However, the fact that Con_π is not the standard Π_1^0 -consistency statement, clause 4. of the definition is not fulfilled. Hence it is the failure of this clause that prevents there being a strong interpretation of PA in itself, in accordance with the Second Incompleteness Theorem (if PA would strongly interpret itself, it would prove its own consistency statement).

Properties of strong interpretations

In this section, we prove some consequences of a theory T strongly interpreting a theory S via I and γ.

An induction on the complexity of a formula shows that for any formula φ of S containing at most n free variables, all variables $v_0, \dots v_n$, and for all $x_0, \dots, x_n$ in the extension of δ, T proves

$$I\left(\varphi\left(x_0, \dots x_n\right)\right) \leftrightarrow$$

$$\gamma\left(\mathrm{Sbs}\left(\kappa\left(x_0\right), \overline{\ulcorner v_0 \urcorner}, \dots, \mathrm{Sbs}\left(\kappa\left(x_n\right), \overline{\ulcorner v_n \urcorner}, \overline{\ulcorner \varphi\left(v_0, \dots v_n\right) \urcorner}\right) \dots\right)\right).$$

In particular if φ is a sentence,

$$T \vdash I(\varphi) \leftrightarrow \gamma\left(\overline{\ulcorner \varphi \urcorner}\right).$$

i.e. γ satisfies the minimal requirement for a truth predicate.

An important property of a strong interpretation of S in T is that

$$T \vdash \forall x\left(\mathrm{Bew}_\sigma(x) \rightarrow \gamma(x)\right)$$

where σ is the formula numerating the axioms of S in T. This means that T proves the formal statement that all theorems of S are true.

The proof relies on Theorem 4.6.(v) in Feferman's article. This theorem formalises the method of using induction on the length of the derivation in order to prove properties of derivable formulas. For any formula χ of T, we have:

$$T \vdash \forall x\left(\left(\sigma(x) \rightarrow \chi(x)\right) \wedge \left(\lambda(x) \rightarrow \chi(x)\right)\right) \wedge$$

$$\forall xy\left(\chi\left(x * \overline{\ulcorner \rightarrow \urcorner} * y\right) \wedge \chi(x) \rightarrow \chi(y)\right)$$

$$\rightarrow \forall z\left(\mathrm{Bew}_\sigma(z) \rightarrow \chi(z)\right)$$

where λ is a formula that numerates a set of logical axioms in T.

Taking γ for χ in the above formula, we thus have to prove the following in order to get the desired result:

$$T \vdash \forall x\,(\sigma(x) \to \gamma(x)) \tag{1}$$

$$T \vdash \forall x\,(\lambda(x) \to \gamma(x)) \tag{2}$$

$$T \vdash \forall xy\,\left(\left(\gamma\left(x * \overline{\ulcorner\to\urcorner} * y\right) \wedge \gamma(x)\right) \to \gamma(y)\right) \tag{3}$$

(1) Immediate by the definition of strong interpretability.

(2) We assume that $\lambda\,(x)$ is an intensionally correct formalisation of the set of axioms of predicate logic in Hilbert style, defined as a finite disjunction of sentences like e.g. $\exists y \exists z \left(x = \left(y * \overline{\ulcorner\to\urcorner} * \left(z * \overline{\ulcorner\to\urcorner} * y\right)\right)\right)$, corresponding to the axiom schema $(\varphi \to (\psi \to \varphi))$. Let $\mathcal{M}$ be a model of T, and $m \in \mathrm{dom}\,(\mathcal{M})$ s.t. $\mathcal{M} \vDash \lambda\,[m]$. We will consider an exemplary case where m has the form $a * \overline{\ulcorner\to\urcorner} * \left(b * \overline{\ulcorner\to\urcorner} * a\right)$, with $a, b \in \mathrm{dom}\,(\mathcal{M})$ possibly nonstandard. We want to show that $\mathcal{M} \vDash \gamma\,[m]$, i.e. that $\mathcal{M} \vDash \gamma\left[a * \overline{\ulcorner\to\urcorner} * \left(b * \overline{\ulcorner\to\urcorner} * a\right)\right]$. By the requirement that γ commutes with the propositional connectives in the internally quantifed sense , we have that

$$T \vdash \forall x \forall y \left(\gamma\left(x * \overline{\ulcorner\to\urcorner} * \left(y * \overline{\ulcorner\to\urcorner} * x\right)\right) \leftrightarrow (\gamma\,(x) \to (\gamma\,(y) \to \gamma\,(x)))\right)$$

whence it suffices to prove that $\mathcal{M} \vDash \gamma\,[a] \to (\gamma\,[b] \to \gamma\,[a])$. But the latter is a propositional tautology and hence we are done.

(3) Immediate by the requirement. that γ commutes with the propositional connectives in the internally quantified sense.

From the above, it also follows that $T \vdash \neg\mathrm{Bew}_\sigma\left(\overline{\ulcorner\bot\urcorner}\right)$. We have $T \vdash \forall x\,(\neg\gamma\,(x) \to \neg\mathrm{Bew}_\sigma\,(x))$ by propositional logic, and hence also $T \vdash \neg\gamma\left(\overline{\ulcorner\bot\urcorner}\right) \to \neg\mathrm{Bew}_\sigma\left(\overline{\ulcorner\bot\urcorner}\right)$. On the other hand, we also have $T \vdash \neg\gamma\left(\overline{\ulcorner\bot\urcorner}\right) \leftrightarrow \neg I\,(\bot)$, i.e., since I commutes with the propositional connectives, $T \vdash \neg\gamma\left(\overline{\ulcorner\bot\urcorner}\right) \leftrightarrow \top$, whence $T \vdash \neg\gamma\left(\overline{\ulcorner\bot\urcorner}\right)$. Thus if T strongly interprets S, then T proves the formal statement of S's consistency. This also implies that the relation of strong interpretability between theories whose models are models PA is irreflexive: by Gödel's Second Incompleteness Theorem, no such theory can prove its own consistency statement. In the next section, we will see that also the opposite of the above statement is the case: if T proves the formal statement of S's consistency, then T strongly interprets S.

Since the axioms of PA are a subset of the axioms of S, the above proof also goes through if instead of σ we work with a formula numerating the axioms of PA in T. It follows that $T \vdash \forall x\,(\mathrm{Bew}_{\mathrm{PA}}(x) \to \gamma(x))$, and $T \vdash \mathrm{Con}_{\mathrm{PA}}$, i.e. T proves that all theorems of PA are true in models of S which are internally constructed inside models of T.

