The Strength of Compositional Truth

MSc Thesis *(Afstudeerscriptie)*

written by

Fangjing Xiong

under the supervision of **Prof Dr Albert Visser** and **Prof Dr Dick de Jongh**, and submitted to the Examinations Board in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam*.

Date of the public defense:	**Members of the Thesis Committee:**
July 7th, 2025	Dr Balder ten Cate (Chair)
	Prof Dr Albert Visser (Supervisor)
	Prof Dr Dick de Jongh (Supervisor)
	Prof Dr Rineke Verbrugge
	Dr Thomas Schindler

Institute for Logic, Language and Computation

There is nothing but quotations left for us.

Jorge Luis Borges, *The Book of Sand and Shakespeare's Memory*

Abstract

A compositional theory of truth with the induction principle extended to sentences containing the truth predicate is not conservative over the base theory. It is unknown whether compositionality or extended induction contributes more to the nonconservativity result. This thesis follows Heck [Hec18]'s clarification by studying the strength of compositional truth, in particular, we study the strength of compositional truth in an induction-free environment.

We establish two conservation results concerning compositional truth without extended induction (whose axioms are denoted as CT) in an induction-free fragment of Peano Arithmetic PA^-. We denote PA^- with the compositional truth axioms $\mathsf{CT}[\mathsf{PA}^-]$. First, generalizing a model-theoretic proof of Enayat and Visser [EV15] for the conservativity of $\mathsf{CT}[\mathsf{PA}]$, we show that $\mathsf{CT}[\mathsf{PA}^-]$ is syntactically conservative over PA^-. Second, by generalizing Kaye [Kay91]'s modification of Lachlan [Lac81]'s proof of Lachlan's theorem to PA^-, we establish that $\mathsf{CT}[\mathsf{PA}^-]$ is not semantically conservative over PA^-. Built from a lemma by Mateusz Łełyk, we also show that every computably enumerable extension of PA^- has a non-recursively saturated model. On the technical side, the landscape regarding conservativity and recursive saturation for PA^- is very similar to that of PA. On the philosophical side, compositional truth is much stronger than extended induction — it alone is sufficient to enforce recursive saturation.

Acknowledgements

I wish to first express my deepest gratitude to my supervisors, Prof. Albert Visser and Prof. Dick de Jongh. Albert introduced me to the wonderful field of arithmetical truth, and guided me patiently while giving me the freedom to explore. I still remember receiving his detailed replies to emails late at night, and the encouragement he gave me to develop my own philosophical viewpoints. His insights always helped when my thoughts were in a muddle. His sophisticated sense of humor made our weekly meetings even more enjoyable. Dick responded very quickly to my initial sketchy ideas on truth, thereby making this thesis possible. He ensured that everything –whether it is my understanding of theorems, the presentation of materials, or logistics – was on the right track. This thesis project would not have initiated, nor would it have been completed, without them.

I must also thank Dick for his support throughout the last two years. He helped me navigate the terrain of mathematical logic hand in hand, and our regular meetings and conversations made the ILLC feel like home.

I would also like to thank my committee members, Prof. Rineke Verbrugge and Dr. Thomas Schindler, for reading my thesis, raising insightful questions during the defense, and providing meticulous comments on my drafts.

My gratitude extends as well to my friends in Amsterdam. I wish to thank Klarise Marais for our weekly working sessions at Lab42, Ruiting Jiang for our discussions on truth, and Yuan Ma for our time together in the apartment near Amstelstation. Two other friends I wish to thank – though I fail to think of a fitting continuation for the verse – are Minzhe Li and Tenyo Takahashi.

Finally, to Mom and Dad: thank you for everything.

Contents

Introduction

This thesis is seated at a juncture of several mathematical-philosophical themes.

Axiomatic Theories of Truth. Tarski's undefinability theorem (Tarski [Tar56]) states that defining a truth predicate for any language satisfying reasonable expressive constraints requires resources that exceed those of the language whose truth predicate is being defined. This should abolish the project of defining truth in formal languages. This thesis follows an alternative route — the axiomatic route — where truth is assumed to be a primitive, undefined notion. A theory of truth is thus not a list of necessary and sufficient conditions for a truthbearer to fall under the extension of the truth predicate, but a set of axioms that describe its content.

