

# Self-referential theories

Samuel A. Alexander\*

*Department of Mathematics, the Ohio State University*

September 30, 2019

## Abstract

We study the structure of families of theories in the language of arithmetic extended to allow these families to refer to one another and to themselves. If a theory contains schemata expressing its own truth and expressing a specific Turing index for itself, and contains some other mild axioms, then that theory is untrue. We exhibit some families of true self-referential theories that barely avoid this forbidden pattern.

## 1 Introduction

This is a paper about families of r.e. theories, each capable of referring to itself and the others. Many of this paper's results first appeared in the author's dissertation [1]. There, they were stated in terms of families of interacting mechanical knowing agents. Here, we will speak instead of families of self-referential r.e. theories. We hope this will more directly expose the underlying mathematics.

In epistemology, it is well-known that a (suitably idealized) truthful knowing machine capable of arithmetic, logic, and self-reflection, cannot know its own truth and its own code. This is due, in various guises, to authors such as Lucas [7], Benacerraf [3], Reinhardt [10], Penrose [8], and Putnam [9]. In terms of self-referential theories, a true theory satisfying certain assumptions cannot contain schemata stating its own truth and its own Gödel number (if such a theory did exist, we could program a machine knower that knows precisely its consequences). Reinhardt conjectured, and Carlson proved [5], a truthful machine knower can know (in a local sense, i.e., expressed by infinite schemata rather than a single axiom) that it is truthful and has some code, without knowing which. A true self-referential theory can (in a local sense) state its own truth and recursive enumerability. We showed [2] that, alternatively, a truthful machine can (in a local sense) exactly know its own code, if not required to know its own truth. A true theory can state (in a local sense) its own Gödel number.

Our goal is to generalize the above consistency results to multiple theories. The paper contains four main findings. In the following list of promises, except where otherwise stated,  $\prec$  is an r.e. well-founded partial-order on  $\omega$ , and *expresses* is meant in the local (infinite schema) sense.

1. There are true theories  $(T_i)_{i \in \omega}$  such that  $T_i$  expresses a Gödel number of  $T_j$  (all  $i, j$ ) and  $T_i$  expresses the truth of  $T_j$  (all  $j \prec i$ ).
2. There are true theories  $(T_i)_{i \in \omega}$  such that  $T_i$  expresses a Gödel number of  $T_j$  ( $j \prec i$ ), the truth of  $T_j$  ( $j \preceq i$ ), and the fact that  $T_j$  has some Gödel number (all  $i, j$ ).
3. If  $\prec$  is ill-founded, and if we extend the base language to include a predicate for computable ordinals and require the theories to include rudimentary facts about them, then 1 and 2 fail.
4. Finally, if we do not extend the base language as in 3, then there do exist ill-founded r.e. partial orders  $\prec$  such that 1 and 2 hold.

Our proofs of 1 and 2 are constructive, but the proof of 4 is nonconstructive. In short, if 4 were false, either of 1 or 2 could be used to define the set  $WF$  of r.e. well-founded partial orders of  $\omega$  using nothing but arithmetic and a truth predicate  $\text{Tr}$  for arithmetic. This is impossible since  $WF$  is  $\Pi_1^1$ -complete and  $\text{Tr}$  is  $\Delta_1^1$ .

---

\*Email: alexander@math.ohio-state.edu

## 2 Preliminaries

To us, *theory* and *schema* mean *set of sentences* (a *sentence* is a formula with no free variables).

**Definition 1.** (Standard Definitions)

1. When a first-order structure is clear from context, an *assignment* is a function  $s$  mapping first-order variables into the universe of that structure. If  $x$  is a variable and  $u$  is an element of the universe,  $s(x|u)$  is the assignment that agrees with  $s$  except that it maps  $x$  to  $u$ .
2. We write  $\mathcal{M} \models \phi[s]$  to indicate that the first-order structure  $\mathcal{M}$  satisfies the formula  $\phi$  relative to the assignment  $s$ . We write  $\mathcal{M} \models \phi$  just in case  $\mathcal{M} \models \phi[s]$  for every assignment  $s$ . If  $T$  is a theory,  $\mathcal{M} \models T$  means that  $\mathcal{M} \models \phi$  for every  $\phi \in T$ .
3. We write  $\text{FV}(\phi)$  for the set of free variables of  $\phi$ .
4. We write  $\phi(x|t)$  for the result of substituting term  $t$  for variable  $x$  in  $\phi$ .
5.  $\mathcal{L}_{\text{PA}}$  is the language of Peano arithmetic, with constant symbol  $0$  and function symbols  $S, +, \cdot$  with the usual arities. If  $\mathcal{L}$  extends  $\mathcal{L}_{\text{PA}}$ , an  $\mathcal{L}$ -structure *has standard first-order part* if it has universe  $\mathbb{N}$  and interprets  $0, S, +$  and  $\cdot$  as intended.
6. We define  $\mathcal{L}_{\text{PA}}$ -terms  $\bar{n}$  ( $n \in \mathbb{N}$ ), called *numerals*, so that  $\bar{0} \equiv 0$  and  $\overline{n+1} \equiv S(\bar{n})$ .
7. We fix a computable bijection  $\langle \bullet, \bullet, \bullet \rangle : \mathbb{N}^3 \rightarrow \mathbb{N}$ . Being computable, this is  $\mathcal{L}_{\text{PA}}$ -definable, so we may freely act as if  $\mathcal{L}_{\text{PA}}$  contained a function symbol for this bijection. Similarly we may act as if  $\mathcal{L}_{\text{PA}}$  contained a binary predicate symbol  $\bullet \in W_\bullet$  for membership in the  $n$ th r.e. set  $W_n$ .
8. Whenever a computable language is clear from context,  $\phi \mapsto \ulcorner \phi \urcorner$  denotes Gödel numbering.
9. A *valid* formula is one that is true in every structure.
10. A *universal closure* of  $\phi$  is a sentence  $\forall x_1 \cdots \forall x_n \phi$  where  $\text{FV}(\phi) \subseteq \{x_1, \dots, x_n\}$ . We write  $\text{ucl}(\phi)$  to denote a generic universal closure of  $\phi$ .

Note that if  $\mathcal{M}$  is a structure and  $\psi$  is a universal closure of  $\phi$ , in order to prove  $\mathcal{M} \models \psi$  it suffices to let  $s$  be an arbitrary assignment and show  $\mathcal{M} \models \phi[s]$ .

To formalize self-referential theories, we employ an extension of first-order logic where languages may contain new unary connective symbols. This logic is borrowed from [5].

**Definition 2.** (The Base Logic) A language  $\mathcal{L}$  of the *base logic* is a first-order language  $\mathcal{L}_0$  together with a class of symbols called *operators*. Formulas of  $\mathcal{L}$  are defined as usual, with the clause that  $\mathbf{T}_i \models \phi$  is a formula whenever  $\phi$  is a formula and  $\mathbf{T}_i \models$  is an operator. Syntactic parts of Definition 1 extend to the base logic in obvious ways (we define  $\text{FV}(\mathbf{T}_i \models \phi) = \text{FV}(\phi)$ ). An  $\mathcal{L}$ -structure  $\mathcal{M}$  is a first-order  $\mathcal{L}_0$ -structure  $\mathcal{M}_0$  together with a function that takes one operator  $\mathbf{T}_i \models$ , one  $\mathcal{L}$ -formula  $\phi$ , and one assignment  $s$ , and outputs either True or False—in which case we write  $\mathcal{M} \models \mathbf{T}_i \models \phi[s]$  or  $\mathcal{M} \not\models \mathbf{T}_i \models \phi[s]$ , respectively—satisfying the following three requirements.

1. Whether or not  $\mathcal{M} \models \mathbf{T}_i \models \phi[s]$  does not depend on  $s(x)$  if  $x \notin \text{FV}(\phi)$ .
2. If  $\phi$  and  $\psi$  are *alphabetic variants* (meaning that one is obtained from the other by renaming bound variables so as to respect the binding of the quantifiers), then  $\mathcal{M} \models \mathbf{T}_i \models \phi[s]$  if and only if  $\mathcal{M} \models \mathbf{T}_i \models \psi[s]$ .
3. For variables  $x$  and  $y$  such that  $y$  is substitutable for  $x$  in  $\mathbf{T}_i \models \phi$ ,  $\mathcal{M} \models \mathbf{T}_i \models \phi(x|y)[s]$  if and only if  $\mathcal{M} \models \mathbf{T}_i \models \phi[s(x|s(y))]$ .

The definition of  $\mathcal{M} \models \phi[s]$  for arbitrary  $\mathcal{L}$ -formulas is obtained from this by induction. Semantic parts of Definition 1 extend to the base logic in obvious ways.

Traditionally the operator  $\mathbf{T}_i \models$  would be written  $K_i$ , and the formula  $K_i \phi$  would be read like “agent  $i$  knows  $\phi$ ”. For the present paper, the added intuition would not be worth the philosophical distraction.

**Theorem 3.** (Completeness and compactness) Suppose  $\mathcal{L}$  is an r.e. language in the base logic.

1. The set of valid  $\mathcal{L}$ -formulas is r.e.
2. For any r.e.  $\mathcal{L}$ -theory  $\Sigma$ ,  $\{\phi : \Sigma \models \phi\}$  is r.e.
3. There is an effective procedure, given (a Gödel number of) an r.e.  $\mathcal{L}$ -theory  $\Sigma$ , to find (a Gödel number of)  $\{\phi : \Sigma \models \phi\}$ .
4. If  $\Sigma$  is an  $\mathcal{L}$ -theory and  $\Sigma \models \phi$ , there are  $\sigma_1, \dots, \sigma_n \in \Sigma$  such that<sup>1</sup>  $\sigma_1 \rightarrow \dots \rightarrow \sigma_n \rightarrow \phi$  is valid.

*Proof.* By interpreting the base logic within first-order logic (for details see [1]).  $\square$

**Definition 4.** If  $\mathcal{L}$  is a first-order language and  $I$  is an index set, let  $\mathcal{L}(I)$  be the language (in the base logic) consisting of  $\mathcal{L}$  along with operators  $\mathbf{T}_i$  for all  $i \in I$ .

In case  $I$  is a singleton,  $\mathcal{L}_{\text{PA}}(I)$  is a form of Shapiro's [11] language of Epistemic Arithmetic.

**Definition 5.**

- For any  $\mathcal{L}_{\text{PA}}(I)$ -formula  $\phi$  with  $\text{FV}(\phi) = \{x_1, \dots, x_n\}$ , and for assignment  $s$  (into  $\mathbb{N}$ ), let  $\phi^s$  be the sentence

$$\phi^s \equiv \phi(x_1 | \overline{s(x_1)}) \cdots (x_n | \overline{s(x_n)})$$

obtained by replacing all free variables in  $\phi$  by numerals for their  $s$ -values.

- For any language  $\mathcal{L}$  extending  $\mathcal{L}_{\text{PA}}$ , if  $\mathcal{M}$  is an  $\mathcal{L}$ -structure, then  $\mathcal{M}$  is said to *interpret formulas by substitution* if  $\mathcal{M}$  has standard first-order part and the following property holds: for every  $\mathcal{L}$ -formula  $\phi$  and assignment  $s$ ,  $\mathcal{M} \models \phi[s]$  if and only if  $\mathcal{M} \models \phi^s$ .

For example, if  $s(x) = 0$  and  $s(y) = 2$  then  $(\forall z(x = y + z))^s \equiv \forall z(0 = S(S(0)) + z)$ .

**Definition 6.** If  $\mathbf{T} = (T_i)_{i \in I}$  is an  $I$ -indexed family of  $\mathcal{L}_{\text{PA}}(I)$ -theories and  $\mathcal{N}$  is an  $\mathcal{L}_{\text{PA}}(I)$ -structure, we say  $\mathcal{N} \models \mathbf{T}$  if  $\mathcal{N} \models T_i$  for all  $i \in I$ .

**Definition 7.** Suppose  $\mathbf{T} = (T_i)_{i \in I}$  is an  $I$ -indexed family of  $\mathcal{L}_{\text{PA}}(I)$ -theories. The *intended structure* for  $\mathbf{T}$  is the  $\mathcal{L}_{\text{PA}}(I)$ -structure  $\mathcal{M}_{\mathbf{T}}$  with standard first-order part, interpreting the operators  $\mathbf{T}_i$  ( $i \in I$ ) as follows:

$$\mathcal{M}_{\mathbf{T}} \models \mathbf{T}_i \phi[s] \text{ if and only if } T_i \models \phi^s.$$

If  $\mathcal{M}_{\mathbf{T}} \models \mathbf{T}$ , we say  $\mathbf{T}$  is *true*.

**Lemma 8.** For any family  $\mathbf{T} = (T_i)_{i \in I}$  of  $\mathcal{L}_{\text{PA}}(I)$ -theories,  $\mathcal{M}_{\mathbf{T}}$  interprets formulas by substitution.

*Proof.* In other words, we must show that for every  $\mathcal{L}_{\text{PA}}(I)$ -formula  $\phi$  and assignment  $s$ ,  $\mathcal{M}_{\mathbf{T}} \models \phi[s]$  if and only if  $\mathcal{M}_{\mathbf{T}} \models \phi^s$ . The proof is a straightforward induction.  $\square$

**Definition 9.** By the *axioms of Peano arithmetic for  $\mathcal{L}_{\text{PA}}(I)$*  we mean the axioms of Peano arithmetic, with induction extended to  $\mathcal{L}_{\text{PA}}(I)$ .

**Lemma 10.** For any  $\mathcal{L}_{\text{PA}}(I)$ -structure  $\mathcal{M}$ , if  $\mathcal{M}$  interprets formulas by substitution, then  $\mathcal{M}$  satisfies the axioms of Peano arithmetic for  $\mathcal{L}_{\text{PA}}(I)$ .

*Proof.* Let  $\mathcal{M}$  be any  $\mathcal{L}_{\text{PA}}(I)$ -structure which interprets formulas by substitution. This means  $\mathcal{M}$  has standard-first order part and for every formula  $\phi$  and assignment  $s$ ,  $\mathcal{M} \models \phi[s]$  if and only if  $\mathcal{M} \models \phi^s$ .

Let  $\sigma$  be an axiom of Peano arithmetic for  $\mathcal{L}_{\text{PA}}(I)$ . If  $\sigma$  is not an instance of induction, then  $\mathcal{M} \models \sigma$  since  $\mathcal{M}$  has standard first-order part. But suppose  $\sigma$  is  $\text{ucl}(\phi(x|0) \rightarrow \forall x(\phi \rightarrow \phi(x|S(x))) \rightarrow \forall x\phi)$ . To see  $\mathcal{M} \models \sigma$ , let  $s$  be an arbitrary assignment and assume  $\mathcal{M} \models \phi(x|0)[s]$  and  $\mathcal{M} \models \forall x(\phi \rightarrow \phi(x|S(x)))[s]$ . By assumption,  $\mathcal{M} \models \phi^{s(x|0)}$  and  $\forall m \in \mathbb{N}$ , if  $\mathcal{M} \models \phi^{s(x|m)}$  then  $\mathcal{M} \models \phi(x|S(x))^{s(x|m)}$ . Evidently  $\phi(x|S(x))^{s(x|m)} \equiv \phi^{s(x|m+1)}$ . By mathematical induction,  $\forall m \in \mathbb{N}$ ,  $\mathcal{M} \models \phi^{s(x|m)}$ . By assumption,  $\mathcal{M} \models \forall x\phi[s]$ .  $\square$

<sup>1</sup>We write  $A \rightarrow B \rightarrow C$  for  $A \rightarrow (B \rightarrow C)$ , and likewise for longer chains.

**Definition 11.** Suppose  $\mathbf{T} = (T_i)_{i \in I}$  is a family  $\mathcal{L}_{\text{PA}}(I)$ -theories. If  $\mathbf{T}^+ = (T_i^+)_{i \in I}$  is another such family, we say  $\mathbf{T} \subseteq \mathbf{T}^+$  if  $T_i \subseteq T_i^+$  for every  $i \in I$ . If  $T$  is a single  $\mathcal{L}_{\text{PA}}(I)$ -theory, we say  $T \subseteq \mathbf{T}$  if  $T \subseteq T_i$  for all  $i \in I$ . If  $\mathbf{T}^1 = (T_i^1)_{i \in I}$  and  $\mathbf{T}^2 = (T_i^2)_{i \in I}$  are families of  $\mathcal{L}_{\text{PA}}(I)$ -theories,  $\mathbf{T}^1 \cup \mathbf{T}^2$  is the family  $\mathbf{T}' = (T_i')_{i \in I}$  where each  $T_i' = T_i^1 \cup T_i^2$ . Arbitrary unions  $\bigcup_{n \in \mathbb{N}} \mathbf{T}^n$  are defined similarly.