2.3 Strong interpretability as an accessibility relation

This section establishes the property of the relation of strong interpretability allowing us to think of models of PA as nodes in a Kripke structure (see also 2.1). Let $\mathcal{N}$ and $\mathcal{M}$ be models of PA. Then

$$\mathcal{N} \vDash \mathrm{Bew}_{\mathrm{PA}}(\overline{\ulcorner\varphi\urcorner}) \Rightarrow \text{ for all } \mathcal{M} \text{ s.t. } \mathcal{M} \preccurlyeq \mathcal{N},\, \mathcal{M} \vDash \varphi.$$

$$\mathcal{N} \vDash \neg\mathrm{Bew}_{\mathrm{PA}}\left(\overline{\ulcorner\varphi\urcorner}\right) \Rightarrow \text{ there is } \mathcal{M} \text{ s.t. } \mathcal{M} \preccurlyeq \mathcal{N} \text{ and } \mathcal{M} \vDash \neg\varphi$$

Besides making it possible to construct the Big Model, the validity of this equation has another virtue: it makes the provability predicate $\mathrm{Bew}_{\mathrm{PA}}$ behave in a very reasonable way.

As mentioned in section 2.1, a consequence of the Diagonal Lemma is that inside any model $\mathcal{M}$ of PA, we have, for some sentences φ, both $\mathcal{M} \vDash \varphi$ and $\mathcal{M} \vDash \neg\mathrm{Bew}_{\mathrm{PA}}\left(\overline{\ulcorner\varphi\urcorner}\right)$, or $\mathcal{M} \vDash \varphi$ and $\mathcal{M} \vDash \mathrm{Bew}_{\mathrm{PA}}(\overline{\ulcorner\neg\varphi\urcorner})$. But if a model "claims" that a sentence is true, but at the same time not provable, or that a sentence is true but its negation is provable, it might seem that there is something wrong with the provability predicate.

The validity of the above equations shows that there is at least one context that makes the provability predicate behave as neatly as we might expect. The best that we can get, in this sense, is a provability predicate in one model whose behaviour is meaningful with respect to sentences in *other* models, namely ones which are t-internal to it. If a model $\mathcal{N}$ claims that φ is provable, then φ is true in all its t-internal models, and if $\mathcal{N}$ claims that φ is not provable, then there is at least one model t-internal to $\mathcal{N}$ where φ is false. Thus, whereas the self-reflexive capacities of PA have a limit, in that the statement $\mathrm{Bew}_{\mathrm{PA}}(\overline{\ulcorner\varphi\urcorner}) \leftrightarrow \varphi$ is not provable for all sentences φ[7], any model of PA has "perfect knowledge" about all models that are t-internal to it, at least when it comes to provability in PA. In the following, I will outline the proofs of the above statements.

In order to show that

$$\mathcal{N} \vDash \mathrm{Bew}_{\mathrm{PA}}(\overline{\ulcorner\varphi\urcorner}) \Rightarrow \text{ for all } \mathcal{M} \text{ s.t. } \mathcal{M} \preccurlyeq \mathcal{N},\, \mathcal{M} \vDash \varphi$$

assume $\mathcal{N} \vDash \mathrm{Bew}_{\mathrm{PA}}(\overline{\ulcorner\varphi\urcorner})$, $\mathcal{M} \preccurlyeq \mathcal{N}$, $\mathcal{N}$ is a model of T, and $\mathcal{M}$ a model of S. By the result in 2.2, we have that $T \vdash \forall x\, (\mathrm{Bew}_{\mathrm{PA}}(x) \rightarrow \gamma(x))$. Thus $\mathcal{N} \vDash \mathrm{Bew}_{\mathrm{PA}}(\overline{\ulcorner\varphi\urcorner})$ implies $\mathcal{N} \vDash \gamma\left(\overline{\ulcorner\varphi\urcorner}\right)$. Since also $T \vdash I\,(\varphi) \leftrightarrow \gamma\left(\overline{\ulcorner\varphi\urcorner}\right)$ (see section 2.2), $\mathcal{N} \vDash I(\varphi)$. Therefore $\mathcal{M} \vDash \varphi$ (since $\mathcal{M}$ is t-internal to $\mathcal{N}$, this follows by construction of $\mathcal{M}$ inside $\mathcal{N}$).

For the other direction, we show

$$\mathcal{N} \vDash \neg\mathrm{Bew}_{\mathrm{PA}}\left(\overline{\ulcorner\varphi\urcorner}\right) \Rightarrow \text{ there is } \mathcal{M} \text{ s.t. } \mathcal{M} \preccurlyeq \mathcal{N} \text{ and } \mathcal{M} \vDash \neg\varphi$$

Since $\mathcal{N}$ is a model of $\mathrm{PA} + \neg\mathrm{Bew}_{\mathrm{PA}}\left(\overline{\ulcorner\varphi\urcorner}\right)$, it suffices to show that

$$\mathrm{PA} + \neg\varphi \prec \mathrm{PA} + \neg\mathrm{Bew}_{\mathrm{PA}}(\overline{\ulcorner\varphi\urcorner})$$

[7] This biconditional is provable only for provable Σ_1-sentences.

Note that $\neg\mathrm{Bew}_{\mathrm{PA}}(\ulcorner\varphi\urcorner)$ is a consistency statement for $\mathrm{PA} + \neg\varphi$. First, it states that not everything is provable in PA, implying the consistency of PA. In particular, it states that φ is not provable in PA, and this implies the consistency of $\mathrm{PA} + \neg\varphi$. The desired result is thus a direct consequence of the formalised Completeness Theorem for first-order theories, according to which T strongly interprets S, given that T contains an ordinary Π^0_1 consistency statement for S (as was shown in section 2.2, also the converse holds). The next section contains a proof sketch of this theorem.

2.4 Formalised Gödel's Completeness Theorem

Let S be a first-order theory and $\sigma(x)$ a formula that numerates the axioms of S in PA. Then

$$S \prec \mathrm{PA} + \mathrm{Con}_\sigma,$$

where Con_σ is an ordinary Π^0_1-consistency statement for S (e.g. $\neg\mathrm{Bew}_\sigma\left(\ulcorner\bot\urcorner\right)$) depending on the representation of S's axioms by σ inside PA.

This theorem is a formalised version of Gödel's Completeness Theorem which states that every consistent theory has a model. In the formalised version, *our* perspective is replaced by that of PA, and the intuitive meaning of the theorem is: PA "knows" that if S is consistent (Con_σ), then S has a model (a model of S can be internally constructed inside a model of $\mathrm{PA} + \mathrm{Con}_\sigma$). Traditionally (see Feferman, Lindström), the proof of the theorem is used to obtain an interpretation I of S in $\mathrm{PA} + \mathrm{Con}_\sigma$. However, the additional requirement that the consistency statement is the ordinary Π^0_1-consistency statement suffices for PA to *strongly* interpret S.

Our proof of the formalised Gödel's Completeness Theorem resembles closely Henkin's proof of Gödel's Completeness Theorem. After giving an outline of the latter, we will indicate how it can be carried out inside PA.