One is generally free in laying down axioms of truth. A minimal criterion for axiomatic theories is that they should derive what is called "the Tarski biconditionals", theorems of the shape $T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$, where φ is a schematic variable for sentences in the language, and "$\ulcorner\urcorner$" maps each sentence to its name. Therefore, the minimal truth theory is usually assumed to be just the set of all Tarski Biconditionals, denoted as $\mathsf{TT}[B]$, where B is the base theory. This thesis focuses on two further design choices, i.e., compositionality (that truth is compositional with respect to logical connectives, denoted as CT) and extended induction (that one can reason by mathematical induction on formulas containing the truth predicate, denoted as superscript $^+$). This gives four candidates, whose performance will be evaluated on (i) how well they model the content of the natural language truth predicate, and (ii) how well they suit the maxim of certain philosophical positions.

The Conservativeness Argument Against Deflationism. Shapiro [Sha98] and Ketland [Ket99] independently made what is dubbed "the conservativeness argument" against deflationism. The argument can be framed from the perspective of axiomatic truth theories as proposing a novel desideratum for any deflationism-friendly truth theory: conservativity over the base theory. The motivation is as follows. Deflationism, broadly construed, takes truth as a semantically insubstantial notion that only plays an expressive role in everyday discourses (e.g., it can be used to form universal generalizations like "All Tarski said about truth is true.")[1] Therefore, according to deflationism, adding a theory of truth should not be able to provide more semantic insight into the base theory. In particular, it should not prove any new theorems that are not originally provable in the theory without the truth predicate. Because if it did, then one must invoke the truth predicate in explaining the content and the proof of the new theorem. Truth, therefore, plays an explanatory role that exceeds the expressive role deflationists assign it.

Conservativity does not coordinate well with other general requirements for truth theories. We note two. First, any truth theory is expected to prove generalizations that one could make with the natural language truth predicate. For example, if the truth theory proves that for every sentence bivalence holds, i.e., it is true or its negation is true, one would expect it could also prove the universal generalizations of the bivalence principle. The theory of truth with only the biconditionals

[1] The characterization is perhaps too crude. But since deflationism is not the central topic of our thesis, we refer the reader to the SEP page [ASW23] for a more comprehensive overview.

has been charged by Tarski [Tar56] and Gupta [Gup93] for its deductive weakness. We dub the requirement "Deductive strength." Second, some (e.g. [Fef91], [Sha98]) consider the principle of induction fundamental to our understanding of numbers — and syntax, since syntax is usually represented in the context of axiomatic theories of truth as numbers, therefore any truth theory whose base theory contains induction (e.g. Peano Arithmetic PA) should automatically extended the principle to sentences with truth predicates. This is the requirement of extended induction. One obtains the following unfortunate survey of theories when combining the three individually justified requirements for truth:

	Deductive Strength	Extended Induction	Conservative
$\mathsf{TT}[\mathsf{PA}]$	No	No	Yes
$\mathsf{TT}^{+}[\mathsf{PA}]$	No	Yes	Yes
$\mathsf{CT}[\mathsf{PA}]$	Yes	No	Yes
$\mathsf{CT}^{+}[\mathsf{PA}]$	Yes	Yes	No

Where each theory fails to meet at least one expectation. The non-conservativity of $\mathsf{CT}^{+}[\mathsf{PA}]$ is worth explaining. Since PA is known to be able to represent syntactical objects like formulas and proofs as numbers, one can reason by induction on the length of proofs. Since all PA axioms are true and truth is preserved by modus ponens and universal generalization, the conclusion of all PA-proofs is true. Since there exists at least one false arithmetical sentence $0 = 1$, PA is consistent. By Gödel's second incompleteness theorem, PA cannot prove its own consistency, so $\mathsf{CT}^{+}[\mathsf{PA}]$ is not a conservative extension over PA. A trilemma is forced upon the deflationist: either one opt with an expressively adequate truth theory that proves significantly more than what the deflationist agenda allows it to prove, or to maintain a theory so deductively impoverished that it cannot serve to model the content of a natural language truth predicate, or to hold a philosophically unnatural view of the induction principle[2].