**Definition 12.** Suppose  $\mathbf{T} = (T_i)_{i \in I}$  is a family of  $\mathcal{L}_{\text{PA}}(I)$ -theories. For each  $i \in I$ , we say  $T_i$  is  $\mathbf{T}_i \models$ -closed if  $\mathbf{T}_i \models \phi \in T_i$  whenever  $\phi \in T_i$ . We say  $\mathbf{T}$  is *closed* if each  $T_i$  is  $\mathbf{T}_i \models$ -closed.

**Definition 13.** If  $I$  is an r.e. index set, a family  $\mathbf{T} = (T_i)_{i \in I}$  is *r.e.* just in case  $\{(\phi, i) : \phi \in T_i\}$  is r.e.

### 3 Generic Axioms

If  $\mathbf{T}$  is a family of theories whose truth was in doubt, and if we state a theorem removing that doubt, we often state more: that  $\mathbf{T} \cup \mathbf{S}$  is true, where  $\mathbf{S}$  is some background theory of provability, including non-controversial things like Peano arithmetic or the schema  $\text{ucl}(\mathbf{T}_i \models (\phi \rightarrow \psi) \rightarrow \mathbf{T}_i \models \phi \rightarrow \mathbf{T}_i \models \psi)$ . The choice of  $\mathbf{S}$  is somewhat arbitrary, or at best based on tradition. We will avoid this arbitrary choice by stating results in the form: “ $\mathbf{T}$  is true together with any background theory of provability such that...”

**Definition 14.** A family  $\mathbf{T}$  of  $\mathcal{L}_{\text{PA}}(\omega)$ -theories is *closed-r.e.-generic* if  $\mathbf{T}$  is r.e. and  $\mathcal{M}_{\mathbf{T}'} \models \mathbf{T}$  for every closed r.e. family  $\mathbf{T}' \supseteq \mathbf{T}$  of  $\mathcal{L}_{\text{PA}}(\omega)$ -theories.

**Lemma 15.** If  $\mathbf{T}$  is a union of closed-r.e.-generic families and  $\mathbf{T}$  is r.e., then  $\mathbf{T}$  is closed-r.e.-generic.

*Proof.* Straightforward. □

**Definition 16.** For  $j \in I$  and for  $T$  an  $\mathcal{L}_{\text{PA}}(I)$ -theory, we write  $[T]_j$  for the family  $\mathbf{T} = (T_i)_{i \in I}$  where  $T_j = T$  and  $T_i = \emptyset$  for all  $i \neq j$ .

The following lemma provides building blocks that can be combined in diverse ways, via Lemma 15, to form background theories of provability. In the following lemma, part 4 is the whole reason for the “closed” in “closed-r.e.-generic”, and part 7 is part of the reason for the “r.e.”.

**Lemma 17.** For any  $i, j \in \omega$ , each of the following families is closed-r.e.-generic.

1.  $[S]_i$  where  $S$  is: ( $j$ -Deduction) the schema  $\text{ucl}(\mathbf{T}_j \models (\phi \rightarrow \psi) \rightarrow \mathbf{T}_j \models \phi \rightarrow \mathbf{T}_j \models \psi)$ .
2.  $[S]_i$  where  $S$  is: (Assigned Validity) the schema  $\phi^s$  ( $\phi$  valid,  $s$  an assignment).
3.  $[\text{Assigned Validity}]_j \cup [S]_i$  where  $S$  is: ( $j$ -Validity)  $\text{ucl}(\mathbf{T}_j \models \phi)$  for  $\phi$  valid.
4.  $[\text{Assigned Validity}]_j \cup [j\text{-Validity}]_j \cup [j\text{-Deduction}]_j \cup [S]_i$  where  $S$  is:  
 $(j\text{-Introspection})$  the schema  $\text{ucl}(\mathbf{T}_j \models \phi \rightarrow \mathbf{T}_j \models \mathbf{T}_j \models \phi)$ .
5.  $[S]_i$  where  $S$  is the set of axioms of Peano arithmetic for  $\mathcal{L}_{\text{PA}}(\omega)$ .
6.  $[S]_i$  where  $S$  is any r.e. set of true arithmetic sentences.
7.  $[S]_i$  where  $S$  is: ( $j$ -SMT) (See [5] and [10])  $\text{ucl}(\exists e \forall x (\mathbf{T}_j \models \phi \leftrightarrow x \in W_e)), e \notin \text{FV}(\phi)$ .
8.  $\mathbf{T} \cup [S]_i$  where  $\mathbf{T} = (T_k)_{k \in \omega}$  is closed-r.e.-generic and  $S$  is the schema  $\mathbf{T}_j \models \phi$  ( $\phi \in T_j$ ).

*Proof.*

(1) Let  $\mathbf{T}' = (T'_k)_{k \in \omega}$  be any closed r.e. family of  $\mathcal{L}_{\text{PA}}(\omega)$ -theories such that  $\mathbf{T}' \supseteq [S]_i$  where  $S$  is  $j$ -Deduction. We must show  $\mathcal{M}_{\mathbf{T}'} \models [S]_i$ . In other words we must show  $\mathcal{M}_{\mathbf{T}'} \models \text{ucl}(\mathbf{T}_j \models (\phi \rightarrow \psi) \rightarrow \mathbf{T}_j \models \phi \rightarrow \mathbf{T}_j \models \psi)$  for any  $\phi, \psi$ . Let  $s$  be an assignment and assume  $\mathcal{M}_{\mathbf{T}'} \models \mathbf{T}_j \models (\phi \rightarrow \psi)[s]$  and  $\mathcal{M}_{\mathbf{T}'} \models \mathbf{T}_j \models \phi[s]$ , we must show  $\mathcal{M}_{\mathbf{T}'} \models \mathbf{T}_j \models \psi[s]$ . By Definition of  $\mathcal{M}_{\mathbf{T}'}$ ,  $T'_j \models (\phi \rightarrow \psi)^s$  and  $T'_j \models \phi^s$ . Clearly  $(\phi \rightarrow \psi)^s \equiv \phi^s \rightarrow \psi^s$  so by modus ponens  $T'_j \models \psi^s$ , that is,  $\mathcal{M}_{\mathbf{T}'} \models \psi[s]$ .

(2) Let  $\mathbf{T}' = (T'_k)_{k \in \omega}$  be a closed r.e. superset of  $[S]_i$  where  $S$  is Assigned Validity. We must show  $\mathcal{M}_{\mathbf{T}'} \models [S]_i$ . If  $\phi \in [S]_i$  then  $\phi$  is  $\phi_0^s$  for some valid  $\phi_0$  and some assignment  $s$ . Since  $\phi_0$  is valid,  $\mathcal{M}_{\mathbf{T}'} \models \phi_0[s]$ . By Lemma 8,  $\mathcal{M}_{\mathbf{T}'} \models \phi_0^s$ .

(3) By Theorem 3,  $[\text{Assigned Validity}]_j \cup [j\text{-Validity}]_i$  is r.e. Let  $\mathbf{T}' = (T'_k)_{k \in \omega}$  be any closed r.e. family of  $\mathcal{L}_{\text{PA}}(\omega)$ -theories such that  $T'_j$  contains Assigned Validity and  $T'_i$  contains  $j$ -Validity. We must show  $\mathcal{M}_{\mathbf{T}'}$  satisfies Assigned Validity and  $j$ -Validity. For Assigned Validity, let  $\phi$  be valid and  $s$  an assignment. Since  $\phi$  is valid,  $\mathcal{M}_{\mathbf{T}'} \models \phi[s]$ , so by Lemma 8,  $\mathcal{M}_{\mathbf{T}'} \models \phi^s$  as desired. For  $j$ -Validity, let  $\phi$  be valid and  $s$  an assignment. Since  $T'_j$  contains Assigned Validity,  $T'_j \models \phi^s$ , so by definition of  $\mathcal{M}_{\mathbf{T}'}$ ,  $\mathcal{M}_{\mathbf{T}'} \models \mathbf{T}_j \models \phi[s]$ .

(4) Recursive enumerability is by Theorem 3. Let  $\mathbf{T}' = (T'_k)_{k \in \omega}$  be any closed r.e. family of  $\mathcal{L}_{\text{PA}}(\omega)$ -theories such that  $T'_j$  contains Assigned Validity,  $j$ -Validity and  $j$ -Deduction, and  $T'_i$  contains  $j$ -Introspection. That  $\mathcal{M}_{\mathbf{T}'}$  satisfies Assigned Validity and  $j$ -Validity is as in (3). That  $\mathcal{M}_{\mathbf{T}'}$  satisfies  $j$ -Deduction is straightforward. For  $j$ -Introspection, let  $s$  be an assignment and assume  $\mathcal{M}_{\mathbf{T}'} \models \mathbf{T}_j \models \phi[s]$ , we will show  $\mathcal{M}_{\mathbf{T}'} \models \mathbf{T}_j \models \mathbf{T}_j \models \phi[s]$ . Since  $\mathcal{M}_{\mathbf{T}'} \models \mathbf{T}_j \models \phi[s]$ ,  $T'_j \models \phi^s$ . By Theorem 3, there are  $\sigma_1, \dots, \sigma_n \in T'_j$  such that  $\sigma_1 \rightarrow \dots \rightarrow \sigma_n \rightarrow \phi^s$  is valid. Since  $T'_j$  contains  $j$ -Validity,  $T'_j \models \mathbf{T}_j \models (\sigma_1 \rightarrow \dots \rightarrow \sigma_n \rightarrow \phi^s)$ . By repeated applications of  $j$ -Deduction contained in  $T'_j$ ,  $T'_j \models \mathbf{T}_j \models \sigma_1 \rightarrow \dots \rightarrow \mathbf{T}_j \models \sigma_n \rightarrow \mathbf{T}_j \models (\phi^s)$ . Since  $\mathbf{T}'$  is closed,  $T'_j$  is  $\mathbf{T}_j \models$ -closed and so contains  $\mathbf{T}_j \models \sigma_1, \dots, \mathbf{T}_j \models \sigma_n$ . So  $T'_j \models (\mathbf{T}_j \models \phi)^s$  and  $\mathcal{M}_{\mathbf{T}'} \models \mathbf{T}_j \models \mathbf{T}_j \models \phi[s]$ .

(5) Follows from Lemma 10.

(6) Trivial.

(7) Suppose  $\mathbf{T}' = (T'_i)_{i \in \omega}$  is a closed r.e. family of  $\mathcal{L}_{\text{PA}}(\omega)$ -theories and  $\mathbf{T}' \supseteq [S]_i$  where  $S$  is  $j$ -SMT. We must show  $\mathcal{M}_{\mathbf{T}'} \models [S]_i$ . That is, given  $\phi$  with  $e \notin \text{FV}(\phi)$ , we must show  $\mathcal{M}_{\mathbf{T}'} \models \text{ucl}(\exists e \forall x (\mathbf{T}_j \models \phi \leftrightarrow x \in W_e))$ . Let  $s$  be an assignment and let  $x_1, \dots, x_k = \text{FV}(\phi) \setminus \{x\}$ . Since  $T'_j$  is r.e., by the  $S$ - $m$ - $n$  theorem there is some  $n$  such that  $W_n = \{m : T'_j \models \phi(x|\overline{m})(x_1|\overline{s(x_1)}) \dots (x_k|\overline{s(x_k)})\}$ . Since  $e \notin \text{FV}(\phi)$ , and  $\mathcal{M}_{\mathbf{T}'}$  has standard first-order part, it follows that  $\mathcal{M}_{\mathbf{T}'} \models \forall x (\mathbf{T}_j \models \phi \leftrightarrow x \in W_e)[s(e|n)]$ .

(8) Suppose  $\mathbf{T}' = (T'_i)_{i \in \omega} \supseteq \mathbf{T} \cup [S]_i$  where  $\mathbf{T} = (T_i)_{i \in \omega}$  is closed-r.e.-generic and  $S$  is the schema  $\mathbf{T}_j \models \phi$  ( $\phi \in T_j$ ). Right away  $\mathcal{M}_{\mathbf{T}'} \models \mathbf{T}$  because  $\mathbf{T}$  is closed-r.e.-generic. It remains to show that  $\mathcal{M}_{\mathbf{T}'} \models [S]_i$ , i.e., that  $\mathcal{M}_{\mathbf{T}'} \models S$ . Fix  $\phi \in T_j$  and let  $s$  be any assignment. Since  $\phi$  is a sentence,  $\phi \equiv \phi^s$  and thus  $T_j \models \phi^s$ . Since  $T'_j \supseteq T_j$ ,  $T'_j \models \phi^s$ . By definition of  $\mathcal{M}_{\mathbf{T}'}$ ,  $\mathcal{M}_{\mathbf{T}'} \models \mathbf{T}_j \models \phi[s]$ .  $\square$

## 4 First Consistency Result: Prioritizing Exact Codes

The following theorem fulfils the first promise from the introduction.

**Theorem 18.** Suppose  $\prec$  is an r.e. well-founded partial order on  $\omega$  and  $\mathbf{T}^0 = (T_i^0)_{i \in \omega}$  is closed-r.e.-generic. For each  $n \in \mathbb{N}$ , let  $\mathbf{T}(n) = (T_i(n))_{i \in \omega}$  where each  $T_i(n)$  is the smallest  $\mathbf{T}_i \models$ -closed theory containing the following:

1. The axioms in  $T_i^0$ .
2.  $\forall x (\mathbf{T}_j \models \phi \leftrightarrow \langle \overline{\phi}, \overline{j}, x \rangle \in W_{\overline{n}})$  whenever  $j \in \omega$ ,  $\text{FV}(\phi) \subseteq \{x\}$ .
3.  $\text{ucl}(\mathbf{T}_j \models \phi \rightarrow \phi)$  whenever  $j \prec i$ .

There is some  $n \in \mathbb{N}$  such that  $\mathbf{T}(n)$  is true.

*Proof.* By the  $S$ - $m$ - $n$  Theorem, there is a total computable  $f : \mathbb{N} \rightarrow \mathbb{N}$  such that  $\forall n \in \mathbb{N}$ ,

$$W_{f(n)} = \{ \langle \overline{\phi}, \overline{i}, m \rangle : \text{FV}(\phi) \subseteq \{x\} \text{ and } T_i(n) \models \phi(x|\overline{m}) \}.$$

Using the Recursion Theorem, fix  $n \in \mathbb{N}$  such that  $W_{f(n)} = W_n$ . For brevity write  $\mathbf{T}$  for  $\mathbf{T}(n)$  and  $T_i$  for  $T_i(n)$ . We will show  $\mathcal{M}_{\mathbf{T}} \models \mathbf{T}$ . This is a self-referential statement: to show  $T_i$  is true includes showing  $\mathcal{M}_{\mathbf{T}} \models \text{ucl}(\mathbf{T}_j \models \phi \rightarrow \phi)$ , which is essentially the statement that  $T_j$  is true. Hence the restriction  $j \prec i$ , which allows induction since  $\prec$  is well founded. We will show, by  $\prec$ -induction on  $i$ , that  $\mathcal{M}_{\mathbf{T}} \models T_i$  for every  $i \in \omega$ . Fix  $i \in \omega$  and assume  $\mathcal{M}_{\mathbf{T}} \models T_j$  for all  $j \prec i$ . Suppose  $\sigma \in T_i$ , we will show  $\mathcal{M}_{\mathbf{T}} \models \sigma$ .

**Case 1:**  $\sigma \in T_i^0$ . Then  $\mathcal{M}_{\mathbf{T}} \models \sigma$  because  $\mathbf{T}^0$  is closed-r.e.-generic and  $\mathbf{T} \supseteq \mathbf{T}^0$  is closed r.e.