Henkin's proof of Gödel's Completeness Theorem

Let S be a consistent theory in the language L_S. In order to find a model for S, we expand the language of S to L and construct an L-theory Γ s.t. $S \subseteq \Gamma$, Γ is maximal (i.e. for every L-sentence φ, either $\varphi \in \Gamma$ or $\neg\varphi \in \Gamma$), and if $\exists x\psi \in \Gamma$ for some L-formula ψ, then there is some constant in L s.t. $\psi[c/x] \in \Gamma$. We say that a theory with these properties is a Henkin theory. The syntactical information contained in Γ is then rich enough to allow us to construct a model for S on the basis of it.

In constructing the language L, we have to make sure that there is a witnessing constant for every existential sentence of L. Bearing in mind that we have to formalise this procedure in PA, it is convenient if the constants are typographically associated with the formulas they will witness in Γ. For that we consider a fixed constant symbol c. Given a formula $\exists x\varphi$, the witnessing constant for that formula will be the symbol $c[\exists x\varphi]$. Note that given any reasonable coding, this choice of constants will imply that for any sentence φ, $\ulcorner\varphi\urcorner < \ulcorner c[\varphi]\urcorner$.

The constants and sentences of L will be constructed interdependently:

1. if φ is a sentence of L, then $c[\varphi]$ is a constant of L
2. the formulas of L are constructed by the standard rules

This definition ensures that whenever φ is a sentence of L, the constant $c[\varphi]$ is a term of L, and also that $\ulcorner\varphi\urcorner < \ulcorner c[\varphi]\urcorner$.

Next, we will recursively construct the Henkin theory Γ by making sure simultaneously that it is maximal consistent and contains witnessing constants for all existential formulas. For that, consider an enumeration of L-sentences according to their codes: $\varphi_{j_0}, \varphi_{j_1}, \varphi_{j_2}, \ldots$ where $\ulcorner\varphi_{j_k}\urcorner = j_k$ for all k, and j_0 is the smallest code of an L-sentence, j_1 the next smallest code of an L-sentence, etc. We can now recursively define theories Γ_i for $i \in \omega$. Let $\Gamma_0 := \emptyset$. Assuming we have already defined Γ_i for $j < i$, we define Γ_i by taking care of the L-sentence with code i. In case there is no L- sentence φ s.t. $i = \ulcorner\varphi\urcorner$, we let $\Gamma_{i+1} := \Gamma_i$. In case $i = \ulcorner\varphi\urcorner$ for some L-sentence φ, we make sure that either φ or $\neg\varphi$ is in Γ_i, and in case φ is an existential sentence, we make sure that Γ_i contains a witnessing constant for φ. More formally, if $i = j_k$ for some k, i.e. if $i = \ulcorner\varphi_i\urcorner$, we define

$$\Gamma_i := \begin{cases} \Gamma_{i-1} \cup \{\neg\varphi_i\} & \text{if } S \nvdash \bigwedge_{\theta_j \in \Gamma_{i-1}} \theta_j \to \varphi_i \\ \Gamma_{i-1} \cup \{\varphi_i\} & \text{if } S \vdash \bigwedge_{\theta_j \in \Gamma_{i-1}} \theta_j \to \varphi_i, \\ & \text{and } \varphi_i \text{ does not have the form } \exists x\psi \\ \Gamma_{i-1} \cup \{\varphi_i \wedge \psi[c[\varphi_i]/x]\} & \text{if } S \vdash \bigwedge_{\theta_j \in \Gamma_{i-1}} \theta_j \to \varphi_i, \\ & \text{and } \varphi_i \text{ has the form } \exists x\psi \end{cases}$$

and let

$$\Gamma := \bigcup_{i \in \mathbb{N}} \Gamma_i.$$

Although in usual circumstances, the above definition could be made to look more intuitive, e.g. by letting $\Gamma_0 := S$ (from which it would immediately follow that $S \subseteq \Gamma$), we have to keep in mind that the goal of formalising this procedure inside PA. In particular, it is good to keep each of the steps finite (where possible), so instead of adding all sentences of S immediately, we add each on the appropriate stage (indexed by its code).

Note that at stage, i $c[\varphi_i]$ is a new constant, i.e. it occurs neither in φ_i nor in Γ_{i-1}. The first is due to the fact that $i = \ulcorner\varphi_i\urcorner < \ulcorner c[\varphi_i]\urcorner$ by construction of the individual constants. If $c[\varphi_i]$ would occur in φ_i, we would have $\ulcorner\varphi_i\urcorner > \ulcorner c[\varphi_i]\urcorner$ for any reasonable coding. And $c[\varphi_i]$ does not occur in Γ_{i-1} because all constants occurring in Γ_i have smaller codes than $c[\varphi_i]$. By choice of the enumeration of L-sentences, for all $j < i$, $\ulcorner\varphi_j\urcorner < \ulcorner\varphi_i\urcorner$, whence also $\ulcorner c[\varphi_j]\urcorner < \ulcorner c[\varphi_i]\urcorner$ for all $j < i$.

We will now show that Γ is a Henkin theory.

First, we have that $S \subseteq \Gamma$. If $\varphi_i \in S$ then $S \vdash \varphi_i$, hence also for any ψ, $S \vdash \psi \to \varphi_i$ and so φ_i is added to Γ at stage i.

We show by induction on i that Γ_i is consistent for all i. So assume that Γ_i is consistent and suppose for contradiction that $\Gamma_{i+1} \vdash \perp$. Then i must be the code of some L-sentence φ_i, otherwise we would have $\Gamma_i = \Gamma_{i+1}$.

- In case $\Gamma_{i+1} = \Gamma_i \cup \{\neg\varphi_{i+1}\}$, $S \nvdash \bigwedge_{\theta_j \in \Gamma_i} \theta_j \to \varphi_{i+1}$ by definition of Γ_{i+1}. But if $\Gamma_{i+1} = \Gamma_i \cup \{\neg\varphi_{i+1}\} \vdash \bot$, also $\Gamma_i \vdash \neg\neg\varphi_{i+1}$, i.e. $\bigwedge_{\theta_j \in \Gamma_i} \theta_j \vdash \varphi_{i+1}$, whence also $S \vdash \bigwedge_{\theta_j \in \Gamma_i} \theta_j \to \varphi_{i+1}$ which is a contradiction.