The Logical Strength of Compositionality. One can find in the table above that neither the induction principle alone ($\mathsf{TT}^{+}[\mathsf{PA}]$), nor compositionality alone ($\mathsf{CT}[\mathsf{PA}]$), can induce non-conservativity. Therefore, the dialogue between the deflationists and the anti-deflationists moved into a standstill. In general, deflationists wish to blame the non-conservativity result on extended induction, thus to occupy the position of $\mathsf{CT}[\mathsf{PA}]$, while their opponents wish to show the opposite, that extended induction is not the one to blame for non-conservativity.

Heck [Hec18] proposes a way to disentangle the contributions of each theory to the non-conservativity result, namely, to measure the logical strength of both in isolation. This means comparing the strength of $\mathsf{TT}^{+}[\mathsf{PA}]$ and $\mathsf{CT}[\mathsf{PA}]$ with PA. It turns out that even though the compositional truth theory without extended induction axioms is conservative over the base theory, the resulting theory is still much stronger, in the sense that it is not interpretable in the base theory. On the other hand, even if we add induction to the Tarski biconditionals, the resulting truth theory is, in most cases, interpretable in the base theory, which means that it is not logically stronger.

Plan. This thesis builds on Heck's project of disentangling the contribution of induction and compositionality to the non-conservativity result. In particular, we specify the logical strength of compositional truth by proving several conservativity results of compositional truth without extended induction in an induction-free fragment of Peano Arithmetic.

To motivate the choice of base and truth theory, we outline a certain technical nuisance surrounding the induction schema in interaction with truth theories. Recall that the usual setup for axiomatic truth theories assumes the truth predicate as a predicate on numbers, given a presupposed encoding of syntactical objects. Hence, in any model $\mathcal{M}$

[2] Most philosophers go with accepting option 3 and argue that there is nothing unnatural in restricting the induction schema.

of the base theory T, the truth predicate has the same ontological status as arithmetical predicates like "is prime" or "is even" — the extensions of both are sets of numbers. Since classical first-order logic cannot distinguish between different intensions of predicates with the same extension, we therefore operate in a setup that conflates two ontological categories. It fails to reflect that truth in natural language is a predicate on truth-bearers — usually assumed to be sentences or propositions, but are never numbers — and "is even" always predicates numbers. More generally, the equivocation makes one unable to distinguish the object theory whose truth we are reasoning about, and the syntactic theory in which we reason about the object theory. Consequently, the induction schema plays a double role of both an arithmetical induction schema on numbers, and a syntactical induction schema on the structure of sentences.

The representation of syntactical objects as numbers is possibly passed down from the age of Gödel's incompleteness proofs, where it was necessary and an invention of genius. However, it turns out to be a rather problematic feature for axiomatic theories of truth, because the entanglement of syntactical induction and arithmetic induction allows us to prove controversial statements about truth.

Suppose we want to investigate the functioning of syntactical induction in truth theories. We observe that a fixed amount of syntactic induction is sufficient in showing the consistency of the base theory. Take $\mathsf{PA}(\mathsf{T})$, i.e., PA formulated in $\mathcal{L}_{\mathsf{PA}} \cup \{\mathsf{T}\}$ and whose induction schema extends to formulas containing truth, as an example. Recall the inductive argument that establishes everything provable in PA in x steps is true. Since the inductive hypothesis "if φ is the conclusion of a proof in PA encoded by a number less than n, then φ is true" can be formalized with only one unbounded existential quantifier, the proof can be done in a fragment of $\mathsf{PA}(\mathsf{T})$ where one restricts the induction scheme to formulas with at most one unbounded existential quantifier, known as $\mathrm{I}\Sigma_1(\mathsf{T})$.