**Case 2:**  $\sigma$  is  $\forall x(\mathbf{T}_j \models \phi \leftrightarrow \langle \overline{\Gamma\phi}, \bar{j}, x \rangle \in W_{\bar{n}})$  for some  $j \in \omega$ ,  $\text{FV}(\phi) \subseteq \{x\}$ . Let  $s$  be an assignment,  $m \in \mathbb{N}$ . The following are equivalent.

$$\begin{array}{ll}
\mathcal{M}_{\mathbf{T}} \models \mathbf{T}_j \models \phi[s(x|m)] & \\
T_j \models \phi^{s(x|m)} & \text{(Definition of } \mathcal{M}_{\mathbf{T}}) \\
T_j \models \phi(x|\bar{m}) & \text{(Since } \text{FV}(\phi) \subseteq \{x\}) \\
\langle \overline{\Gamma\phi}, \bar{j}, m \rangle \in W_{\bar{n}} & \text{(By definition of } n) \\
\mathcal{M}_{\mathbf{T}} \models \langle \overline{\Gamma\phi}, \bar{j}, \bar{m} \rangle \in W_{\bar{n}} & (\mathcal{M}_{\mathbf{T}} \text{ has standard first-order part}) \\
\mathcal{M}_{\mathbf{T}} \models \langle \overline{\Gamma\phi}, \bar{j}, x \rangle \in W_{\bar{n}}[s(x|m)]. & \text{(Lemma 8)}
\end{array}$$

**Case 3:**  $\sigma$  is  $\text{ucl}(\mathbf{T}_j \models \phi \rightarrow \phi)$  for some  $j \prec i$ . Let  $s$  be an assignment and assume  $\mathcal{M}_{\mathbf{T}} \models \mathbf{T}_j \models \phi[s]$ . This means  $T_j \models \phi^s$ . By our  $\prec$ -induction hypothesis,  $\mathcal{M}_{\mathbf{T}} \models T_j$ , so  $\mathcal{M}_{\mathbf{T}} \models \phi^s$ . By Lemma 8,  $\mathcal{M}_{\mathbf{T}} \models \phi[s]$ .

**Case 4:**  $\sigma$  is only present in  $T_i$  because of the clause that  $T_i$  is  $\mathbf{T}_i$ -closed. Then  $\sigma$  is  $\mathbf{T}_i \models \sigma_0$  for some  $\sigma_0 \in T_i$ . Being in  $T_i$ ,  $\sigma_0$  is a sentence, so for any assignment  $s$ ,  $\sigma_0 \equiv \sigma_0^s$ ,  $T_i \models \sigma_0^s$ , and finally  $\mathcal{M}_{\mathbf{T}} \models \mathbf{T}_i \models \sigma_0[s]$ .

By  $\prec$ -induction,  $\mathcal{M}_{\mathbf{T}} \models T_i$  for all  $i \in \omega$ . This shows  $\mathcal{M}_{\mathbf{T}} \models \mathbf{T}$ , that is,  $\mathbf{T}$  is true.  $\square$

The first promise from the introduction is met: for any r.e. well-founded partial order  $\prec$  on  $\omega$ , there are theories  $(T_n)_{n \in \omega}$  such that  $\forall i, j, k \in \omega$  with  $j \prec i$ ,  $T_i$  expresses the truth of  $T_j$ , and  $T_i$  expresses a Gödel number of  $T_k$ . In order to fulfil the second promise we will extend Carlson's notion of *stratification* to the case of multiple operators, and introduce *stratifiers*, a tool used to deal with subtleties that arise when multiple self-referential theories refer to one another.

In [2] the technique behind Theorem 18 was used to exhibit a machine that knows its own code.

## 5 Stratification

For the second promise from the introduction, we need to prove a result like Theorem 18 where  $T_i$  includes  $\text{ucl}(\mathbf{T}_j \models \phi \rightarrow \phi)$  for all  $j \preceq i$ , not just  $j \prec i$ . This rules out the direct  $\prec$ -induction of the type used above. Induction on formula complexity will not work either: we would need to show all of  $T_i$  consistent just to show  $\mathcal{M}_{\mathbf{T}} \models \mathbf{T}_i \models (1 = 0) \rightarrow (1 = 0)$ . Instead, we will use ordinal induction. But there are no ordinals anywhere in sight. To obtain ordinals to induct on, we will modify the theories we care about, in a process called *stratification*. We will start with some informal motivational remarks. Readers who would like to advance directly to the formal definitions can safely skip Subsection 5.1.

### 5.1 Motivation for Stratification

As explained above, we would like to invoke ordinal induction, but there are no ordinals in sight. In order to make ordinal induction relevant, we will do the following. We will extend the background language to contain not only the operators  $\mathbf{T}_i \models$  ( $i \in \omega$ ), but also operators  $\mathbf{T}_i^\alpha \models$  ( $i \in \omega$ ,  $\alpha \in \epsilon_0 \cdot \omega$ ). And instead of focusing directly on  $T_i$ , we will focus on a theory  $U_i$  such that the result  $U_i^-$  of erasing superscripts from  $U_i$  is  $U_i^- = T_i$ . The intended interpretation of  $\mathbf{T}_i^\alpha \models \phi[s]$  will be  $U_i \cap \alpha \models \phi^s$ , where  $U_i \cap \alpha$  is the set of axioms of  $U_i$  whose superscripts are  $< \alpha$ . Thus, we may think of  $U_i$  as a version of  $T_i$  with extra information about the structure of  $T_i$ . We will show that for all naturals  $m > n > 0$ ,  $U_i \cap (\epsilon_0 \cdot m)$  and  $U_i \cap (\epsilon_0 \cdot n)$  have essentially the same consequences, except for the superscripts occurring in those consequences. We will use this, after proving that  $U_i$  holds, to conclude that  $T_i$  also holds.

Suppose we would like  $T_i$  to contain the axiom  $\mathbf{T}_i \models (1 + 1 = 2)$ . Then, as we carry out the procedure in the above paragraph, we would ensure that  $U_i$  contain all sentences of the form  $\mathbf{T}_i^\alpha \models (1 + 1 = 2)$ . This would have the side effect that for any  $\beta > \alpha$ ,  $U_i \cap \beta \models \mathbf{T}_i^\alpha \models (1 + 1 = 2)$ , so that  $\mathbf{T}_i^\beta \models \mathbf{T}_i^\alpha \models (1 + 1 = 2)$  would hold in structures with the intended interpretation.

Next, suppose that for every arithmetical sentence  $\phi$ , we would like  $T_i$  to include

$$\mathbf{T}_i \models \phi \rightarrow \mathbf{T}_i \models \mathbf{T}_i \models \phi.$$

Then we would arrange that  $U_i$  contain

$$\mathbf{T}_i^\alpha \models \phi \rightarrow \mathbf{T}_i^\beta \models \mathbf{T}_i^\alpha \models \phi$$

(whenever  $\beta > \alpha$ ). The reason for the  $\beta$  is as follows. The intended interpretation of  $\mathbf{T}_i^\alpha \models \phi$  shall be  $U_i \cap \alpha \models \phi$ . Thus, it would make no sense to put the axiom  $\mathbf{T}_i^\alpha \models \phi \rightarrow \mathbf{T}_i^\alpha \models \mathbf{T}_i^\alpha \models \phi$  into  $U_i$ : the fact that  $U_i \cap \alpha \models \phi$  does not generally imply that  $U_i \cap \alpha \models \mathbf{T}_i^\alpha \models \phi$ , since  $U_i \cap \alpha$  is limited to formulas in which all superscripts are  $< \alpha$ . At least  $\mathbf{T}_i^\alpha \models \phi \rightarrow \mathbf{T}_i^\beta \models \mathbf{T}_i^\alpha \models \phi$  is plausible.

Again, suppose that for some  $j \prec i$ , we would like for  $T_i$  to include

$$\mathbf{T}_i \models (\mathbf{T}_j \models (1 = 0) \rightarrow (1 = 0)).$$

We would arrange that  $U_i$  contain (for all  $\alpha$ ):

$$\mathbf{T}_i^\alpha \models (\mathbf{T}_j \models (1 = 0) \rightarrow (1 = 0)).$$

Note the lack of superscript on  $\mathbf{T}_j \models$ . The intuition is that  $U_i$  is a version of  $T_i$  with extra information about the structure of  $T_i$  (namely, that said structure arises from an increasing family of theories), but without any additional information about the structure of  $T_j$ .

Similarly, suppose we would like  $T_i$  to include

$$\mathbf{T}_j \models (\mathbf{T}_i \models (1 = 0)) \rightarrow \mathbf{T}_i \models (1 = 0).$$

We would arrange that  $U_i$  contain (for each  $\alpha$ ):

$$\mathbf{T}_j \models (\mathbf{T}_i \models (1 = 0)) \rightarrow \mathbf{T}_i^\alpha \models (1 = 0).$$

Note the lack of superscript on the  $\mathbf{T}_i \models$  within the scope of  $\mathbf{T}_j \models$ . As above, the intuition is that  $U_i$  is a version of  $T_i$  with extra information about the structure of  $T_i$ . It does not have any extra information about the structure of  $T_j$ —not even about what  $T_j$  says about  $T_i$ . This is important because, when  $j \prec i$ , we would like  $T_i$  to contain axioms declaring, essentially, the Gödel number of  $T_j$ . This Gödel number would be hardcoded into such axioms, and thus there would be no hope of such axioms remaining true if  $T_j$  were changed.

## 5.2 Stratification Formal Details

To get a foothold for induction, instead of considering a particular theory  $T_i$ , we will be considering copies of  $T_i$  with ordinal-number superscripts added. To recover information about the original  $T_i$  from these modified theories, we will need to use sophisticated results from [4] about the structure of the ordinals.

**Definition 19.** We define a binary relation  $\leq_1$  on Ord by transfinite recursion so that for all  $\alpha, \beta \in \text{Ord}$ ,  $\alpha \leq_1 \beta$  if and only if  $\alpha \leq \beta$  and  $(\alpha, \leq, \leq_1)$  is a  $\Sigma_1$ -elementary substructure of  $(\beta, \leq, \leq_1)$ .

The following theorem is based on calculations from [4]. It was used by Carlson to prove Reinhardt's conjecture [5]. We state it here without proof.

**Theorem 20.**

1. The binary relation  $\leq_1$  is a recursive partial ordering on  $\epsilon_0 \cdot \omega$ .
2. For all positive integers  $m \leq n$ ,  $\epsilon_0 \cdot m \leq_1 \epsilon_0 \cdot n$ .
3. For any  $\alpha \leq \beta \in \text{Ord}$ ,  $\alpha \leq_1 \beta$  if and only if the following statement is true. For every finite set  $X \subseteq \alpha$  and every finite set  $Y \subseteq [\alpha, \beta)$ , there is a set  $X < \tilde{Y} < \alpha$  such that  $X \cup \tilde{Y} \cong_{(\leq, \leq_1)} X \cup Y$ .

The usefulness of Theorem 20 will appear in Theorem 32, but first we need some machinery.

**Definition 21.** Let  $\mathcal{I} = ((\epsilon_0 \cdot \omega) \times \omega) \sqcup \omega$ . Thus  $\mathcal{L}_{\text{PA}}(\mathcal{I})$  contains operators  $\mathbf{T}_{(\alpha, i)} \models$  for all  $\alpha \in \epsilon_0 \cdot \omega$ ,  $i \in \omega$ , along with operators  $\mathbf{T}_i \models$  for all  $i \in \omega$ . As abbreviation, we write  $\mathbf{T}_i^\alpha \models$  for  $\mathbf{T}_{(\alpha, i)} \models$ , and refer to  $\alpha$  as its *superscript*.

**Definition 22.** For any  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -formula  $\phi$ ,  $\text{On}(\phi) \subseteq \epsilon_0 \cdot \omega$  denotes the set of superscripts appearing in  $\phi$ .

**Definition 23.** Suppose  $i \in \omega$ . The  $i$ -stratified formulas of  $\mathcal{L}_{\text{PA}}(\mathcal{I})$  are defined as follows (where  $\phi$  ranges over  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -formulas).

1. If  $\phi$  is  $\mathbf{T}_j \models \phi_0$  for some  $j \neq i$ , then  $\phi$  is  $i$ -stratified if and only if  $\phi$  is an  $\mathcal{L}_{\text{PA}}(\omega)$ -formula.
2. If  $\phi$  is  $\mathbf{T}_j^\alpha \models \phi_0$  for some  $j \neq i$ , then  $\phi$  is not  $i$ -stratified.
3. If  $\phi$  is  $\mathbf{T}_i \models \phi_0$ , then  $\phi$  is not  $i$ -stratified.
4. If  $\phi$  is  $\mathbf{T}_i^\alpha \models \phi_0$ , then  $\phi$  is  $i$ -stratified if and only if  $\phi_0$  is  $i$ -stratified and  $\alpha > \text{On}(\phi_0)$ .
5. If  $\phi$  is  $\neg \phi_0$ ,  $\phi_1 \rightarrow \phi_2$ , or  $\forall x \phi_0$ , then  $\phi$  is  $i$ -stratified if and only if its immediate subformula(s) are.
6. If  $\phi$  is atomic, then  $\phi$  is  $i$ -stratified.

An  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -theory  $T$  is  $i$ -stratified if  $\phi$  is  $i$ -stratified whenever  $\phi \in T$ . An  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -formula  $\phi$  is *very  $i$ -stratified* if  $\phi$  is  $i$ -stratified and  $\text{On}(\phi) \subseteq \{\epsilon_0 \cdot 1, \epsilon_0 \cdot 2, \dots\}$ .

For example:

- $\mathbf{T}_7^\omega \models \mathbf{T}_7^5 \models (1 = 0) \rightarrow \mathbf{T}_8 \models (1 = 0)$  is 7-stratified but not 6- or 8-stratified.
- $\mathbf{T}_7^5 \models \mathbf{T}_7^\omega \models (1 = 0)$  is not 7-stratified, nor is  $\mathbf{T}_7^5 \models \mathbf{T}_7 \models (1 = 0)$ .
- $\mathbf{T}_7^5 \models \mathbf{T}_8 \models \mathbf{T}_7 \models (1 = 0)$  is 7-stratified but  $\mathbf{T}_7^5 \models \mathbf{T}_8 \models \mathbf{T}_7^4 \models (1 = 0)$  is not.

We will not make use of the following lemma, but we state it to further illuminate Definition 23.

**Lemma 24.** Suppose  $\phi$  is an  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -formula,  $i \in \omega$ . Then  $\phi$  is  $i$ -stratified if and only if all of the following conditions hold.

1. For all  $j \in \omega$  and  $\alpha \in \epsilon_0 \cdot \omega$ , if  $\mathbf{T}_j^\alpha \models$  occurs in  $\phi$ , then  $j = i$ .
2. Every occurrence of  $\mathbf{T}_i \models$  in  $\phi$  is inside the scope of  $\mathbf{T}_j \models$  for some  $j \neq i$ .
3.  $\mathbf{T}_i^\alpha \models$  never occurs in  $\phi$  inside the scope of  $\mathbf{T}_j \models$ , for any  $\alpha \in \epsilon_0 \cdot \omega$  or any  $j \in \omega$ .
4. For all  $\alpha, \beta \in \epsilon_0 \cdot \omega$ , if  $\mathbf{T}_i^\alpha \models$  occurs in  $\phi$  inside the scope of  $\mathbf{T}_i^\beta \models$ , then  $\beta > \alpha$ .

*Proof.* Straightforward. □

**Definition 25.** Suppose  $X \subseteq \epsilon_0 \cdot \omega$  and  $h : X \rightarrow \epsilon_0 \cdot \omega$  is order preserving. For each  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -formula  $\phi$ , let  $h(\phi)$  be the  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -formula obtained by applying  $h$  to every superscript in  $\phi$  that is in  $X$ .