- If $\Gamma_{i+1} = \Gamma_i \cup \{\varphi_{i+1}\}$, then $S \vdash \bigwedge_{\theta_j \in \Gamma_i} \theta_j \to \varphi_{i+1}$. If also $\Gamma_{i+1} = \Gamma_i \cup \{\varphi_{i+1}\} \vdash \bot$, then $\bigwedge_{\theta_j \in \Gamma_i} \theta_j \vdash \neg\varphi_{i+1}$, whence $\vdash \bigwedge_{\theta_j \in \Gamma_i} \theta_j \to \neg\varphi_{i+1}$, and also $S \vdash \bigwedge_{\theta_j \in \Gamma_i} \theta_j \to \neg\varphi_{i+1}$. So $S \cup \Gamma_i$ is inconsistent. So there is some sentence φ where $\overline{\ulcorner\varphi\urcorner} = j < i+1$ s.t. $S \vdash \varphi$ and $\Gamma_i \vdash \neg\varphi$. By construction of Γ_i, we have that $\varphi \in \Gamma_i$, contradicting the assumption that Γ_i is consistent.

- Finally, if $\Gamma_{i+1} = \Gamma_i \cup \{\varphi_{i+1} \wedge \psi\,[c\,[\varphi_{i+1}]\,/x]\}$, then $S \vdash \bigwedge_{\theta_j \in \Gamma_i} \theta_j \to \varphi_{i+1}$ and φ_{i+1} has the form $\exists x\psi$. If $\Gamma_i \cup \{\varphi_{i+1} \wedge \psi\,[c\,[\varphi_{i+1}]\,/x]\} \vdash \bot$, and $\Gamma_i \nvdash \neg\varphi_{i+1}$, then $\Gamma_i \vdash \neg\,\{\psi\,[c\,[\varphi_{i+1}]\,/x]\}$. Since $c\,[\varphi_i]$ does not occur in Γ_i or in φ_i, this implies that $\Gamma_i \vdash \forall x \neg\psi$, i.e. $\vdash \bigwedge_{\theta_j \in \Gamma_i} \theta_j \to \neg\varphi_i$ and so also $S \vdash \bigwedge_{\theta_j \in \Gamma_i} \theta_j \to \neg\varphi_i$. By a similar argument as above, this can be seen to contradict the consistency of Γ_i.

Then also Γ is consistent. If $\Gamma \vdash \bot$, then there is a finite subset Γ' of Γ s.t. $\Gamma' \vdash \bot$. Let $k = \max\{\ulcorner\varphi\urcorner \mid \varphi \in \Gamma'\}$. Then $\Gamma' \subseteq \Gamma_{k+1}$, contradicting the observation above.

Γ is maximal: by construction, we have either $\varphi \in \Gamma$ or $\neg\varphi \in \Gamma$ for every L-sentence φ. It follows that for every L- sentence φ, $\varphi \in \Gamma$ if and only if $\Gamma \vdash \varphi$. If $\varphi \in \Gamma$, then obviously $\Gamma \vdash \varphi$. For the other direction, suppose that $\Gamma \vdash \varphi$. We have that either $\varphi \in \Gamma$ or $\neg\varphi \in \Gamma$. But if $\neg\varphi \in \Gamma$, then $\Gamma \vdash \neg\varphi$, contradicting the consistency of Γ.

By construction, we also have that if $\exists x\varphi \in \Gamma$, there is a constant c in L s.t. $\varphi\,[c\,[\exists x\varphi]\,/x] \in \Gamma$.

To obtain the model $\mathcal{M}$ of S we let $\operatorname{dom}(\mathcal{M}) = \{c\,[\varphi] \mid \varphi \text{ is an } L\text{-sentence}\}$[8], and use membership in Γ to define the interpretations of sentences. If φ and ψ are L-sentences, we let

$$c\,[\varphi]^{\mathcal{M}} = c\,[\varphi]$$

and e.g. for a two-place predicate P,

$$\left(c\,[\varphi]^{\mathcal{M}}, c\,[\psi]^{\mathcal{M}}\right) \in P^{\mathcal{M}} \iff Pc\,[\varphi]\,c\,[\psi] \in \Gamma$$

Using the fact that Γ is a Henkin theory, we can then show by induction that for all L-sentences φ,

$$\mathcal{M} \vDash \varphi \iff \varphi \in \Gamma.$$

Since $S \subseteq \Gamma$, it follows that $\mathcal{M}$ is a model of S.

[8] We will ignore the issue that of having $c\,[\varphi] = c\,[\psi] \in \Gamma$ for different L-sentences φ and ψ, in which case $c\,[\varphi]$ and $c\,[\psi]$ should refer to the same object in $\operatorname{dom}(\mathcal{M})$. This can be solved by taking the elements of the domain to be equivalence classes of constants instead of constants themselves. In the formal version, the situation can be solved by taking the smallest representatives of each equivalence class, i.e. the smallest codes.

The formalised Completeness Theorem

Let σ numerate the axioms of S in PA, i.e. for all sentences φ,

$$\varphi \in S \iff \mathrm{PA} \vdash \sigma\left(\overline{\ulcorner\varphi\urcorner}\right).$$

In the formalised version of the Completeness theorem, we want to construct a model of S inside a model of $\mathrm{PA} + \mathrm{Con}_\sigma$. The elements of this model will be elements (of a model of $\mathrm{PA}+\mathrm{Con}_\sigma$) which function as *codes* of the individual constants that were added to L_S to obtain L. The set $\{\ulcorner\varphi\urcorner \mid \varphi \text{ is an } L \text{ -sentence}\}$ is recursively enumerable, hence we have a formula $\nu(x)$ which numerates the elements of this set in PA, i.e.

$$\varphi_j \text{ is an } L\text{-sentence} \iff \mathrm{PA} \vdash \nu\left(\overline{j}\right)$$

where j refers to the enumeration of L-sentences according to their codes (i.e. $\ulcorner\varphi_j\urcorner = j$). Also, let ν_F be the formula numerating the formulas of L containing one free variable.

By that, we also have a formula κ that numerates the new individual constants in PA, i.e.

$$\ulcorner c\left[\varphi_j\right]\urcorner = a \iff \mathrm{PA} \vdash \kappa\left(\overline{j}\right) = \overline{a}$$

We will now define the formula $\gamma'(z, y)$ which will intuitively correspond, inside PA, to the recursive definition of the theories Γ_i. The first variable, z, will represent a finite sequence whose elements are the (codes of) sentences that have already been added. In order to make the construction of γ' simple, we assume that exactly one sentence is added at each stage. So, if at stage i, i is not the code of any sentence, we just add $\top$. If x codes an existential sentence, we add the conjunction of this existential sentence and its instance containing the witnessing constant. In that way, z represents the sentences that have been added so far, and at the same time also the stage of construction. The second argument y represents the code of the last sentence that has been added to this finite sequence. For reasons of readability, we have written down the formula in a semi-formal language.

$$\gamma'(z, y) := \mathrm{FinSeq}(z) \wedge (z)_{\mathrm{length}(z)-1} = y \wedge$$
$$\forall x < \mathrm{length}(z)$$