Lemma 1. $\mathsf{CT}[\mathrm{I}\Sigma_1] + \mathsf{T}(\mathsf{PA})$ *proves* $\mathsf{Con}(\mathsf{PA})$.

where $\mathsf{T}(\mathsf{PA})$ is the formalization that all axioms of PA are true. Conceptually, in the formulation of Lemma 1, we restrict the role of the $\mathrm{I}\Sigma_1$-extended induction to be syntactical, as an induction principle on proofs. Hence, $\mathsf{CT}[\mathrm{I}\Sigma_1]$ alone only establishes the inductive case of the proof. We still need $\mathsf{T}(\mathsf{PA})$ for establishing the base case (which, as we shall see later, relies on $\mathrm{I}\Sigma_1$ arithmetic induction.) Since the inductive argument is the same for other base theories, $\mathrm{I}\Sigma_1$ is sufficient for proving the consistency of any theory:

Theorem 2. *Suppose* $T \supseteq \mathrm{I}\Sigma_1$. *Then* $\mathsf{CT}[T] + \mathsf{T}(U)$ *proves* $\mathsf{Con}(U)$.

where U can be seen as the object theory, and T provides a syntactic theory. However, the attempt is futile as the syntactic and arithmetic induction is manifested in the object language as the same induction schema, so one cannot forbid the allegedly Σ_1 syntactical induction in $\mathsf{CT}[\mathrm{I}\Sigma_1]$ to function as an Σ_1 arithmetical induction. Σ_1 arithmetical induction can prove, among many things, that universal generalization preserves truth. Important for our purpose, with both Σ_1 arithmetical and syntactical induction, we have:

Lemma 3. $\mathsf{CT}[\mathrm{I}\Sigma_1]$ *proves that all axioms of* PA *are true.*

which establishes the base case of the previous syntactical inductive argument. The $\mathsf{T}(\mathsf{PA})$ in Lemma 1 is superfluous.

Corollary 4. $\mathsf{CT}[\mathrm{I}\Sigma_1]$ *proves* $\mathsf{Con}(\mathsf{PA})$.

which is weird in the sense that the compositional truth of a weaker theory proves the consistency of a stronger theory[3].

Naturally, one would like to study the properties of a logical axiom in a "controlled" environment. Just like conducting *ceteris paribus* experiments in the sciences, we wish to avoid the complexities induced by technical features that are highly specific to induction in the investigation of compositionality by making the contribution of induction "fixed", and

[3]We refer to Heck [Hec18] and Leigh and Nicolai [LN13] for a detailed analysis of the entanglement.

vice versa. In practice, we follow a method common in the study of "weak arithmetic" — arithmetical systems weaker than PA — where one examines the logical power of a schematic axiom by restricting it to certain sets of formulas. Hence, our question "How does compositional truth and induction interact?" is broken down into a series of questions: "How does compositional truth and induction as restricted to set X interact?[4]", where X is some subset of arithmetical formula. This thesis focuses on the simplest task of investigating the behavior of a compositional truth theory in a base theory that lacks induction altogether. We follow the arithmetician's standard choice of induction-free arithmetic — the theory of the positive part of discretely ordered rings PA^-. We aim to answer the following two questions on the strength of compositionality:

1. Is compositional truth conservative in PA^-?
2. Are there any other notions of conservativeness? What is the behavior of compositional truth in PA^- with respect to these other notions?

A Short Answer: The strength of compositional truth does not manifest itself in the sense of being able to prove more theorems when added to the base theory, but it does have the power of significantly restricting the number of models of the theory.

The short answer will be elaborated in full detail. The first question will be addressed in Chapter 2 by generalizing the model-theoretic proof of the conservativeness of the compositional truth for PA by Enayat & Visser [EV15]. Many claim that the result is generalizable to weaker systems. The thesis vindicates these claims by giving a detailed implementation of the proof starting from the arithmetization of syntax, which is made possible by Jeřábek [Jeř12]'s result that PA^- is sequential. In Chapter 3, we will look into the notion of model-theoretic conservativity, which intuitively concerns the number of models for a theory. The notion has been studied widely in the mathematical literature (e.g., see [Lac81],[Kay91],[Wci17]) but relatively underdeveloped in the philosophical literature on deflationism[5]. As we will see, as is the case with PA, compositionality in itself is sufficient to enforce recursive saturation in PA^-. With an extra insight by Lelyk, we also show that $\mathsf{CT}[\mathsf{PA}^-]$ is not semantically conservative, and any compositional truth to any computably enumerable (thereafter c.e.) extension of PA^- is not semantically conservative.