For example if  $X = \{1, \omega\}$ ,  $h(1) = 0$ , and  $h(\omega) = \omega \cdot 2 + 1$ , then

$$h(\mathbf{T}_i^0 \models (1 = 0) \rightarrow \mathbf{T}_i^1 \models (1 = 0) \rightarrow \mathbf{T}_i^\omega \models (1 = 0)) \equiv \mathbf{T}_i^0 \models (1 = 0) \rightarrow \mathbf{T}_i^0 \models (1 = 0) \rightarrow \mathbf{T}_i^{\omega \cdot 2 + 1} \models (1 = 0).$$

**Definition 26.** Suppose  $X \subseteq \epsilon_0 \cdot \omega$  and  $h : X \rightarrow \epsilon_0 \cdot \omega$  is order preserving. For any  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -structure  $\mathcal{N}$ , we define an  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -structure  $h(\mathcal{N})$  that has the same universe as  $\mathcal{N}$ , agrees with  $\mathcal{N}$  on  $\mathcal{L}_{\text{PA}}(\omega)$ , and interprets  $\mathcal{L}_{\text{PA}}(\mathcal{I}) \setminus \mathcal{L}_{\text{PA}}(\omega)$  so that

$$h(\mathcal{N}) \models \mathbf{T}_i^\alpha \models \phi[s] \text{ if and only if } \mathcal{N} \models h(\mathbf{T}_i^\alpha \models \phi)[s].$$

**Lemma 27.** Suppose  $X \subseteq \epsilon_0 \cdot \omega$ ,  $h : X \rightarrow \epsilon_0 \cdot \omega$  is order preserving, and  $\mathcal{N}$  is an  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -structure. For any  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -formula  $\phi$  and assignment  $s$ ,  $h(\mathcal{N}) \models \phi[s]$  if and only if  $\mathcal{N} \models h(\phi)[s]$ .

*Proof.* By induction. □

**Corollary 28.** Suppose  $X \subseteq \epsilon_0 \cdot \omega$  and  $h : X \rightarrow \epsilon_0 \cdot \omega$  is order preserving. For any valid  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -formula  $\phi$ ,  $h(\phi)$  is valid.



*Proof.* For any  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -structure  $\mathcal{N}$  and assignment  $s$ ,  $h(\mathcal{N}) \models \phi[s]$  by validity, so  $\mathcal{N} \models h(\phi)[s]$  by Lemma 27.  $\square$

**Definition 29.** If  $X \subseteq \text{Ord}$  and  $h : X \rightarrow \text{Ord}$ , we call  $h$  a *covering* if  $h$  is order preserving and whenever  $x, y \in X$  and  $x \leq_1 y$ ,  $h(x) \leq_1 h(y)$ .

**Definition 30.** Suppose  $i \in \omega$ . A set  $T$  of  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -sentences is  *$i$ -unistratified* if the following conditions hold:

1.  $T$  is  $i$ -stratified.
2. (Uniformity) Whenever  $\phi \in T$ ,  $X \subseteq \epsilon_0 \cdot \omega$ ,  $\text{On}(\phi) \subseteq X$ , and  $h : X \rightarrow \epsilon_0 \cdot \omega$  is a covering, then  $h(\phi) \in T$ .

**Definition 31.** If  $T$  is an  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -theory and  $\alpha \in \epsilon_0 \cdot \omega$ , let  $T \cap \alpha$  be the set  $\{\phi \in T : \text{On}(\phi) \subseteq \alpha\}$  of sentences in  $T$  that do not contain any superscripts  $\geq \alpha$ .

**Theorem 32.** (The Collapse Theorem) Suppose  $T$  is an  $i$ -unistratified  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -theory.

1. If  $n$  is a positive integer and  $\text{On}(\phi) \subseteq \epsilon_0 \cdot n$ , then  $T \models \phi$  if and only if  $T \cap (\epsilon_0 \cdot n) \models \phi$ .
2. If  $\alpha \leq_1 \beta$  and  $\text{On}(\phi) \subseteq \alpha$ , then  $T \cap \alpha \models \phi$  if and only if  $T \cap \beta \models \phi$ .

*Proof.* Note that since  $T$  is  $i$ -unistratified, in particular  $T$  is  $i$ -stratified. We will prove (1), the proof of (2) is similar.

( $\Leftarrow$ ) Immediate since  $T \cap (\epsilon_0 \cdot n) \subseteq T$ .

( $\Rightarrow$ ) Assume  $T \models \phi$ . By Theorem 3 there are  $\sigma_1, \dots, \sigma_k \in T$  such that

$$\Phi \equiv \sigma_1 \rightarrow \dots \rightarrow \sigma_k \rightarrow \phi$$

is valid. Let  $X = \text{On}(\Phi) \cap (\epsilon_0 \cdot n)$ ,  $Y = \text{On}(\Phi) \cap [\epsilon_0 \cdot n, \infty)$ , note  $|X|, |Y| < \infty$ .

Since  $Y$  is finite, there is some integer  $n' > n$  such that  $Y \subseteq \epsilon_0 \cdot n'$ . By Theorem 20 part 2,  $\epsilon_0 \cdot n \leq_1 \epsilon_0 \cdot n'$ . By Theorem 20 part 3, there is some  $X < \tilde{Y} < \epsilon_0 \cdot n$  such that  $X \cup \tilde{Y} \cong_{(\leq, \leq_1)} X \cup Y$ .

Let  $h : X \cup Y \rightarrow X \cup \tilde{Y}$  be a  $(\leq, \leq_1)$ -isomorphism. Since  $\text{On}(\phi) \subseteq \epsilon_0 \cdot n$ ,  $h(\phi) = \phi$ . By Corollary 28,

$$h(\Phi) \equiv h(\sigma_1) \rightarrow \dots \rightarrow h(\sigma_k) \rightarrow \phi$$

is valid. Since  $T$  is  $i$ -unistratified,  $h(\sigma_1), \dots, h(\sigma_k) \in T$ . Finally since  $\text{range}(h) < \epsilon_0 \cdot n$ ,  $h(\sigma_1), \dots, h(\sigma_k) \in T \cap (\epsilon_0 \cdot n)$ , showing  $T \cap (\epsilon_0 \cdot n) \models \phi$ .  $\square$

Loosely speaking, what we have done in Theorem 32 is we have taken a proof of  $\phi$  and we have *collapsed* the proof, shrinking its ordinals by using Theorem 20 part 3.

**Definition 33.** For every  $i \in \omega$  we define the following  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -schema:

- ( $i$ -Collapse)  $\text{ucl}(\mathbf{T}_i^\alpha \models \phi \leftrightarrow \mathbf{T}_i^\beta \models \phi)$  whenever  $\mathbf{T}_i^\alpha \models \phi$  is  $i$ -stratified and  $\alpha \leq_1 \beta$ .

**Definition 34.** For any  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -formula  $\phi$ ,  $\phi^-$  is the result of erasing all superscripts from  $\phi$ . If  $T$  is an  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -theory,  $T^- = \{\sigma^- : \sigma \in T\}$ .

For example, if  $\phi$  is  $\mathbf{T}_5^\omega \models (1 = 0) \rightarrow \mathbf{T}_5^{\omega+1} \models \mathbf{T}_5^\omega \models (1 = 0)$ , then  $\phi^-$  is  $\mathbf{T}_5 \models (1 = 0) \rightarrow \mathbf{T}_5 \models \mathbf{T}_5 \models (1 = 0)$ .

**Lemma 35.** If  $T$  is  $i$ -unistratified then for every  $\phi \in T$  there is some  $\psi \in T$  such that  $\psi$  is very  $i$ -stratified and  $\psi^- \equiv \phi^-$ .

*Proof.* Let  $X = \text{On}(\phi) = \{\alpha_1 < \dots < \alpha_n\}$ ,  $Y = \{\epsilon_0 \cdot 1, \dots, \epsilon_0 \cdot n\}$ , and define  $h : X \rightarrow Y$  by  $h(\alpha_j) = \epsilon_0 \cdot j$ . Clearly  $h$  is order preserving; by Theorem 20 part 2,  $h$  is a covering. Since  $T$  is  $i$ -unistratified,  $T$  contains  $\psi \equiv h(\phi)$ . Clearly  $\psi$  is very  $i$ -stratified and  $\psi^- \equiv \phi^-$ .  $\square$

**Definition 36.** For any  $\mathcal{L}_{\text{PA}}(\omega)$ -structure  $\mathcal{N}$ , we define an  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -structure  $\mathcal{N}^-$  that has the same universe as  $\mathcal{N}$ , agrees with  $\mathcal{N}$  on  $\mathcal{L}_{\text{PA}}(\omega)$ , and interprets  $\mathcal{L}_{\text{PA}}(\mathcal{I}) \setminus \mathcal{L}_{\text{PA}}(\omega)$  as follows. For any  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -formula  $\phi$ ,  $\alpha \in \epsilon_0 \cdot \omega$ ,  $i \in \mathbb{N}$ , and assignment  $s$ ,

$$\mathcal{N}^- \models \mathbf{T}_i^\alpha \phi[s] \text{ if and only if } \mathcal{N} \models (\mathbf{T}_i^\alpha \phi)^-[s].$$

**Lemma 37.** Suppose  $\mathcal{N}$  is an  $\mathcal{L}_{\text{PA}}(\omega)$ -structure. For every  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -formula  $\phi$  and assignment  $s$ ,  $\mathcal{N}^- \models \phi[s]$  if and only if  $\mathcal{N} \models \phi^-[s]$ .

*Proof.* By induction. □

**Corollary 38.** If  $\phi$  is a valid  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -formula, then  $\phi^-$  is a valid  $\mathcal{L}_{\text{PA}}(\omega)$ -formula.

*Proof.* Similar to the proof of Corollary 28. □

A converse-like statement holds for Corollary 38 as well.

**Lemma 39.** For any valid  $\mathcal{L}_{\text{PA}}(\omega)$ -sentence  $\phi$  and  $i \in \omega$ , there is a valid very  $i$ -stratified  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -sentence  $\psi$  such that  $\psi^- \equiv \phi$ .

*Proof.* Let  $\psi \mapsto \psi^+$  be the function taking  $\mathcal{L}_{\text{PA}}(\omega)$ -formulas to  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -formulas defined as follows.

1. If  $\psi$  is atomic, or of the form  $\mathbf{T}_j \psi_0$  with  $j \neq i$ , then  $\psi^+ \equiv \psi$ .
2. If  $\psi$  is  $\mathbf{T}_i \psi_0$ , then  $\psi^+ \equiv \mathbf{T}_i^{\epsilon_0 \cdot n} \psi_0^+$ , where  $n = \min\{m \in \mathbb{N} : \epsilon_0 \cdot m > \text{On}(\psi_0^+)\}$ .
3. If  $\psi$  is  $\neg \psi_0$ ,  $\psi_0 \rightarrow \psi_1$ , or  $\forall x \psi_0$ , then  $\psi^+$  is  $\neg \psi_0^+$ ,  $\psi_1^+ \rightarrow \psi_2^+$ , or  $\forall x \psi_0^+$ , respectively.

It is straightforward to show  $\phi^+$  is very  $i$ -stratified. We claim  $\phi^+$  is valid. Let  $\mathcal{M}$  be any  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -structure, we will show  $\mathcal{M} \models \phi^+$ . Let  $\mathcal{M}^+$  be the  $\mathcal{L}_{\text{PA}}(\omega)$ -structure with the same universe as  $\mathcal{M}$ , which agrees with  $\mathcal{M}$  on the interpretation of arithmetic and of  $\mathbf{T}_j \psi$  for  $j \neq i$ , and which interprets  $\mathbf{T}_i \psi$  as follows:

$$\mathcal{M}^+ \models \mathbf{T}_i \psi[s] \text{ if and only if } \mathcal{M} \models (\mathbf{T}_i \psi)^+[s].$$

Since  $\phi$  is valid,  $\mathcal{M}^+ \models \phi$ . It follows that  $\mathcal{M} \models \phi^+$ . □

**Definition 40.** Let  $i \in \omega$ . We define the following  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -schemas.

- ( $i$ -Strativalidity)  $\text{ucl}(\mathbf{T}_i^\alpha \phi)$  whenever  $\phi$  is a valid  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -formula and  $\mathbf{T}_i^\alpha \phi$  is  $i$ -stratified.
- ( $i$ -Stratideduction)  $\text{ucl}(\mathbf{T}_i^\alpha (\phi \rightarrow \psi) \rightarrow \mathbf{T}_i^\alpha \phi \rightarrow \mathbf{T}_i^\alpha \psi)$  whenever this formula is  $i$ -stratified.

**Definition 41.** An  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -theory  $T$  is *i-straticlosed* if the following conditions hold:

1.  $T$  is  $i$ -unistratified.
2.  $T$  includes  $i$ -Strativalidity,  $i$ -Stratideduction and  $i$ -Collapse.
3. For every  $\phi \in T$ , if  $\mathbf{T}_i^\alpha \phi$  is  $i$ -stratified then  $\mathbf{T}_i^\alpha \phi \in T$ .

A family  $\mathbf{T} = (T_i)_{i \in \omega}$  is *straticlosed* if each  $T_i$  is  $i$ -straticlosed.

The following theorem serves as an omnibus of results from Section 5 of [5].

**Theorem 42.** (Proof Stratification) Suppose  $T$  is an  $i$ -straticlosed  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -theory. Then:

1. Whenever  $T \cap \alpha \models \phi$ ,  $\mathbf{T}_i^\alpha \phi$  is an  $i$ -stratified sentence, and  $\beta > \alpha$ , then  $T \cap \beta \models \mathbf{T}_i^\alpha \phi$ .
2. For any very  $i$ -stratified  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -sentences  $\rho$  and  $\sigma$ , if  $\rho^- \equiv \sigma^-$  then  $T \models \rho \leftrightarrow \sigma$ .
3. For any very  $i$ -stratified  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -sentence  $\phi$ ,  $T \models \phi$  if and only if  $T^- \models \phi^-$ .

*Proof.* Note that since  $T$  is  $i$ -straticlosed, in particular  $T$  is  $i$ -unistratified and hence,  $i$ -stratified.

**Claim 0:** Any time  $T \models \mathbf{T}_i^\alpha (\rho \leftrightarrow \sigma)$  and this is  $i$ -stratified,  $T \models \mathbf{T}_i^\alpha \rho \leftrightarrow \mathbf{T}_i^\alpha \sigma$ .

Assume the hypotheses. By  $i$ -Strativalidity,  $T \models \mathbf{T}_i^\alpha \models ((\rho \leftrightarrow \sigma) \rightarrow (\rho \rightarrow \sigma))$ . By  $i$ -Stratideduction,

$$\begin{aligned} T &\models \mathbf{T}_i^\alpha \models ((\rho \leftrightarrow \sigma) \rightarrow (\rho \rightarrow \sigma)) \rightarrow \mathbf{T}_i^\alpha \models (\rho \leftrightarrow \sigma) \rightarrow \mathbf{T}_i^\alpha \models (\rho \rightarrow \sigma) \\ \text{and } T &\models \mathbf{T}_i^\alpha \models (\rho \rightarrow \sigma) \rightarrow \mathbf{T}_i^\alpha \models \rho \rightarrow \mathbf{T}_i^\alpha \models \sigma. \end{aligned}$$

It follows that  $T \models \mathbf{T}_i^\alpha \models \rho \rightarrow \mathbf{T}_i^\alpha \models \sigma$ . The reverse implication is similar.

**Claim 1:** If  $T \cap \alpha \models \phi$ ,  $\mathbf{T}_i^\alpha \models \phi$  is an  $i$ -stratified sentence, and  $\beta > \alpha$ , then  $T \cap \beta \models \mathbf{T}_i^\alpha \models \phi$ .

Given  $T \cap \alpha \models \phi$ , there are  $\sigma_1, \dots, \sigma_n \in T \cap \alpha$  such that  $\sigma_1 \rightarrow \dots \rightarrow \sigma_n \rightarrow \phi$  is valid. By instances of  $i$ -Strativalidity and  $i$ -Stratideduction contained in  $T \cap \beta$ ,  $T \cap \beta \models \mathbf{T}_i^\alpha \models \phi$ .

**Claim 2:** If  $\rho$  and  $\sigma$  are very  $i$ -stratified  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -sentences and  $\rho^- \equiv \sigma^-$ , then  $T \models \rho \leftrightarrow \sigma$ .

By induction on  $\rho$ . Note that  $\rho$  is not of the form  $\mathbf{T}_j^\alpha \models \rho_0$  (with  $j \neq i$ ), as that is not  $i$ -stratified. If  $\rho$  is  $\mathbf{T}_j^\alpha \models \rho_0$  then  $\rho \equiv \rho^- \equiv \sigma^- \equiv \sigma$  and the claim is immediate.