1. $(z)_x = \overline{\ulcorner\top\urcorner}$ if $\neg\nu(x)$
2. if $\nu(x)$:
    (a) $(z)_x = \overline{\ulcorner\neg\urcorner} * x$ if $\neg\mathrm{Bew}_\sigma\left(\left(\left((z)_0 * \overline{\ulcorner\wedge\urcorner} * \cdots * \overline{\ulcorner\wedge\urcorner}\,(z)_{x-1}\right) * \overline{\ulcorner\rightarrow\urcorner} * x\right)\right.$
    (b) if $\mathrm{Bew}_\sigma\left(\left(\left((z)_0 * \overline{\ulcorner\wedge\urcorner} * \cdots * \overline{\ulcorner\wedge\urcorner}\,(z)_{x-1}\right) * \overline{\ulcorner\rightarrow\urcorner} * x\right)\right.$:
        i. $(z)_x = x$ if $\neg\exists vu\left(\mathrm{Var}(v) \wedge \nu_F(u) \wedge x = \overline{\exists} * v * u\right)$
        ii. $(z)_x = x {*} \overline{\ulcorner\wedge\urcorner} {*} \mathrm{Sbs}(\kappa(x), v, u)$ if $\exists vu\left(\mathrm{Var}(v) \wedge \nu_F(u) \wedge x = \overline{\exists} * v * u\right)$

Even more informally (here we speak of formulas instead of their codes),

1. if x does not code a formula, the x'th element of the sequence is $\top$.
2. if x codes a formula φ, then the x'th element of the sequence is:
    (a) $\neg\varphi$ if $\neg\varphi$ is consistent with the previous elements in the sequence
    (b) if the previous elements in the sequence prove φ, (according to the axioms of S) the x'th element of the sequence is
        i. φ if φ is not an existential sentence
        ii. $\varphi \wedge [x/c[\varphi]]\,\psi$ if φ is an existential sentence $\exists x\psi$

The construction of γ' illustrates why it is convenient for the constants of L to be typographically associated to the existential sentences they are supposed to witness. Using a more traditional form of the completeness proof, we would have to explicitly require the constant c witnessing $\varphi_n = \exists z\psi$ to be new with respect to Γ_n and φ_n. Following our version of the completeness proof, however, we obtain the new constant straightforwardly by using $\kappa(\overline{n})$, without having to think of formulas of PA which would express the fact that a certain constant is new. We can now define the formula $\gamma(x)$ by letting

$$\gamma(x) := \exists z\gamma'(z, x) \vee \operatorname{ext}(x)$$

where $\operatorname{ext}(x)$ is an abbreviation for the formula

$$\exists vu\left(\operatorname{Var}(v) \wedge \nu_F(u) \wedge x = \overline{\exists} * v * u\right) \wedge \exists w\gamma'\left(w, x * \overline{\ulcorner\wedge\urcorner} * \operatorname{Sbs}(\kappa(x), v, u)\right).$$

The existential case needs to be treated separately because an existential formula is never added to Γ alone, but only as a conjunct together with its witnessing instance (the other conjunct will appear at some later stage in the enumeration of L-sentences where it will be added as a single sentence).

Intuitively, $\gamma(x)$ means that x codes a member of Γ. Note that since Γ is not recursively enumerable, γ can not numerate Γ inside PA. We could say that PA's knowledge of Γ via γ is similar to our knowledge of Γ. Although we know facts about Γ, such that it contains S, and is maximal consistent, we do not know which particular elements it contains; we even cannot list these elements. The same applies to PA.

Definition of I

In order to define the interpretation I, we first set $\delta(x) := \exists y(x = \kappa(y))$. Thus some element a of a model of $\mathrm{PA} + \mathrm{Con}_\sigma$ is in δ - in the domain of the internally constructed model of S - if and only if a is the code of some constant c_y. Similarly like we had to prove that the syntactical information in Γ was rich enough to allow us to construct a model, we have to see that PA proves enough facts about γ for it to function as a truth predicate.

In particular, we have that $\mathrm{PA} + \mathrm{Con}_\sigma$ proves the following statements (the free variables below should be viewed as universally quantified):

$$\sigma(x) \to \gamma(x) \tag{4}$$

corresponding to the fact that $S \subseteq \Gamma$. In order to establish this, Con_σ has to be an ordinary Π^0_1-consistency statement. $\mathrm{PA} + \mathrm{Con}_\sigma$ also proves

$$\mathrm{Con}_\gamma \tag{5}$$

corresponding to the result in the informal proof that Γ is consistent if S is. Corresponding to the result that Γ is maximal, we have that

$$\nu(x) \to \left(\gamma\left(\overline{\ulcorner\neg\urcorner} * x\right) \leftrightarrow \neg\gamma(x)\right) \tag{6}$$

and

$$\nu(x) \wedge \nu(y) \to \left(\gamma\left(x * \overline{\ulcorner\to\urcorner} * y\right) \leftrightarrow (\gamma(x) \to \gamma(y))\right). \tag{7}$$

Corresponding to the property that Γ contains witnesses for all existential statements, i.e.. that

$$\exists x\psi \in \Gamma \iff \psi(c[\exists x\psi]) \in \Gamma$$

we have

$$\nu_F(x) \wedge \mathrm{Var}(y) \to \left(\gamma\left(\overline{\ulcorner\exists\urcorner} * y * x\right) \leftrightarrow (\exists z(\delta(z) \to \gamma(\mathrm{Sbs}(z, y, x)))))\right) \tag{8}$$

In the informal statement, the symbol $c[\exists x\psi]$ functions both as a syntactical object - an individual constant - and an element in the domain of a model that is defined through Γ. In a similar way, the variable z in the formal version has two functions. In the first occurrence, we are concerned with z as denoting an element of the domain of the internal model of S (constructed inside a model of $\mathrm{PA} + \mathrm{Con}_\sigma$). In the second occurrence, we are concerned with z as coding some individual constant c. Since z is in the domain if and only if it codes an individual constant by definition, the two perspectives are entangled, exactly as in the informal version, where whether c is an element of the domain or an individual constant is just a matter of perspective.

As a formula numerating the elements of a maximal consistent set, γ is a good candidate for the truth predicate that we require from a strong interpretation. In fact, note that the properties of γ listed above correspond to clauses 2.-4. in the definition of a strong interpretation. In order to also get clause 1. we just have to define the interpretation I in a suitable way. Whereas in the informal version, we use membership in Γ to construct the model $\mathcal{M}$ of S, we will now use γ to define the interpretation I of S inside $\mathrm{PA} + \mathrm{Con}_\sigma$.

To define I using γ, we let, for a two-place predicate P, and x and y in δ,

$$I(P(x, y)) = \gamma\left(\overline{\ulcorner P\urcorner} * x * y\right).$$

Thus if $x = \overline{\ulcorner c_a\urcorner}$ and $y = \overline{\ulcorner c_b\urcorner}$ for some a, b (this is required for x and y to be in the extension of δ), then $I(P(x, y)) = \gamma\left(\overline{\ulcorner P(c_m, c_n)\urcorner}\right)$. Then $\gamma\left(\overline{\ulcorner P\urcorner} * x * y\right)$

is the formula $\psi_P(x, y)$ required in the definition of an interpretation. We also require I to commute with the propositional connectives, and with the quantifiers in a relativised sense, i.e. for all sentences φ and ψ,

1. $I(\neg\varphi) = \neg I(\varphi)$
2. $I(\varphi \rightarrow \psi) = I(\varphi) \rightarrow I(\psi)$
3. $I(\exists x\varphi) = \exists x\,(\delta(x) \wedge I(\varphi))$

By these requirements, I is determined for all formulas of S. By the definition of I, and since $\mathrm{PA}+\mathrm{Con}_\sigma$ proves the properties of γ as numerating the elements of a maximal consistent set, $\mathrm{PA}+\mathrm{Con}_\sigma$ proves for any formula φ of S containing at most the free variables $v_0, \dots v_n$, and for all $x_0, \dots, x_n$ in the extension of δ,

$$I(\varphi(x_0, \dots x_n)) \leftrightarrow \gamma\left(\mathrm{Sbs}\left(x_0, \overline{\ulcorner v_0 \urcorner}, \dots, \mathrm{Sbs}\left(x_n, \overline{\ulcorner v_n \urcorner}, \overline{\ulcorner \varphi(v_0, \dots v_n) \urcorner}\right) \dots\right)\right)$$

(where the atomic case follows by the definition of I). This corresponds to the result in the informal proof that for all L-sentences φ,

$$\mathcal{M} \vDash \varphi \iff \varphi \in \Gamma$$

In particular, if φ is a sentence of S, then

$$\mathrm{PA} + \mathrm{Con}_\sigma \vdash I(\varphi) \leftrightarrow \gamma(\overline{\ulcorner \varphi \urcorner})$$

i.e. γ satisfies the minimal requirement for a truth predicate. Thus, PA "knows" that: a sentence φ is true in a model of S ($I(\varphi)$ - because I is used to construct a model of S inside a model of $\mathrm{PA} + \mathrm{Con}_\sigma$) if and only if φ is included in the Henkin theory Γ containing S ($\gamma(\overline{\ulcorner \varphi \urcorner})$).

Since $\mathrm{PA} + \mathrm{Con}_\sigma \vdash \forall x(\sigma(x) \rightarrow \gamma(x))$, and σ numerates the axioms of S in $\mathrm{PA} + \mathrm{Con}_\sigma$, we have, for all axioms ψ of S:

$$\mathrm{PA} + \mathrm{Con}_\sigma \vdash \gamma(\overline{\ulcorner \psi \urcorner})$$

and hence also

$$\mathrm{PA} + \mathrm{Con}_\sigma \vdash I(\psi)$$

It follows that I satisfies all the requirements of an interpretation, and together with γ we have a strong interpretation of S in $\mathrm{PA} + \mathrm{Con}_\sigma$.

3 Solovay's proof

The proof of GL's arithmetical completeness proceeds by contraposition, and makes use of the fact that GL is complete with respect to finite transitive and irreflexive Kripke frames. Given a sentence φ s.t. GL $\nvdash \varphi$, we want to find a realisation $*$ s.t. PA $\nvdash \varphi^*$. Due to the semantical completeness of GL, there is a model $\mathfrak{M} = \langle W_{\mathfrak{M}}, R, V_{\mathfrak{M}} \rangle$, where $W_{\mathfrak{M}} = \{1, \ldots, n\}$; R is transitive and irreflexive, and $\mathfrak{M}, 1 \nvDash \varphi$.

The core of Solovay's proof is the construction of sentences $S_0, S_1 \ldots, S_n$ of $\mathcal{L}_{\mathrm{PA}}$ (I will refer to these as Solovay sentences from now on) that make it possible to simulate $\mathfrak{M}$ inside PA. By translating a sentence letter into PA as the disjunction of all S_i s.t. the sentence letter is true at the world i in $\mathfrak{M}$, i.e. putting $p^* = \bigvee_{i:iVp} S_i$ for a sentence letter p, the simulation consists in the following result:

Lemma 1 : For each subsentence χ of φ:

- if $\mathfrak{M}, i \vDash \chi$, then PA $\vdash \mathrm{S_i} \rightarrow \chi^*$
- if $\mathfrak{M}, i \nvDash \chi$, then PA $\vdash \mathrm{S_i} \rightarrow \neg\chi^*$

The satisfiability relation of the Kripke model is thus simulated by a provable implication whose antecedent is the sentence S_i simulating the node i, and the consequent the translation of the sentence true at the node i.

The assumption $\mathfrak{M}, 1 \nvDash \varphi$ will then imply PA $\vdash S_1 \rightarrow \neg\varphi^*$, whence, by propositional logic and the Hilbert-Bernays-Löb derivability conditions (see section 1),

$$\mathrm{PA} \vdash \neg\mathrm{Bew}_{\mathrm{PA}}(\overline{\ulcorner \neg S_1 \urcorner}) \rightarrow \neg\mathrm{Bew}_{\mathrm{PA}}(\overline{\ulcorner \varphi^* \urcorner})$$

In order to get the desired result, there is the true (but undecidable) sentence S_0 s.t.

$$\mathrm{PA} \vdash \mathrm{S_0} \rightarrow \neg\mathrm{Bew}_{\mathrm{PA}}(\overline{\ulcorner \neg S_1 \urcorner})$$

whereby then PA $\vdash \mathrm{S_0} \rightarrow \neg\mathrm{Bew}_{\mathrm{PA}}(\overline{\ulcorner \varphi^* \urcorner})$. Since PA is sound, $\mathrm{S_0} \rightarrow \neg\mathrm{Bew}_{\mathrm{PA}}(\overline{\ulcorner \varphi^* \urcorner})$ is true. And since also S_0 is true, also $\neg\mathrm{Bew}_{\mathrm{PA}}(\overline{\ulcorner \varphi^* \urcorner})$ has to be true. But the truth of $\neg\mathrm{Bew}_{\mathrm{PA}}(\overline{\ulcorner \varphi^* \urcorner})$ just means that φ^* is not a theorem of PA.

3.1 Construction of the Solovay sentences

The model $\mathfrak{M}$ is extended to a model $\mathfrak{M}' = \langle W', R', V' \rangle$ by adding the node 0 that has access to all the nodes of the original model, and carries the same propositional information as node 1.