The only nontrivial remaining case is when  $\rho$  is  $\mathbf{T}_i^\alpha \models \rho_0$ . Since  $\rho$  is very  $i$ -stratified, this implies  $\alpha = \epsilon_0 \cdot n$  (some positive integer  $n$ ) and  $\rho_0$  is very  $i$ -stratified. Since  $\sigma^- \equiv \rho^-$  and  $\sigma$  is very stratified, this implies  $\sigma \equiv \mathbf{T}_i^{\epsilon_0 \cdot m} \models \sigma_0$  for some positive integer  $m$  and very  $i$ -stratified  $\sigma_0$  with  $\sigma_0^- \equiv \rho_0^-$ . Assume  $m \leq n$ , the other case is similar.

By induction,  $T \models \rho_0 \leftrightarrow \sigma_0$ . By compactness, there is a natural  $\ell \geq n$  such that  $T \cap (\epsilon_0 \cdot \ell) \models \rho_0 \leftrightarrow \sigma_0$ . By Claim 1,  $T \models \mathbf{T}_i^{\epsilon_0 \cdot \ell} \models (\rho_0 \leftrightarrow \sigma_0)$ ; Claim 0 then gives  $T \models \mathbf{T}_i^{\epsilon_0 \cdot \ell} \models \rho_0 \leftrightarrow \mathbf{T}_i^{\epsilon_0 \cdot \ell} \models \sigma_0$ . The claim now follows since  $T$  contains  $i$ -Collapse and  $\epsilon_0 \cdot m \leq \epsilon_0 \cdot n \leq \epsilon_0 \cdot \ell$  (Theorem 20 part 2).

**Claim 3:** If  $\phi$  is a very  $i$ -stratified  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -sentence and  $T \models \phi$ , then  $T^- \models \phi^-$ .

By compactness, find  $\sigma_1, \dots, \sigma_n \in T$  such that  $\sigma_1 \rightarrow \dots \rightarrow \sigma_n \rightarrow \phi$  is valid. By Corollary 38, so is  $\sigma_1^- \rightarrow \dots \rightarrow \sigma_n^- \rightarrow \phi^-$ , witnessing  $T^- \models \phi^-$ .

**Claim 4:** If  $\phi$  is a very  $i$ -stratified  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -sentence and  $T^- \models \phi^-$ , then  $T \models \phi$ .

By compactness, there is a valid sentence

$$\Phi \equiv \sigma_1^- \rightarrow \dots \rightarrow \sigma_n^- \rightarrow \phi^-$$

where each  $\sigma_j \in T$ . By Lemma 39, there is a valid very  $i$ -stratified  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -sentence  $\Psi$  such that  $\Psi^- \equiv \Phi$ . And because  $\Psi^- \equiv \Phi$ , this implies

$$\Psi \equiv \sigma_1^* \rightarrow \dots \rightarrow \sigma_n^* \rightarrow \phi^*$$

where each  $(\sigma_j^*)^- \equiv \sigma_j^-$ ,  $(\phi^*)^- \equiv \phi^-$ , and  $\sigma_1^*, \dots, \sigma_n^*, \phi^*$  are very  $i$ -stratified.

By Lemma 35, there are very  $i$ -stratified  $\sigma_1^{**}, \dots, \sigma_n^{**} \in T$  with each  $(\sigma_j^{**})^- \equiv \sigma_j^- \equiv (\sigma_j^*)^-$ . By Claim 2,  $T \models \phi^* \leftrightarrow \phi$ , and for  $j = 1, \dots, n$ ,  $T \models \sigma_j^{**} \leftrightarrow \sigma_j^*$ . Thus

$$T \models (\sigma_1^{**} \rightarrow \dots \rightarrow \sigma_n^{**} \rightarrow \phi) \leftrightarrow \Psi,$$

and since  $\Psi$  is valid and the  $\sigma_j^{**} \in T$ , this shows  $T \models \phi$ .  $\square$

**Definition 43.** If  $\mathbf{T} = (T_i)_{i \in \omega}$  is a straticlosed family of  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -theories, its *stratification*, written  $\text{Str}(\mathbf{T})$ , is the family  $\text{Str}(\mathbf{T}) = (S_i)_{i \in \mathcal{I}}$ , where for every  $i \in \omega$ ,  $S_i = T_i^-$  and  $\forall \alpha \in \epsilon_0 \cdot \omega$ ,  $S_{(\alpha, i)} = T_i \cap \alpha$ .

**Theorem 44.** (The Stratification Theorem) Suppose  $\mathbf{T} = (T_i)_{i \in \omega}$  is a straticlosed family of  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -theories. For any  $i \in \omega$ , any very  $i$ -stratified  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -formula  $\phi$ , and any assignment  $s$ ,  $\mathcal{M}_{\text{Str}(\mathbf{T})} \models \phi[s]$  if and only if  $\mathcal{M}_{\text{Str}(\mathbf{T})} \models \phi^-[s]$ .

*Proof.* By induction on  $\phi$ . The only nontrivial case is when  $\phi$  is  $\mathbf{T}_i^\alpha \models \psi$ . Since  $\phi$  is very  $i$ -stratified,  $\psi$  is very  $i$ -stratified and we may write  $\alpha = \epsilon_0 \cdot n$  for some positive integer  $n$ ,  $\text{On}(\psi) \subseteq \epsilon_0 \cdot n$ . The following are equivalent.

$$\begin{aligned} \mathcal{M}_{\text{Str}(\mathbf{T})} &\models \mathbf{T}_i^{\epsilon_0 \cdot n} \models \psi[s] \\ T_i \cap (\epsilon_0 \cdot n) &\models \psi^s && \text{(Definition of } \mathcal{M}_{\text{Str}(\mathbf{T})}) \\ T_i &\models \psi^s && \text{(Theorem 32)} \\ T_i^- &\models (\psi^s)^- && \text{(Theorem 42)} \\ T_i^- &\models (\psi^-)^s && \text{(Clearly } (\psi^s)^- \equiv (\psi^-)^s) \\ \mathcal{M}_{\text{Str}(\mathbf{T})} &\models \mathbf{T}_i \models \psi^-[s]. && \text{(Definition of } \mathcal{M}_{\text{Str}(\mathbf{T})}) \end{aligned}$$

□

## 6 Stratifiers

In order to apply theorems from the previous section, it is necessary to work with families  $\mathbf{T} = (T_i)_{i \in \omega}$  where each  $T_i$  is  $i$ -stratified. If we want  $T_i^-$  to (locally) express the truthfulness of  $T_j^-$ , we cannot simply add a schema like  $\text{ucl}(\mathbf{T}_j \models \phi \rightarrow \phi)$  to  $T_i$ , because this is not necessarily  $i$ -stratified: for example, the particular instance  $\mathbf{T}_j \models \mathbf{T}_i \models (1 = 0) \rightarrow \mathbf{T}_i \models (1 = 0)$  is not  $i$ -stratified. But neither is, say,  $\mathbf{T}_j \models \mathbf{T}_i^\alpha \models (1 = 0) \rightarrow \mathbf{T}_i^\alpha \models (1 = 0)$ , where  $\mathbf{T}_i^\alpha$  occurs within the scope of  $\mathbf{T}_j$ . We will use a schema  $\text{ucl}(\mathbf{T}_j \models \phi \rightarrow \phi^+)$ , where  $\bullet^+$  varies over what we call  $i$ -stratifiers.

**Definition 45.** Suppose  $X \subseteq \epsilon_0 \cdot \omega$ ,  $|X| = \infty$ , and  $i \in \omega$ . The  $i$ -stratifier given by  $X$  is the function  $\phi \mapsto \phi^+$  taking  $\mathcal{L}_{\text{PA}}(\omega)$ -formulas to  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -formulas as follows.

1. If  $\phi$  is atomic or of the form  $\mathbf{T}_j \models \phi_0$  with  $j \neq i$ , then  $\phi^+ \equiv \phi$ .
2. If  $\phi$  is  $\mathbf{T}_i \models \phi_0$  then  $\phi^+ \equiv \mathbf{T}_i^\alpha \models \phi_0^+$  where  $\alpha = \min\{x \in X : x > \text{On}(\phi_0^+)\}$ .
3. If  $\phi$  is  $\neg\psi$ ,  $\psi \rightarrow \rho$ , or  $\forall x\psi$ , then  $\phi^+$  is  $\neg\psi^+$ ,  $\psi^+ \rightarrow \rho^+$  or  $\forall x\psi^+$ , respectively.

By an  $i$ -stratifier we mean an  $i$ -stratifier given by some  $X$ . By the  $i$ -veristratifier we mean the  $i$ -stratifier given by  $X = \{\epsilon_0 \cdot 1, \epsilon_0 \cdot 2, \dots\}$ .

For example, if  $\bullet^+$  is the  $i$ -veristratifier and  $j \neq i$  then

$$(\mathbf{T}_j \models \mathbf{T}_i \models (1 = 0) \rightarrow \mathbf{T}_i \models \mathbf{T}_i \models (1 = 0))^+ \equiv \mathbf{T}_j \models \mathbf{T}_i \models (1 = 0) \rightarrow \mathbf{T}_i^{\epsilon_0 \cdot 2} \models \mathbf{T}_i^{\epsilon_0} \models (1 = 0).$$

**Lemma 46.** Suppose  $Z \subseteq \epsilon_0 \cdot \omega$ ,  $h : Z \rightarrow \epsilon_0 \cdot \omega$  is order preserving,  $i \in \omega$ , and  $\bullet^+$  is an  $i$ -stratifier. For any  $\mathcal{L}_{\text{PA}}(\omega)$ -formula  $\theta$  with  $\text{On}(\theta^+) \subseteq Z$ , there is a computable  $i$ -stratifier  $\bullet^*$  with  $\theta^* \equiv h(\theta^+)$ .

*Proof.* Let  $X_0 = \{h(\alpha) : \alpha \in \text{On}(\theta^+)\}$ , let  $X = X_0 \cup \{\alpha \in \epsilon_0 \cdot \omega : \alpha > X_0\}$ , and let  $\bullet^*$  be the  $i$ -stratifier given by  $X$ . By induction, for every subformula  $\theta_0$  of  $\theta$ ,  $\theta_0^* \equiv h(\theta_0^+)$ . □

**Definition 47.** By a *stratifier-set*, we mean a finite set

$$I = \{\bullet^{+1}, \dots, \bullet^{+k}\}$$

where each  $\bullet^{+p}$  is an  $i_p$ -stratifier for some  $i_p \in \omega$ , and  $i_1, \dots, i_k$  are distinct. With  $I$  as above, we write  $\text{Indices}(I)$  for  $\{i_1, \dots, i_k\}$ . We say  $I$  is *computable* if each  $\bullet^{+p}$  is computable.

For example, if  $\bullet^{+1}$  is a 1-stratifier,  $\bullet^{+2}$  is a 5-stratifier, and  $\bullet^{+3}$  is a 2-stratifier, then  $I = \{\bullet^{+1}, \bullet^{+2}, \bullet^{+3}\}$  is a stratifier-set and  $\text{Indices}(I) = \{1, 5, 2\}$ . For a non-example, if  $\bullet^{*1}$  and  $\bullet^{*2}$  are distinct 1-stratifiers, then  $\{\bullet^{*1}, \bullet^{*2}\}$  is not a stratifier-set, because it fails the distinctness condition.

**Definition 48.**

1. Suppose  $\mathcal{N}$  is an  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -structure and  $I$  is a stratifier-set. We define an  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -structure  $\mathcal{N}^I$  as follows. The universe and interpretation of arithmetic of  $\mathcal{N}^I$  agree with those of  $\mathcal{N}$ , as do the interpretations of  $\mathbf{T}_i \models$  ( $i \notin \text{Indices}(I)$ ) and  $\mathbf{T}_i^\alpha \models$  (any  $\alpha, i$ ). For each  $i \in \text{Indices}(I)$ , let  $\bullet^+ \in I$  be the corresponding  $i$ -stratifier, and let  $\mathcal{N}^I$  interpret  $\mathbf{T}_i \models$  as follows. For any  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -formula  $\phi$  and assignment  $s$ , we consider two cases.
  - (a) If  $\phi$  is an  $\mathcal{L}_{\text{PA}}(\omega)$ -formula, then  $\mathcal{N}^I \models \mathbf{T}_i \models \phi[s]$  if and only if  $\mathcal{N} \models (\mathbf{T}_i \models \phi)^+[s]$ .
  - (b) If  $\phi$  is not an  $\mathcal{L}_{\text{PA}}(\omega)$ -formula, then  $\mathcal{N}^I \models \mathbf{T}_i \models \phi[s]$  if and only if  $\mathcal{N} \models \mathbf{T}_i \models \phi[s]$ .
2. For any  $i \in \omega$ , any  $i$ -stratifier  $\bullet^+$ , and any  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -structure  $\mathcal{N}$ , let  $\mathcal{N}^+ = \mathcal{N}^I$  where  $I = \{\bullet^+\}$  is the stratifier-set containing only  $\bullet^+$ .

Case 1b in Definition 48 is somewhat arbitrary. We will only ever really care about whether  $\mathcal{N}^I \models \mathbf{T}_i \models \phi[s]$  when  $\mathbf{T}_i \models \phi$  is  $j$ -stratified for some  $j$ . If  $\phi$  is not an  $\mathcal{L}_{\text{PA}}(\omega)$ -formula then  $\mathbf{T}_i \models \phi$  is not  $j$ -stratified for any  $j$ .

**Lemma 49.** (Compare Lemma 37) Suppose  $\mathcal{N}$  is an  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -structure,  $i \in \omega$ , and  $\bullet^+$  is an  $i$ -stratifier. For every  $\mathcal{L}_{\text{PA}}(\omega)$ -formula  $\phi$  and assignment  $s$ ,  $\mathcal{N}^+ \models \phi[s]$  if and only if  $\mathcal{N} \models \phi^+[s]$ .

*Proof.* By induction. □

**Lemma 50.** For any  $\mathcal{L}_{\text{PA}}(\omega)$ -formula  $\phi$ , any  $i \in \omega$ , and any  $i$ -stratifier  $\bullet^+$ ,  $\phi$  is valid if and only if  $\phi^+$  is valid.

*Proof.*

( $\Rightarrow$ ) Assume  $\phi$  is valid. For any  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -structure  $\mathcal{N}$  and assignment  $s$ ,  $\mathcal{N}^+ \models \phi[s]$  by validity, so  $\mathcal{N} \models \phi^+[s]$  by Lemma 49.

( $\Leftarrow$ ) By Corollary 38. □

**Lemma 51.** Suppose  $\mathcal{M}$  is an  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -structure,  $I_0$  is a stratifier-set,  $i \in \omega$ ,  $i \notin \text{Indices}(I_0)$ , and  $\bullet^+$  is an  $i$ -stratifier. Let  $I = I_0 \cup \{\bullet^+\}$ . Then  $\mathcal{M}^I = (\mathcal{M}^{I_0})^+$ . Furthermore,  $\mathcal{M}^+$  and  $\mathcal{M}^I$  agree on the interpretation of  $\mathbf{T}_i \models$ .

*Proof.* Straightforward. □

**Lemma 52.** Suppose  $i \in \omega$  and suppose  $\mathcal{M}$  is an  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -structure with the property that for every  $i$ -stratified  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -formula  $\phi$  and assignment  $s$ ,  $\mathcal{M} \models \phi[s]$  if and only if  $\mathcal{M} \models \phi^-[s]$ . Suppose  $I$  is a stratifier-set such that  $i \notin \text{Indices}(I)$ . Then for every  $i$ -stratified  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -formula  $\phi$  and assignment  $s$ ,  $\mathcal{M}^I \models \phi[s]$  if and only if  $\mathcal{M}^I \models \phi^-[s]$ .

*Proof.* By induction on  $\phi$ . Let  $s$  be an assignment. The only interesting cases are the following.

**Case 1:**  $\phi$  is  $\mathbf{T}_j \models \psi$  for some  $j$ . Then  $\phi^- \equiv \phi$  and the claim is trivial.

**Case 2:**  $\phi$  has the form  $\mathbf{T}_j^\alpha \models \psi$  for some  $j \neq i$ . Impossible, this is not  $i$ -stratified.