S_i is the arithmetized statement that a certain function h, whose range is W', has the limit i. A node i in $\mathfrak{M}'$ is then simulated by a sentence of PA asserting that i is the limit of h. Informally, the function h is defined as:

$$h(0) = 0$$

$$h(m+1) = \begin{cases} j & \text{if } Prf\left(m, \overline{\ulcorner \neg S_j \urcorner}\right) \text{ and } h(m)R'j \\ h(m) & \text{else} \end{cases}$$

Then S_i can be defined as:

$$S_i := \exists y \forall x \left(x \geq y \rightarrow \exists z \left(z = \bar{i} \wedge h\left(x\right) = z\right)\right)$$

The function h "climbs up" the accessibility relation R' of $\mathfrak{M}'$: theoretically, each new value is the successor (in $\mathfrak{M}'$) of the previous one. The function will start at 0, and assume another value j only if it is provable that it will *not* stay there, i.e. that j will *not* be the limit of this function. Since the range of the function is finite, and R' is irreflexive and transitive, it is clear that the function will never leave its initial value 0. Since S_j, which is used to define the function h in the first place, makes a claim about the limit of h, the definition of h is obviously self-referential. However, h can be given a nice formal definition by making use of the Generalised Diagonal Lemma.

3.2 Properties of the Solovay sentences

1. PA $\vdash \mathrm{S_i} \rightarrow \neg \mathrm{S_j}$ if $0 \leq i < j \leq n$
2. PA $\vdash S_0 \vee S_1 \vee \cdots \vee S_n$
3. PA $\vdash S_i \rightarrow \neg \mathrm{Bew}_{\mathrm{PA}}\left(\overline{\ulcorner \neg S_j \urcorner}\right)$ for $iR'j$
4. PA $\vdash S_i \rightarrow \mathrm{Bew}_{\mathrm{PA}}\left(\overline{\ulcorner \neg S_i \urcorner}\right)$ for $i \geq 1$
5. PA $\vdash S_i \rightarrow \mathrm{Bew}_{\mathrm{PA}}(\overline{\ulcorner \bigvee_{j:iR'j} S_j \urcorner})$ for $i \geq 1$
6. For $i \geq 1$, S_i is false (hence, because of 2., S_0 has to be true)
7. For $i \geq 0$, S_i is undecidable in PA

Strange as these properties may seem, each of them is crucial for proving **Lemma** 1 in section 3. In the next section, I will show how these properties also guarantee the existence of a bisimulation between a finite irreflexive and transitive Kripke structure and a Kripke structure whose domain consists of certain nonstandard models of PA.

4 The new perspective

In this section, it will be shown that the finite Kripke structure $\mathfrak{M}$, introduced above as a countermodel to a sentence φ s.t. $\mathrm{GL} \nvdash \varphi$, is bisimilar to the model $\mathfrak{N} = \langle W_{\mathfrak{N}}, \preccurlyeq, V_{\mathfrak{N}} \rangle$, where $W_{\mathfrak{N}}$ is the set of nonstandard models $\mathcal{M}$ of PA s.t. $\mathcal{M} \nvDash S_0$, i.e. $W_{\mathfrak{N}}$ contains all models of PA that do not satisfy the true Solovay sentence S_0. The arithmetical completeness of GL is an immediate consequence of this result.

4.1 The bisimulation

Whereas usually, bisimilar nodes are required to carry identical atomic information, this condition has to be loosened here, as there are no proposition letters in $\mathcal{L}_{\mathrm{PA}}$. Bisimilar nodes are required to carry identical information only in the sense that a propositional letter p is true at a node in $\mathfrak{M}$ if and only if its realisation p^* is satisfied at its related node in $\mathfrak{N}$. The realisation $*$ will be like the one used in Solovay's proof, i.e. $p^* = \bigvee_{i:\mathrm{iV}_{\mathfrak{M}}\mathrm{p}} S_i$. The bisimulation $\mathrm{Z} \subseteq (W_{\mathfrak{N}} \times W_{\mathfrak{M}})$ is defined by making use of the Solovay sentences:

$$(\mathcal{M}, i) \,\epsilon \mathrm{Z} \quad :\Leftrightarrow \mathcal{M} \vDash S_i$$

By using the properties of the Solovay sentences in section 3.2, and the result in section 2.3, the three conditions a bisimulation has to satisfy are easily obtained:

(i)

To show: If $(\mathcal{M}, i) \,\epsilon \mathrm{Z}$, then $\mathfrak{N}, \mathcal{M} \vDash p^* \iff \mathfrak{M}, i \vDash p$

$\Leftarrow$ Assume $\mathfrak{M}, i \vDash p$. By definition of $*$, S_i is a disjunct in p^*, i.e. $p^* = S_i \vee \dots$. Since $(\mathcal{M}, i) \,\epsilon \mathrm{Z}$, we have, by definition of Z, $\mathfrak{N}, \mathcal{M} \vDash S_i$, and thus also, by propositional logic, $\mathfrak{N}, \mathcal{M} \vDash S_i \vee \dots$.

$\Rightarrow$ Assume $\mathfrak{N}, \mathcal{M} \vDash p^*$, i.e. $\mathfrak{N}, \mathcal{M} \vDash S_{j_1} \vee S_{j_2} \vee \dots$ where $\mathfrak{M}, j_k \vDash p$. Thus, there has to be k s.t. $\mathfrak{N}, \mathcal{M} \vDash S_{j_k}$. On the other hand, because $(\mathcal{M}, i) \,\epsilon \mathrm{Z}$, also $\mathfrak{N}, \mathcal{M} \vDash S_i$. But since $\mathfrak{N}, \mathcal{M} \vDash S_i \rightarrow \neg S_j$ for all $i \neq j$ (by 1. in 3.2), we must have that $j_k = i$, and hence $\mathfrak{M}, i \vDash p$.

(ii) (The forth condition)

To show: If $(\mathcal{M}, i) \,\epsilon \mathrm{Z}$ and $\mathcal{N} \preccurlyeq \mathcal{M}$, then there is j s.t. iRj and $(\mathcal{N}, j) \,\epsilon \mathrm{Z}$
Since $(\mathcal{M}, i) \,\epsilon \mathrm{Z}$, $\mathfrak{N}, \mathcal{M} \vDash S_i$. Since $\mathrm{PA} \vdash \mathrm{S_i} \rightarrow \mathrm{Bew}_{\mathrm{PA}}\left(\overline{\ulcorner \bigvee_{\mathrm{j:iRj}} \mathrm{S_j} \urcorner}\right)$[9] (by 5. in 3.2), we have $\mathfrak{N}, \mathcal{M} \vDash \mathrm{Bew}_{\mathrm{PA}}\left(\overline{\ulcorner \bigvee_{j:iRj} S_j \urcorner}\right)$. By the result in section 2.3, $\mathfrak{N}, \mathcal{N} \vDash \bigvee_{j:iRj} S_j$. I.e. there is j, iRj s.t. $\mathcal{N} \vDash S_j$, and hence $(\mathcal{N}, j) \,\epsilon \mathrm{Z}$.

[9] For nodes $i \geq 1$ which interest us here, R' is the same as R.

(iii)(The back condition)

To show: If $(\mathcal{M}, i)\,\epsilon\mathrm{Z}$ and iRj, then there is $\mathcal{N}$ s.t. $\mathcal{N} \preccurlyeq \mathcal{M}$ and $(\mathcal{N}, j)\,\epsilon\mathrm{Z}$
Since $(\mathcal{M}, i)\,\epsilon\mathrm{Z}$, $\mathfrak{N}, \mathcal{M} \vDash S_i$. Since $\mathrm{PA} \vdash \mathrm{S_i} \rightarrow \neg\mathrm{Bew_{PA}}\left(\ulcorner\neg\mathrm{S_j}\urcorner\right)$ if iRj, (by 3. in 3.2), we have $\mathfrak{N}, \mathcal{M} \vDash \neg\mathrm{Bew_{PA}}\left(\ulcorner\neg S_j\urcorner\right)$. By the result in section 2.3, there is $\mathcal{N}$, $\mathcal{N} \preccurlyeq \mathcal{M}$ s.t. $\mathfrak{N}, \mathcal{N} \vDash S_j$ (because $\mathcal{M}$ contains a consistency statement for $\mathcal{N}$). Thus also $(\mathcal{N}, j)\,\epsilon\mathrm{Z}$.

Finishing the proof

Using the bisimulation just established, and the result in section 2.3, it is easily obtained that all nonstandard models $\mathcal{M}$ of PA with $\mathcal{M} \vDash S_i$ are "modally equivalent" to the node i in the finite Kripke model $\mathfrak{M}$, in the sense that $\mathfrak{M}.i \vDash \varphi$ if and only if $\mathfrak{N}, \mathcal{M} \vDash \varphi^*$. The arithmetical completeness of GL is an immediate consequence of this result.

If GL$\nvdash\varphi$ for some φ, then, by completeness of GL with respect to finite transitive converse well-founded frames, there is a model $\mathfrak{M} = \langle\{1, \ldots, n\}, R, V\rangle$ s.t. $\mathfrak{M}, 1 \nvDash \varphi$. Corresponding to this model, there are the Solovay sentences $S_1, \ldots, S_n$.

Given the bisimulation, the node 1 in $\mathfrak{M}$ is bisimilar to all nodes $\mathcal{M}$ (i.e. nonstandard models of PA) in $\mathfrak{N}$ s.t. $\mathcal{M} \vDash S_1$ (models with this property are guaranteed to exist by the fact S_1 is independent of PA). Thus, we have $\mathcal{M} \vDash \neg\varphi^*$ for all such models. But since $\mathcal{M}$ is a model of PA, this means that PA $\nvdash \varphi^*$: if φ^* would be a theorem PA, we would have $\mathcal{M} \vDash \varphi^*$ for all models $\mathcal{M}$ of PA.

4.2 Comparison of the two proofs

Essentially, both proofs rely on the capacity of the Solovay sentences to bring modal properties into the realm of Peano Arithmetic. But whereas Solovay's original proof simulates a finite Kripke model inside the *formal system* of PA, the proof presented in this article establishes a direct correspondence between the Kripke model and an arithmetically defined model. In order to construct this embedding of structures, however, the main technical ingredients of the original proof, i.e. the Solovay sentences, are needed, and also some additional ones, in particular the notion of strong interpretability. As a payoff, the completeness of GL follows in a very intuitive and obvious way: given a model of GL where $\neg\varphi$ is true at a node, we immediately obtain a model of PA where $\neg\varphi*$ is true. By contrast, the original proof has to apply to soundness of PA, and to the *meaning* of $\mathrm{Bew_{PA}}\left(\ulcorner\varphi\urcorner\right)$ in order to get the desired result. Thus, whereas Solovay's original proof is technically more simple, one might find the proof presented in this article intuitively more appealing.

When looking more carefully into the details of the two proofs, we see that the bisimulation is basically a substitute for Lemma 1. In Solovay's original proof, one node of $\mathfrak{M}$ is simulated by a sentence independent of PA. In the proof presented in this article, each node is simulated by a nonstandard model of PA in which that independent sentence holds.

Lemma 1 is proved by induction on the complexity of the modal formula ψ. Like the proof of the bisimulation in 4.1, this proof by induction depends crucially on the definition of p^* as the disjunction of all S_i s.t. $\mathfrak{M}, i \vDash p$, and the properties of the Solovay sentences in section 3.2. In fact, the structural similarity between the two proofs is rather fine-grained. In the table below, I have listed the non-trivial steps of the two proofs, and the properties of the Solovay sentences that these steps rely on.

The Bisimulation	Lemma 1	
$(\mathcal{M}, i)\,\epsilon\mathrm{Z}$, $\mathfrak{M}, i \vDash p \Rightarrow \mathfrak{N}, \mathcal{M} \vDash p^*$	$\mathfrak{M}, i \vDash p \Rightarrow \mathrm{PA} \vdash S_i \to p^*$	Def. of p^*
$(\mathcal{M}, i)\,\epsilon\mathrm{Z}$, $\mathfrak{N}, \mathcal{M} \vDash p^* \Rightarrow \mathfrak{M}, i \vDash p$	$\mathfrak{M}, i \nvDash p \Rightarrow \mathrm{PA} \vdash S_i \to \neg p^*$	Property 1.
The forth condition	$\mathfrak{M}, i \vDash \Box\psi \Rightarrow \mathrm{PA} \vdash \mathrm{S_i} \to \Box\psi^*$	Property 5.
The back condition	$\mathfrak{M}, i \nvDash \Box\psi \Rightarrow \mathrm{PA} \vdash \mathrm{S_i} \to \neg\Box\psi^*$	Property 3.

$\mathrm{Property\,1:}\ \mathrm{PA} \vdash \mathrm{S_i} \to \neg\mathrm{S_j}$ if $0 \leq i < j \leq n$

$\mathrm{Property\,3:}\ \mathrm{PA} \vdash S_i \to \neg\mathrm{Bew_{PA}}\left(\ulcorner\neg S_j\urcorner\right)$ for iRj

$\mathrm{Property\,5:}\ \mathrm{PA} \vdash S_i \to \mathrm{Bew_{PA}}(\overline{\ulcorner\bigvee_{j:iRj} S_j\urcorner})$ for $i \geq 1$

References

[1] Boolos, George, *The Logic of Provability*, Cambridge University Press, New York, 1993.

[2] Feferman, Solomon, *Arithmetization of metamathematics in a generalized setting*, Fundamenta Mathematicae. 49 (1960).

[3] Lindström, Per, *Aspects of Incompleteness, 2nd edition*, Association of Symbolic Logic, Natick, Massachusetts, 2003.