**Case 3:**  $\phi$  has the form  $\mathbf{T}_i^\alpha \models \psi$ . The following are equivalent:

$$\begin{aligned} \mathcal{M}^I &\models \mathbf{T}_i^\alpha \models \psi[s] \\ \mathcal{M} &\models \mathbf{T}_i^\alpha \models \psi[s] && (\mathcal{M} \text{ and } \mathcal{M}^I \text{ agree on } \mathbf{T}_i^\alpha \models) \\ \mathcal{M} &\models (\mathbf{T}_i^\alpha \models \psi)^-[s] && (\text{By hypothesis}) \\ \mathcal{M}^I &\models (\mathbf{T}_i^\alpha \models \psi)^-[s]. && (\text{Since } i \notin \text{Indices}(I), \mathcal{M} \text{ and } \mathcal{M}^I \text{ agree on } \mathbf{T}_i \models) \end{aligned}$$

□

**Lemma 53.** Suppose  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -structure  $\mathcal{M}$  is an instance of Definition 7, and suppose  $I$  is a stratifier-set. Then  $\mathcal{M}^I$  interprets formulas by substitution.

*Proof.* By induction on  $|I|$ . If  $|I| = 0$ , we are done by Lemma 8. Otherwise, we may decompose  $I$  as  $I = I_0 \cup \{\bullet^+\}$  where  $\bullet^+$  is an  $i$ -stratifier. By induction,  $\mathcal{M}^{I_0}$  interprets formulas by substitution (\*). By Lemma 51,  $\mathcal{M}^I = (\mathcal{M}^{I_0})^+$ .

By definition of interpreting formulas by substitution, for every  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -formula  $\phi$  and assignment  $s$ ,  $\mathcal{M}^{I_0} \models \phi[s]$  if and only if  $\mathcal{M}^{I_0} \models \phi^s$ . We must show that for every such  $\phi$  and  $s$ ,  $(\mathcal{M}^{I_0})^+ \models \phi[s]$  if and only if  $(\mathcal{M}^{I_0})^+ \models \phi^s$ .

We induct on  $\phi$ . By Definition 48,  $(\mathcal{M}^{I_0})$  and  $(\mathcal{M}^{I_0})^+$  agree on all symbols except  $\mathbf{T}_i \models$ , and they agree on  $\mathbf{T}_i \models \phi_0$  if  $\phi_0$  is not an  $\mathcal{L}_{\text{PA}}(\omega)$ -formula. Thus the only nontrivial case is when  $\phi$  is of the form  $\mathbf{T}_i \models \phi_0$  for some  $\mathcal{L}_{\text{PA}}(\omega)$ -formula  $\phi_0$ . Any such  $\phi$  is itself an  $\mathcal{L}_{\text{PA}}(\omega)$ -formula and thus susceptible to Lemma 49. The following are equivalent.

$$\begin{aligned} (\mathcal{M}^{I_0})^+ &\models \phi[s] \\ \mathcal{M}^{I_0} &\models \phi^+[s] && (\text{Lemma 49}) \\ \mathcal{M}^{I_0} &\models (\phi^+)^s && (\text{By } (*)) \\ \mathcal{M}^{I_0} &\models (\phi^s)^+ && (\text{Clearly } (\phi^+)^s \equiv (\phi^s)^+) \\ (\mathcal{M}^{I_0})^+ &\models \phi^s. && (\text{Lemma 49}) \end{aligned}$$

□

## 7 Second Consistency Result: Prioritizing Self-Truth

In this section we fulfil the second promise from the introduction. Throughout,  $\prec$  is an r.e. well-founded partial-order of  $\omega$ .

**Definition 54.** If  $i \in \omega$ , we say that a stratifier-set  $I$  is *above*  $i$  if  $\forall j \in \text{Indices}(I), i \prec j$ . We adopt the following convention: if  $I$  is above  $i$  then we will write  $I$  as  $I(i)$  in order to remind ourselves that  $I$  is above  $i$ .

**Definition 55.** (Compare Definition 14) Suppose  $\mathbf{T} = (T_i)_{i \in \omega}$  is an r.e. family of  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -theories and each  $T_i$  is  $i$ -unistratified. We say  $\mathbf{T}$  is *straticlosed-r.e.-generic* if for every straticlosed r.e. family  $\mathbf{U} \supseteq \mathbf{T}$ , every  $i \in \omega$ , and every computable stratifier-set  $I(i)$  above  $i$ ,  $\mathcal{M}_{\text{Str}(\mathbf{U})}^{I(i)} \models T_i$ .

**Lemma 56.** If the family  $\mathbf{T} = (T_i)_{i \in \omega}$  of  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -sets is r.e. and is a union of straticlosed-r.e.-generic families, then  $\mathbf{T}$  is straticlosed-r.e.-generic.

*Proof.* Straightforward. □

**Lemma 57.** (Compare Lemma 17) For any  $i, j \in \mathbb{N}$ , each of the following families is straticlosed-r.e.-generic.

1.  $[i\text{-Stratideduction}]_i$ .
2.  $[j\text{-Deduction}]_i$  (if  $j \not\prec i$ ).
3.  $[S]_i$  (if  $i \prec j$ ) where  $S$  is the following schema ( $\phi, \psi$  range over  $\mathcal{L}_{\text{PA}}(\omega)$ -formulas):

$$(\text{Weak } j\text{-Deduction}) \text{ ucl}(\mathbf{T}_j \models (\phi \rightarrow \psi) \rightarrow \mathbf{T}_j \models \phi \rightarrow \mathbf{T}_j \models (\psi \wedge (\phi \vee \neg \phi))).$$

4.  $[S]_i$  where  $S$  is: ( $i$ -Assigned Strativalidity) the schema  $\phi^s$  ( $\phi$  valid and  $i$ -stratified,  $s$  an assignment).
5.  $[i\text{-Assigned Strativalidity}]_i \cup [i\text{-Strativalidity}]_i$ .
6.  $[i\text{-Assigned Strativalidity}]_i \cup [i\text{-Validity}]_j$  (if  $j \neq i$ ).
7.  $[i\text{-Assigned Strativalidity}]_i \cup [i\text{-Strativalidity}]_i \cup [i\text{-Stratideduction}]_i \cup [i\text{-Introspection}]_j$  ( $j \neq i$ ).
8.  $[i\text{-Assigned Strativalidity}]_i \cup [i\text{-Strativalidity}]_i \cup [i\text{-Stratideduction}]_i \cup [S]_i$  where  $S$  is:

$$(i\text{-Stratrospection}) \text{ ucl}(\mathbf{T}_i^\alpha \models \phi \rightarrow \mathbf{T}_i^\beta \models \mathbf{T}_i^\alpha \models \phi) \text{ whenever this is } i\text{-stratified}.$$

9.  $[S]_i$  where  $S$  is the set of those axioms of Peano arithmetic for  $\mathcal{L}_{\text{PA}}(\mathcal{I})$  that are  $i$ -stratified.
10.  $[S]_i$  where  $S$  is any r.e. set of true arithmetic sentences.
11.  $[j\text{-SMT}]_i$  ( $j \neq i$ ).
12.  $[S]_i$ , where  $S$  is: ( $i$ -Strati-SMT)  $\text{ucl}(\exists e \forall x (\mathbf{T}_i^\alpha \models \phi \leftrightarrow x \in W_e))$  when this is  $i$ -stratified,  $e \notin \text{FV}(\phi)$ .
13.  $\mathbf{T} \cup [S]_i$  where  $\mathbf{T} = (T_k)_{k \in \omega}$  is straticlosed-r.e.-generic and  $S$  is the schema  $\mathbf{T}_i^\alpha \models \phi$  ( $\phi \in T_i$  such that this is  $i$ -stratified).

*Proof.* For unistratifiedness of 4–8, use Corollary 28. Unistratifiedness of the other families is clear. Recursive enumerability follows from the fact that  $\prec$  is r.e. In each case below, let  $\mathbf{U} = (U_k)_{k \in \omega}$  be a straticlosed r.e. family extending the family in question. For brevity, let  $\mathcal{M} = \mathcal{M}_{\text{Str}(\mathbf{U})}$ .

(1) Similar to part 1 of Lemma 17.

(2) Let  $I(i)$  be any computable stratifier-set above  $i$ , we must show  $\mathcal{M}^{I(i)} \models \text{ucl}(\mathbf{T}_j \models (\phi \rightarrow \psi) \rightarrow \mathbf{T}_j \models \phi \rightarrow \mathbf{T}_j \models \psi)$ .

Let  $s$  be an assignment and assume  $\mathcal{M}^{I(i)} \models \mathbf{T}_j \models (\phi \rightarrow \psi)[s]$  and  $\mathcal{M}^{I(i)} \models \mathbf{T}_j \models \phi[s]$ . Since  $I(i)$  is above  $i$  and  $j \not\prec i$ ,  $\mathcal{M}^{I(i)}$  and  $\mathcal{M}$  agree on  $\mathbf{T}_j \models$ , so  $\mathcal{M} \models \mathbf{T}_j \models (\phi \rightarrow \psi)[s]$  and  $\mathcal{M} \models \mathbf{T}_j \models \phi[s]$ . By definition of  $\mathcal{M} = \mathcal{M}_{\text{Str}(\mathbf{U})}$ ,  $U_j^- \models \phi^s \rightarrow \psi^s$  and  $U_j^- \models \phi^s$ , thus  $U_j^- \models \psi^s$ , so  $\mathcal{M} \models \mathbf{T}_j \models \psi[s]$  and thus so does  $\mathcal{M}^{I(i)}$ .

(3) Let  $I(i)$  be any computable stratifier-set above  $i$ , we must show  $\mathcal{M}^{I(i)} \models \text{ucl}(\mathbf{T}_j \models (\phi \rightarrow \psi) \rightarrow \mathbf{T}_j \models \phi \rightarrow \mathbf{T}_j \models (\psi \wedge (\phi \vee \neg \phi)))$ . Let  $s$  be an assignment and assume  $\mathcal{M}^{I(i)} \models \mathbf{T}_j \models (\phi \rightarrow \psi)[s]$  and  $\mathcal{M}^{I(i)} \models \mathbf{T}_j \models \phi[s]$ . If  $j \notin \text{Indices}(I(i))$ , then  $\mathcal{M}^{I(i)}$  and  $\mathcal{M}$  agree on  $\mathbf{T}_j \models$ , so reason as in (2) above. If not, we can write  $I(i) = I_0 \cup \{\bullet^+\}$  where  $\bullet^+$  is a computable  $j$ -stratifier, and Lemma 51 ensures that  $\mathcal{M}^{I(i)}$  and  $\mathcal{M}^+$  agree on  $\mathbf{T}_j \models$ . By definition of  $\mathcal{M}^+$ ,  $\mathcal{M} \models (\mathbf{T}_j \models (\phi \rightarrow \psi))^+[s]$  and  $\mathcal{M} \models (\mathbf{T}_j \models \phi)^+[s]$ . Let  $\alpha, \beta \in \epsilon_0 \cdot \omega$  be such that  $(\mathbf{T}_j \models (\phi \rightarrow \psi))^+ \equiv \mathbf{T}_j^\alpha \models (\phi^+ \rightarrow \psi^+)$  and  $(\mathbf{T}_j \models \phi)^+ \equiv \mathbf{T}_j^\beta \models \phi^+$ . Then  $\mathcal{M} \models \mathbf{T}_j^\alpha \models (\phi^+ \rightarrow \psi^+)[s]$  and  $\mathcal{M} \models \mathbf{T}_j^\beta \models \phi^+[s]$ . This means  $U_j \cap \alpha \models (\phi^+ \rightarrow \psi^+)^s$  and  $U_j \cap \beta \models (\phi^+)^s$ . Since  $\phi$  is a subformula of  $\phi \rightarrow \psi$ , it follows  $\beta \leq \alpha$ , thus  $U_j \cap \alpha \models (\psi^+)^s$ , and by tautology,  $U_j \cap \alpha \models (\psi^+ \wedge (\phi^+ \vee \neg \phi^+))^s$ . So  $\mathcal{M} \models \mathbf{T}_j^\alpha \models (\psi^+ \wedge (\phi^+ \vee \neg \phi^+))[s]$ . By Definition 45,

$$\mathbf{T}_j^\alpha \models (\psi^+ \wedge (\phi^+ \vee \neg \phi^+)) \equiv (\mathbf{T}_j \models (\psi \wedge (\phi \vee \neg \phi)))^+$$

(this is the reason for the  $\phi \vee \neg \phi$  clause) and finally  $\mathcal{M}^+ \models \mathbf{T}_j \models (\psi \wedge (\phi \vee \neg \phi))[s]$ .

(4) Similar to part 2 of Lemma 17.

(5) Similar to part 3 of Lemma 17.

(6) By (4),  $\mathcal{M}^{I(i)} \models i$ -Assigned Stratificability for every computable stratifier-set  $I(i)$  above  $i$ . Let  $J(j)$  be a computable stratifier-set above  $j$ , we must show  $\mathcal{M}^{J(j)} \models i$ -Validity. Let  $\phi$  be a valid  $\mathcal{L}_{\text{PA}}(\omega)$ -formula,  $s$  an assignment.

**Case 1:**  $i \notin \text{Indices}(J(j))$ . Then  $\mathcal{M}^{J(j)}$  and  $\mathcal{M}$  agree on  $\mathbf{T}_i \models$ . Let  $\bullet^+$  be an  $i$ -stratifier. Since  $\phi$  is valid, so is  $\phi^+$  (by Lemma 50), so  $(\phi^+)^s \in U_i$  (since  $[i$ -Assigned Stratificability] $_i$  is part of line 6). Clearly  $((\phi^+)^s)^- \equiv \phi^s$ , so  $\phi^s \in U_i^-$ , thus  $\mathcal{M} \models \mathbf{T}_i \models \phi[s]$ , and so does  $\mathcal{M}^{J(j)}$ .

**Case 2:**  $i \in \text{Indices}(J(j))$ . Thus  $j \prec i$  and we can write  $J(j) = I_0 \cup \{\bullet^+\}$  for some computable  $i$ -stratifier  $\bullet^+$ . By Lemma 51,  $\mathcal{M}^{J(j)}$  and  $\mathcal{M}^+$  agree on  $\mathbf{T}_i \models$ . Let  $\alpha \in \epsilon_0 \cdot \omega$  be such that  $(\mathbf{T}_i \models \phi)^+ \equiv \mathbf{T}_i^\alpha \models \phi^+$ . As in Case 1,  $(\phi^+)^s$  is an instance of  $i$ -Assigned Stratificability, so  $(\phi^+)^s \in U_i$  (since  $[i$ -Assigned Stratificability] $_i$  is part of line 6). In fact by choice of  $\alpha$ ,  $(\phi^+)^s \in U_i \cap \alpha$ , so  $\mathcal{M} \models \mathbf{T}_i^\alpha \models \phi^+[s]$ , that is,  $\mathcal{M} \models (\mathbf{T}_i \models \phi)^+[s]$ . By Lemma 49,  $\mathcal{M}^+ \models \mathbf{T}_i \models \phi[s]$ . Since  $\mathcal{M}^{J(j)}$  and  $\mathcal{M}^+$  agree on  $\mathbf{T}_i \models$ ,  $\mathcal{M}^{J(j)} \models \mathbf{T}_i \models \phi[s]$ .

(7–8) Similar to part 4 of Lemma 17. For 8, use part 3 of Definition 41.

(9) Similar to Lemma 10.

(10) Trivial.

(11–12) Similar to part 7 of Lemma 17. Use the fact that the stratifier-sets in Definition 55 are computable.

(13) Similar to part 8 of Lemma 17; use part 3 of Definition 41.  $\square$

**Definition 58.** If  $\mathbf{T}^0 = (T_i^0)_{i \in \omega}$  where each  $T_i^0$  is an  $\mathcal{L}_{\text{PA}}(\omega)$ -theory, we say  $\mathbf{T}^0$  is *stratifiable-r.e.-generic* if there is some straticlosed-r.e.-generic family  $\mathbf{T} = (T_i)_{i \in \omega}$  of  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -theories such that each  $T_i^- = T_i^0$ .

The following theorem will satisfy the second promise from the introduction: it will exhibit true theories  $(T_i)_{i \in \omega}$  such that  $T_i$  expresses a Gödel number of  $T_j$  ( $j \prec i$ ) and the truth of  $T_j$  ( $j \preceq i$ ). These theories can further be taken so that  $T_i$  expresses the fact that  $T_j$  has some Gödel number (all  $i, j$ ), by Lemma 57 parts 11–12.

**Theorem 59.** Let  $\mathbf{T}^0 = (T_i^0)_{i \in \omega}$  be any stratifiable-r.e.-generic family of  $\mathcal{L}_{\text{PA}}(\omega)$ -theories. For every  $i \in \omega$  and  $n \in \mathbb{N}$ , let  $T_i(n)$  be the smallest  $\mathbf{T}_i \models$ -closed  $\mathcal{L}_{\text{PA}}(\omega)$ -theory containing the following axioms.

1. The axioms contained in  $T_i^0$ .
2. Assigned Validity,  $i$ -Validity and  $i$ -Deduction.
3.  $\text{ucl}(\mathbf{T}_j \models \phi \rightarrow \phi)$  whenever  $j \preceq i$ .
4.  $\forall x(\mathbf{T}_j \models \phi \leftrightarrow \langle \overline{\phi}, \bar{j}, x \rangle \in W_{\bar{n}})$  whenever  $j \prec i$ ,  $\text{FV}(\phi) \subseteq \{x\}$ .

Let each  $\mathbf{T}(n) = (T_i(n))_{i \in \omega}$ . There is some  $n \in \mathbb{N}$  such that  $\mathbf{T}(n)$  is true.

*Proof.* By the  $S$ - $m$ - $n$  Theorem, there is a total computable  $f : \mathbb{N} \rightarrow \mathbb{N}$  such that  $\forall n \in \mathbb{N}$ ,

$$W_{f(n)} = \{ \langle \ulcorner \phi \urcorner, j, m \rangle \in \mathbb{N} : \phi \text{ is an } \mathcal{L}_{\text{PA}}(\omega)\text{-formula, } \text{FV}(\phi) \subseteq \{x\}, \text{ and } T_j(n) \models \phi(x|\overline{m}) \}.$$

By the Recursion Theorem, there is an  $n \in \mathbb{N}$  such that  $W_n = W_{f(n)}$ . We will show  $\mathbf{T}(n)$  is true. For the rest of the proof, we write  $\mathbf{T}$  for  $\mathbf{T}(n)$ ,  $T_i$  for  $T_i(n)$ .

The structure of the proof is as follows.

- First, we will aim at the following preliminary result. For a certain carefully-chosen family  $\mathbf{U} = (U_i)_{i \in \omega}$  of  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -theories (with each  $U_i^- = T_i$ ) and the  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -structure  $\mathcal{M} = \mathcal{M}_{\text{Str}(\mathbf{U})}$ , we will prove that  $\forall i \in \omega, \mathcal{M} \models U_i \cup T_i$ .
- To prove the preliminary result, in order to deal with the difficulty mentioned at the beginning of Section 6, we will prove more than necessary, to obtain a strong  $\prec$ -induction hypothesis. Namely, we will prove, by  $\prec$ -induction, that  $\forall i \in \omega$ , for every stratifier-set  $I(i)$  above  $i$ ,  $\mathcal{M}^{I(i)} \models U_i \cup T_i$ .
  - In order to prove  $\mathcal{M}^{I(i)} \models U_i$ , we will use induction on  $\alpha$  to show that  $\mathcal{M}^{I(i)} \models U_i \cap \alpha$  for all  $\alpha \in \epsilon_0 \cdot \omega$ .
  - Part of proving  $\mathcal{M}^{I(i)} \models U_i \cap \alpha$  will be proving  $\mathcal{M}^{I(i)} \models \text{ucl}(\mathbf{T}_i^{\alpha} \models \phi \rightarrow \phi)$  whenever this is  $i$ -stratified,  $\alpha_0 < \alpha$ . This is where we will use the  $\alpha$ -induction hypothesis.
  - Part of proving  $\mathcal{M}^{I(i)} \models U_i \cap \alpha$  will be proving  $\mathcal{M}^{I(i)} \models \text{ucl}(\mathbf{T}_j \models \phi \rightarrow \phi^+)$  whenever  $j \prec i$ ,  $\phi$  is an  $\mathcal{L}_{\text{PA}}(\omega)$ -formula, and  $\bullet^+$  is an  $i$ -stratifier. This is where we will take advantage of our strong  $\prec$ -induction hypothesis.
- Once we've established  $\mathcal{M}^{I(i)} \models U_i$ , we will essentially be able to conclude  $\mathcal{M}^{I(i)} \models T_i$  using the Stratification Theorem (Theorem 44).
- Having established that  $\forall i \in \omega, \mathcal{M} \models U_i \cup T_i$ , we will use that to prove that  $\mathcal{M}_{\mathbf{T}} \models \mathbf{T}$ , i.e., that  $\mathbf{T}$  is true.

Since  $\mathbf{T}^0$  is stratifiable-r.e.-generic, there is a straticlosed-r.e.-generic family  $\mathbf{V} = (V_i)_{i \in \omega}$  of  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -theories such that each  $V_i^- = T_i^0$ . For every  $i \in \mathbb{N}$ , let  $U_i$  be the smallest  $i$ -stratified  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -theory such that the following hold.

1.  $U_i$  contains  $V_i$ .
2.  $U_i$  contains  $i$ -Assigned Strativalidity,  $i$ -Strativalidity,  $i$ -Stratideduction and  $i$ -Collapse.
3.  $U_i$  contains  $\text{ucl}(\mathbf{T}_i^{\alpha} \models \phi \rightarrow \phi)$  whenever  $\mathbf{T}_i^{\alpha} \models \phi$  is  $i$ -stratified.
4.  $U_i$  contains  $\text{ucl}(\mathbf{T}_j \models \phi \rightarrow \phi^+)$  for every  $\mathcal{L}_{\text{PA}}(\omega)$ -formula  $\phi$ ,  $j \prec i$ , and  $i$ -stratifier  $\bullet^+$ .
5.  $U_i$  contains  $\forall x(\mathbf{T}_j \models \phi \leftrightarrow \langle \ulcorner \phi \urcorner, \vec{j}, x \rangle \in W_{\vec{n}})$  whenever  $j \prec i$ ,  $\text{FV}(\phi) \subseteq \{x\}$  and  $\phi$  is an  $\mathcal{L}_{\text{PA}}(\omega)$ -formula.
6. Whenever  $\phi \in U_i$  and  $\mathbf{T}_i^{\alpha} \models \phi$  is  $i$ -stratified,  $\mathbf{T}_i^{\alpha} \models \phi \in U_i$ .

Let  $\mathbf{U} = (U_i)_{i \in \omega}$ . Observe that  $\mathbf{U}$  is straticlosed and r.e. (to see  $U_i$  is  $i$ -unistratified, use Lemma 46; to see  $\mathbf{U}$  is r.e., use Theorem 20 part 1);  $\mathbf{U} \supseteq \mathbf{V}$ ; and for each  $i \in \omega$ ,  $U_i^- = T_i$ .

Let  $\mathcal{M} = \mathcal{M}_{\text{Str}(\mathbf{U})}$ . Recall that  $\text{Str}(\mathbf{U})$  is the  $\mathcal{L}_{\text{PA}}(\mathcal{I})$  family  $(S_i)_{i \in \mathcal{I}}$  where  $\forall i \in \omega$  and  $\alpha \in \epsilon_0 \cdot \omega$ ,  $S_i = U_i^- = T_i$  and  $S_{(\alpha, i)} = U_i \cap \alpha$ . For the reader's convenience, here is how (by definition)  $\mathcal{M}$  interprets  $\mathbf{T}_i \models$  and  $\mathbf{T}_i^{\alpha} \models$  for all  $i \in \omega$ ,  $\alpha \in \epsilon_0 \cdot \omega$ :

$$\begin{aligned} \mathcal{M} \models \mathbf{T}_i \models \phi[s] &\text{ iff } T_i \models \phi^s, \\ \mathcal{M} \models \mathbf{T}_i^{\alpha} \models \phi[s] &\text{ iff } U_i \cap \alpha \models \phi^s. \end{aligned}$$

We would like to prove the following intermediate result:  $\forall i \in \omega, \mathcal{M} \models U_i \cup T_i$ . For the sake of a stronger induction hypothesis, we will prove that  $\forall i \in \omega$ , for every computable stratifier-set  $I(i)$  above  $i$ ,  $\mathcal{M}^{I(i)} \models U_i \cup T_i$ . This is more than enough because  $\mathcal{M}^{I(i)} = \mathcal{M}$  when  $I(i) = \emptyset$ .

Fix  $i \in \omega$ . By  $\prec$ -induction, we have the following:



(\*) For every  $j \prec i$ , for every computable stratifier-set  $J(j)$  above  $j$ ,  $\mathcal{M}^{J(j)} \models U_j \cup T_j$ .

Let  $I(i)$  be any computable stratifier-set above  $i$ . We must show  $\mathcal{M}^{I(i)} \models U_i \cup T_i$ .

**Claim 1:**  $\forall \alpha \in \epsilon_0 \cdot \omega$ ,  $\mathcal{M}^{I(i)} \models U_i \cap \alpha$ .

By induction on  $\alpha$ . Let  $\sigma \in U_i \cap \alpha$ .

**Case 1:**  $\sigma \in V_i$ . Then  $\mathcal{M}^{I(i)} \models \sigma$  because  $\mathbf{V}$  is straticlosed-r.e.-generic and  $\mathbf{U} \supseteq \mathbf{V}$  is straticlosed and r.e.

**Case 2:**  $\sigma$  is an instance of  $i$ -Assigned Strativalidity,  $i$ -Strativalidity, or  $i$ -Stratideduction. Then  $\mathcal{M}^{I(i)} \models \sigma$  by Lemma 57.

**Case 3:**  $\sigma$  is  $\text{ucl}(\mathbf{T}_i^{\alpha_0} \models \phi \rightarrow \phi)$  for some  $i$ -stratified  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -formula  $\phi$  such that  $\mathbf{T}_i^{\alpha_0} \models \phi$  is  $i$ -stratified. Since  $\sigma \in U_i \cap \alpha$ , this forces  $\alpha_0 < \alpha$ . Let  $s$  be an assignment and assume  $\mathcal{M}^{I(i)} \models \mathbf{T}_i^{\alpha_0} \models \phi[s]$ , then:

$$\begin{aligned} \mathcal{M}^{I(i)} &\models \mathbf{T}_i^{\alpha_0} \models \phi[s] && \text{(Assumption)} \\ \mathcal{M} &\models \mathbf{T}_i^{\alpha_0} \models \phi[s] && (\mathcal{M} \text{ and } \mathcal{M}^{I(i)} \text{ agree on } \mathbf{T}_i^{\alpha_0} \models \text{ by Def. 48}) \\ U_i \cap \alpha_0 &\models \phi^s && \text{(Definition of } \mathcal{M}) \\ \mathcal{M}^{I(i)} &\models \phi^s && \text{(By } \alpha\text{-induction, } \mathcal{M}^{I(i)} \models U_i \cap \alpha_0) \\ \mathcal{M}^{I(i)} &\models \phi[s]. && \text{(Lemma 53)} \end{aligned}$$

**Case 4:**  $\sigma$  is  $\text{ucl}(\mathbf{T}_j \models \phi \rightarrow \phi^+)$  for some  $\mathcal{L}_{\text{PA}}(\omega)$ -formula  $\phi$ ,  $j \prec i$ , and  $i$ -stratifier  $\bullet^+$ . By Lemma 46 we may assume  $\bullet^+$  is computable. Let  $J(j)$  be the computable stratifier-set  $J(j) = I(i) \cup \{\bullet^+\}$ , which is above  $j$  since  $I(i)$  is above  $i$  and  $j \prec i$ . Let  $s$  be an assignment and assume  $\mathcal{M}^{I(i)} \models \mathbf{T}_j \models \phi[s]$ , then:

$$\begin{aligned} \mathcal{M}^{I(i)} &\models \mathbf{T}_j \models \phi[s] && \text{(Assumption)} \\ \mathcal{M} &\models \mathbf{T}_j \models \phi[s] && \text{(Since } j \prec i \text{ and } I(i) \text{ is above } i, \mathcal{M}^{I(i)} \text{ and } \mathcal{M} \text{ agree on } \mathbf{T}_j \models) \\ T_j &\models \phi^s && \text{(Definition of } \mathcal{M}) \\ \mathcal{M}^{J(j)} &\models \phi^s && \text{(Since } \mathcal{M}^{J(j)} \models T_j \text{ by } (*) \\ (\mathcal{M}^{I(i)})^+ &\models \phi^s && \text{(Lemma 51)} \\ \mathcal{M}^{I(i)} &\models (\phi^s)^+ && \text{(Lemma 49)} \\ \mathcal{M}^{I(i)} &\models (\phi^+)^s && \text{(Clearly } (\phi^s)^+ \equiv (\phi^+)^s) \\ \mathcal{M}^{I(i)} &\models \phi^+[s]. && \text{(Lemma 53)} \end{aligned}$$

**Case 5:**  $\sigma$  is  $\forall x(\mathbf{T}_j \models \phi \leftrightarrow \langle \overline{\Gamma \phi^\top}, \bar{j}, x \rangle \in W_{\bar{n}})$  for some  $\mathcal{L}_{\text{PA}}(\omega)$ -formula  $\phi$  with  $\text{FV}(\phi) \subseteq \{x\}$  and  $j \prec i$ . Let  $s$  be any assignment, say  $s(x) = m$ . The following biconditionals are equivalent:

$$\begin{aligned} \mathcal{M}^{I(i)} &\models \mathbf{T}_j \models \phi \leftrightarrow \langle \overline{\Gamma \phi^\top}, \bar{j}, x \rangle \in W_{\bar{n}}[s] \\ \mathcal{M} &\models \mathbf{T}_j \models \phi \leftrightarrow \langle \overline{\Gamma \phi^\top}, \bar{j}, x \rangle \in W_{\bar{n}}[s] && (\mathcal{M}^{I(i)} \text{ and } \mathcal{M} \text{ agree on the symbols in question)} \\ \mathcal{M} &\models \mathbf{T}_j \models \phi[s] \text{ iff } \mathcal{M} \models \langle \overline{\Gamma \phi^\top}, \bar{j}, \bar{m} \rangle \in W_{\bar{n}} && \text{(Lemma 53)} \\ \mathcal{M} &\models \mathbf{T}_j \models \phi[s] \text{ iff } \langle \overline{\Gamma \phi^\top}, j, m \rangle \in W_n && (\mathcal{M} \text{ has standard first-order part)} \\ T_j &\models \phi^s \text{ iff } \langle \overline{\Gamma \phi^\top}, j, m \rangle \in W_n && \text{(Definition of } \mathcal{M}) \\ T_j &\models \phi(x|\bar{m}) \text{ iff } \langle \overline{\Gamma \phi^\top}, j, m \rangle \in W_n. && \text{(Since } \text{FV}(\phi) \subseteq \{x\}) \end{aligned}$$

The latter is true by definition of  $n$ .

**Case 6:**  $\sigma$  is an instance  $\mathbf{T}_i^\beta \models \phi \leftrightarrow \mathbf{T}_i^\gamma \models \phi$  of  $i$ -Collapse (so  $\beta \leq_1 \gamma$  and  $\mathbf{T}_i^\beta \models \phi \leftrightarrow \mathbf{T}_i^\gamma \models \phi$  is  $i$ -stratified). Let  $s$  be an assignment, since  $\mathcal{M}^{I(i)}$  and  $\mathcal{M}$  agree on  $\mathbf{T}_i^\beta \models$  and  $\mathbf{T}_i^\gamma \models$ , we need only show  $\mathcal{M} \models \mathbf{T}_i^\beta \models \phi \leftrightarrow \mathbf{T}_i^\gamma \models \phi[s]$ . In other words we must show  $U_i \cap \beta \models \phi^s$  if and only if  $U_i \cap \gamma \models \phi^s$ . This is by Theorem 32.

**Case 7:**  $\sigma$  is  $\mathbf{T}_i^{\alpha_0} \models \phi$  for some  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -formula  $\phi$  such that  $\mathbf{T}_i^{\alpha_0} \models \phi$  is  $i$ -stratified and  $\phi \in U_i$ . Since  $\mathbf{T}_i^{\alpha_0} \models \phi$  is  $i$ -stratified,  $\text{On}(\phi) \subseteq \alpha_0$ , so  $\phi \in U_i \cap \alpha_0$ . Thus  $\mathcal{M} \models \mathbf{T}_i^{\alpha_0} \models \phi$ , so  $\mathcal{M}^{I(i)} \models \mathbf{T}_i^{\alpha_0} \models \phi$  since  $\mathcal{M}^{I(i)}$  and  $\mathcal{M}$  agree on  $\mathbf{T}_i^{\alpha_0} \models$ .

Cases 1–7 establish  $\mathcal{M}^{I(i)} \models U_i \cap \alpha$ . By arbitrariness of  $\alpha$ , Claim 1 is proved.

**Claim 2:** For any assignment  $s$  and any very  $i$ -stratified  $\mathcal{L}_{\text{PA}}(\mathcal{I})$ -formula  $\phi$ ,  $\mathcal{M}^{I(i)} \models \phi[s]$  if and only if  $\mathcal{M}^{I(i)} \models \phi^-[s]$ .

By Theorem 44, for all such  $s$  and  $\phi$ ,  $\mathcal{M} \models \phi[s]$  if and only if  $\mathcal{M} \models \phi^-[s]$ . The claim now follows from Lemma 52 ( $i \notin \text{Indices}(I(i))$  because  $I(i)$  is above  $i$ ).

**Claim 3:**  $\mathcal{M}^{I(i)} \models T_i$ .

For any  $\sigma \in T_i$ , there is some  $\tau \in U_i$  such that  $\tau^- \equiv \sigma$ ; since  $U_i$  is  $i$ -unistratified, we may take  $\tau$  to be very  $i$ -stratified (Lemma 35). By Claim 1,  $\mathcal{M}^{I(i)} \models U_i$ , so  $\mathcal{M}^{I(i)} \models \tau$ . By Claim 2,  $\mathcal{M}^{I(i)} \models \sigma$ .

For each  $i \in \omega$ , letting  $I(i) = \emptyset$ , Claims 1–3 show that  $\mathcal{M} \models U_i \cup T_i$ . It follows that  $\mathcal{M} \models \mathbf{T}$ . Now, for every  $i \in \omega$ ,  $\mathcal{M}_{\mathbf{T}}$  interprets  $\mathbf{T}_i$  as follows:

$$\mathcal{M}_{\mathbf{T}} \models \mathbf{T}_i \models \phi[s] \text{ iff } T_i \models \phi^s.$$

This is exactly the same way that  $\mathcal{M}$  interprets  $\mathbf{T}_i$ . It follows that  $\mathcal{M}$  and  $\mathcal{M}_{\mathbf{T}}$  agree on  $\mathcal{L}_{\text{PA}}(\omega)$ -formulas. Thus, since  $\mathcal{M} \models \mathbf{T}$ ,  $\mathcal{M}_{\mathbf{T}} \models \mathbf{T}$ , i.e.,  $\mathbf{T}$  is true.  $\square$

## 8 Well-Foundation and Ill-Foundation

The following is a variation on Kleene's  $\mathcal{O}$ .

**Definition 60.** Simultaneously define  $\mathcal{O} \subseteq \mathbb{N}$  and  $|\bullet| : \mathcal{O} \rightarrow \text{Ord}$  so that  $\mathcal{O} \subseteq \mathbb{N}$  is the smallest set such that:

1.  $0 \in \mathcal{O}$  (it represents the ordinal  $|0| = 0$ ).
2.  $\forall n \in \mathcal{O}, 2^n \in \mathcal{O}$  (it represents the ordinal  $|2^n| = |n| + 1$ ).
3. If  $\varphi_e$  (the  $e$ th partial recursive function) is total and  $\text{range}(\varphi_e) \subseteq \mathcal{O}$ , then  $3 \cdot 5^e \in \mathcal{O}$  (it represents the ordinal  $|3 \cdot 5^e| = \sup\{|\varphi_e(0)|, |\varphi_e(1)|, \dots\}$ ).

To avoid technical complications, we have differed from the usual Kleene's  $\mathcal{O}$  in the following way: in the usual definition, in order for  $3 \cdot 5^e$  to lie in  $\mathcal{O}$ , it is also required that  $|\varphi_e(0)| < |\varphi_e(1)| < \dots$ .

**Definition 61.**  $\mathcal{L}_{\text{PA}}^{\mathcal{O}}$  is the language of Peano arithmetic extended by a unary predicate  $\mathcal{O}$ . The following notions are defined by analogy with Section 2:

1. For any assignment  $s$  and  $\mathcal{L}_{\text{PA}}^{\mathcal{O}}(I)$ -formula  $\phi$  with  $\text{FV}(\phi) = \{x_1, \dots, x_n\}$ ,  $\phi^s \equiv \phi(x_1 | \overline{s(x_1)}) \cdots (x_n | \overline{s(x_n)})$ .
2. If  $\mathbf{T} = (T_i)_{i \in I}$  is an  $I$ -indexed family of  $\mathcal{L}_{\text{PA}}^{\mathcal{O}}(I)$ -theories, the *intended structure* for  $\mathbf{T}$  is the  $\mathcal{L}_{\text{PA}}^{\mathcal{O}}(I)$ -structure  $\mathcal{M}_{\mathbf{T}}$  with universe  $\mathbb{N}$ , interpreting symbols of PA as usual and interpreting  $\mathcal{O}$  as  $\mathcal{O}$ , and interpreting  $\mathbf{T}_i$  ( $i \in I$ ) as in Definition 7. For any  $\mathcal{L}_{\text{PA}}^{\mathcal{O}}(I)$ -structure  $\mathcal{N}$ , we write  $\mathcal{N} \models \mathbf{T}$  if  $\forall i \in I$ ,  $\mathcal{N} \models T_i$ . We say  $\mathbf{T}$  is *true* if  $\mathcal{M}_{\mathbf{T}} \models \mathbf{T}$ .

**Definition 62.** If  $I$  is an index set and  $\mathbf{T} = (T_i)_{i \in I}$  is a family of  $\mathcal{L}_{\text{PA}}^{\mathcal{O}}(I)$ -theories, then for any  $i \in I$  such that  $\mathcal{M}_{\mathbf{T}} \models T_i$ , we define the ordinal  $\|T_i\| = \sup\{|m| + 1 : T_i \models \mathcal{O}(\overline{m})\}$ .

The above definition makes sense: since  $\mathcal{M}_{\mathbf{T}} \models T_i$  and  $\mathcal{O}^{\mathcal{M}_{\mathbf{T}}} = \mathcal{O}$ , the supremands are defined.

**Definition 63.** The *basic axioms* of  $\mathcal{O}$  are the following  $\mathcal{L}_{\text{PA}}^{\mathcal{O}}$ -axioms.

1.  $\mathcal{O}(0)$ .
2.  $\mathcal{O}(\overline{n}) \rightarrow \mathcal{O}(\overline{2^n})$ , for every  $n \in \mathbb{N}$ .
3.  $\forall x(\varphi_{\overline{n}}(x) \downarrow \ \& \ \mathcal{O}(\varphi_{\overline{n}}(x))) \rightarrow \mathcal{O}(\overline{3 \cdot 5^n})$ , for every  $n \in \mathbb{N}$ .

We have written the last two lines using infinite schemata to strengthen the following result.

**Theorem 64.** Let  $I$  be an index set,  $\prec$  a binary relation on  $I$ . Suppose  $\mathbf{T} = (T_i)_{i \in I}$  is a family of  $\mathcal{L}_{\text{PA}}^{\mathcal{O}}(I)$ -theories with the following properties:

1.  $\forall i \in I$ ,  $T_i$  contains the axioms of Peano arithmetic.
2.  $\forall i \in I$ ,  $T_i$  contains the basic axioms of  $\mathcal{O}$ .
3.  $\forall i \in I$ ,  $\forall j \prec i$ ,  $\exists n \in \mathbb{N}$  such that  $T_i \models \forall x(\mathbf{T}_j \models \mathcal{O}(x) \leftrightarrow x \in W_{\bar{n}})$ .
4.  $\forall i \in I$ ,  $\forall j \prec i$ ,  $T_i \models \forall x(\mathbf{T}_j \models \mathcal{O}(x) \rightarrow \mathcal{O}(x))$ .

If  $\mathcal{M}_{\mathbf{T}} \models T_i \cup T_j$  (in particular if  $\mathbf{T}$  is true) and  $j \prec i$ , then  $\|T_j\| < \|T_i\|$ .

*Proof.* Assume  $\mathcal{M}_{\mathbf{T}} \models T_i \cup T_j$  and  $j \prec i$ . By hypothesis there is some  $n \in \mathbb{N}$  such that  $T_i \models \forall x(\mathbf{T}_j \models \mathcal{O}(x) \leftrightarrow x \in W_{\bar{n}})$  and  $T_i \models \forall x(\mathbf{T}_j \models \mathcal{O}(x) \rightarrow \mathcal{O}(x))$ . From these,  $T_i \models \forall x(x \in W_{\bar{n}} \rightarrow \mathcal{O}(x))$ .

Since  $\mathcal{M}_{\mathbf{T}} \models T_i$ , in particular  $\mathcal{M}_{\mathbf{T}} \models \forall x(\mathbf{T}_j \models \mathcal{O}(x) \leftrightarrow x \in W_{\bar{n}})$ . This means  $W_n = \{m \in \mathbb{N} : T_j \models \mathcal{O}(\bar{m})\}$ . Since  $T_j$  includes the axiom  $\mathcal{O}(0)$ ,  $W_n \neq \emptyset$ .

Since  $W_n \neq \emptyset$ , by computability theory there is some  $k \in \mathbb{N}$  such that

$$\text{PA} \models (\text{domain}(\varphi_{\bar{k}}) = \mathbb{N}) \wedge (\text{range}(\varphi_{\bar{k}}) = W_{\bar{n}}).$$

Since  $T_i$  includes PA,  $T_i$  also implies as much. Combined with  $T_i \models \forall x(x \in W_{\bar{n}} \rightarrow \mathcal{O}(x))$ , it follows that  $T_i \models \forall x(\varphi_{\bar{k}}(x) \downarrow \ \& \ \mathcal{O}(\varphi_{\bar{k}}(x)))$ . Since  $T_i$  contains the basic axiom  $\forall x(\varphi_{\bar{k}}(x) \downarrow \ \& \ \mathcal{O}(\varphi_{\bar{k}}(x))) \rightarrow \mathcal{O}(\overline{3 \cdot 5^k})$ ,  $T_i \models \mathcal{O}(\overline{3 \cdot 5^k})$ .

To finish the proof, calculate

$$\begin{aligned} \|T_j\| &= \sup\{|m| + 1 : T_j \models \mathcal{O}(\bar{m})\} \\ &= \sup\{|m| : T_j \models \mathcal{O}(\bar{m})\} && (\text{Since } T_j \text{ contains } \mathcal{O}(\bar{n}) \rightarrow \mathcal{O}(\overline{2^n}) \text{ for all } n \in \mathbb{N}) \\ &= \sup\{|m| : m \in W_n\} && (\text{Since } W_n = \{m \in \mathbb{N} : T_j \models \mathcal{O}(\bar{m})\}) \\ &= \sup\{|\varphi_k(0)|, |\varphi_k(1)|, \dots\} && (\text{By choice of } k) \\ &= |3 \cdot 5^k| && (\text{Definition 60}) \\ &< \sup\{|m| + 1 : T_i \models \mathcal{O}(\bar{m})\} && (\text{Since } T_i \models \mathcal{O}(\overline{3 \cdot 5^k})) \\ &= \|T_i\|. \end{aligned}$$

□

**Corollary 65.** (Well-Foundedness of True Self-Referential Theories) Let  $I$ ,  $\mathbf{T}$ ,  $\prec$  be as in Theorem 64. If  $\mathbf{T}$  is true then  $\prec$  is well founded, by which we mean there is no infinite descending sequence  $i_0 \succ i_1 \succ \dots$ .

In particular Corollary 65 says that if  $I$ ,  $\mathbf{T}$ ,  $\prec$  are as in Theorem 64 and  $\mathbf{T}$  is true then  $\prec$  is strict: there is no  $i$  with  $i \prec i$ . This gives a new form (under the additional new assumption of containing/knowing basic rudiments of computable ordinals) of the Lucas–Penrose–Reinhardt argument that a truthful theory (or machine) cannot state (or know) its own truth and its own Gödel number.

We could remove Peano arithmetic from Theorem 64 if we further departed from Kleene and changed line 3 of Definition 60 to read:

3. If  $W_e \subseteq \mathcal{O}$ , then  $3 \cdot 5^e \in \mathcal{O}$  (and  $|3 \cdot 5^e| = \sup\{|n| : n \in W_e\}$ , or  $|3 \cdot 5^e| = 0$  if  $W_e = \emptyset$ )

(and altered Definition 63 accordingly). The previous paragraph would still stand, in fact giving a version of the Lucas–Penrose–Reinhardt argument in which the theory (machine) is not required to contain (know) arithmetic.

We close the paper by showing that Corollary 65 fails without  $\mathcal{O}$ . Let WF be the set of all r.e. well-founded partial orders on  $\omega$  and let Tr be the set of all true  $\mathcal{L}_{\text{PA}}$ -sentences. It is well-known that WF is computability theoretically  $\Pi_1^1$ -complete and Tr is  $\Delta_1^1$ , so WF cannot be defined in  $\mathcal{L}_{\text{PA}} \cup \{\text{Tr}\}$ .

**Theorem 66.** (Ill-Foundedness of True Self-Referential Theories)

1. For any closed-r.e.-generic  $\mathbf{T}^0 = (T_i^0)_{i \in \omega}$ , there is an r.e., ill-founded partial order  $\prec$  on  $\omega$  and an  $n \in \mathbb{N}$  such that  $\mathbf{T}(n)$  is true, where  $\mathbf{T}(n)$  is as in Theorem 18.

2. For any stratifiable-r.e.-generic  $\mathbf{T}^0 = (T_i^0)_{i \in \omega}$ , there is an r.e., ill-founded partial order  $\prec$  on  $\omega$  and an  $n \in \mathbb{N}$  such that  $\mathbf{T}(n)$  is true, where  $\mathbf{T}(n)$  is as in Theorem 59.

*Proof.* We prove (1), (2) is similar. Assume  $\neg(1)$ . If  $\prec$  is any r.e. partial order on  $\omega$ , then, combining  $\neg(1)$  with Theorem 18,  $\prec$  is well founded if and only if the conclusion of Theorem 18 holds for  $\prec$ . Thus it is possible to define WF in  $\mathcal{L}_{\text{PA}} \cup \{\text{Tr}\}$ . Absurd.  $\square$

## References

- [1] Alexander, S. (2013). The Theory of Several Knowing Machines. Doctoral dissertation, the Ohio State University.
- [2] Alexander, S. (2014). A machine that knows its own code. *Studia Logica*, **102**, 567–576.
- [3] Benacerraf, P. (1967). God, the Devil, and Gödel. *The Monist*, **51**, 9–32.
- [4] Carlson, T.J. (1999). Ordinal arithmetic and  $\Sigma_1$ -elementarity. *Archive for Mathematical Logic*, **38**, 449–460.
- [5] Carlson, T.J. (2000). Knowledge, machines, and the consistency of Reinhardt’s strong mechanistic thesis. *Annals of Pure and Applied Logic*, **105**, 51–82.
- [6] Carlson, T.J. (2001). Elementary patterns of resemblance. *Annals of Pure and Applied Logic*, **108**, 19–77.
- [7] Lucas, J.R. (1961). Minds, machines, and Gödel. *Philosophy*, **36**, 112–127.
- [8] Penrose, R. (1989). *The Emperor’s New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford University Press.
- [9] Putnam, H. (2006). After Gödel. *Logic Journal of the IGPL*, **14**, 745–754.
- [10] Reinhardt, W. (1985). Absolute versions of incompleteness theorems. *Nous*, **19**, 317–346.
- [11] Shapiro, S. (1985). Epistemic and Intuitionistic Arithmetic. In: S. Shapiro (ed.), *Intensional Mathematics* (North-Holland, Amsterdam), pp. 11–46.