For someone interested in seeing a particular theorem, she is suggested to read the following chapters:

- For a proof of the Enayat-Visser result for PA^-, read Chapter 1 and Chapter 2.
- For a proof of Lachlan's theorem for PA^-, read Chapter 1 and Section 3.2.

[4]We only focus on theories weaker than PA. Let T a theory in language $\mathcal{L}$, and T^+ a conservative extension of T in language $\mathcal{L}^+$ with $\mathcal{L} \subseteq \mathcal{L}^+$. For any theory $T + U \supseteq T$ in $\mathcal{L}$, T^+ is a conservative extension over $T + U$.

Proof. Consider arbitrary $\mathcal{L}$-sentence φ where $T^+ + (U + T) \vdash \varphi$. Since every proof is finite, we may define the finite conjunction u of $u_i \in U$ that occur in the proof of φ. Therefore $T^+ + u \vdash \varphi$, $T^+ + u + T \vdash \varphi$, so $T^+ + T \vdash u \to \varphi$. By conservativeness of T^+, $T \vdash u \to \varphi$, $T + U \vdash \varphi$. □

In short, conservativity is upwards closed. So for any $T \supseteq \mathsf{PA}$, the conservativity result of $\mathsf{TT}[T]$, $\mathsf{TT}^+[T]$, $\mathsf{CT}[T]$, $\mathsf{CT}^+[T]$ is exactly like PA.

[5]With the exception of Waxman 2017 [Wax17], but whose focus on the intended model makes the notion of model-theoretic conservativity trivial because there is by stipulation only one model.

Chapter 1

Preliminaries

We introduce the varieties of arithmetical theories in Section 1.1, including our theory of interest, PA^-, a subtheory of Peano Arithmetic that is induction-free but still sequential. Different ways of adding a truth predicate into an arithmetical theory will be discussed in Section 1.2. Finally, Section 1.3 introduces methods of comparing the logical strength of theories.

1.1 Fragments of Arithmetic

All theories[1] discussed in this thesis will be formulated in the language of first-order arithmetic $\mathcal{L}_{\mathsf{PA}}$.

Definition 5. *The arithmetical language $\mathcal{L}_{\mathsf{PA}}$ has logical vocabulary $\{\neg, \vee, \exists\}$ and arithmetical vocabulary $\{0, ', +, \times, =, <\}$ (where $'$ is interpreted as successor).*

We assume the underlying logic to be classical first-order logic. We treat $\wedge, \rightarrow, \leftrightarrow, \forall$ as derivative. Therefore the expression $\varphi \wedge \psi$ is an abbreviation of $\neg(\neg\varphi \vee \neg\psi)$, $\varphi \rightarrow \psi$ an abbreviation of $\neg\varphi \vee \psi$, $\varphi \leftrightarrow \psi$ is an abbreviation of $(\varphi \rightarrow \psi) \wedge (\psi \rightarrow \varphi)$, and $\forall x\varphi(x)$ an abbreviation of $\neg\exists x\neg\varphi(x)$.

We also assume $\mathcal{L}_{\mathsf{PA}}$ to be a relational language. It does not have terms except for constant 0 and variables. Therefore $', +, \times$ are not term-forming operations, as they are usually assumed to be in forming the successor of a number x', the sum of numbers $x + y$ and the product of numbers $x \times y$, but are relations. The binary relation $'xy$ has intended meaning "y is a successor of x", and the ternary relation $+xyz$ ($\times xyz$) means "z is the sum (product) of x and y." The relational language and the term language for arithmetic are known to be inter-translatable by the term expansion algorithm. The idea is to replace every occurrence of a term t in the term language in the formula $\varphi(t)$ with a variable, and modify φ accordingly into $\exists x\varphi'(x)$. For example, a claim in the term language $a + b = c \times d$ expands into $\exists e\exists f(+abe \wedge \times cdf \wedge e = f)$. Axioms are added to stipulate that $', +, \times$ are bijections, etc. Given the translation, we still write the axioms in a term language for readability.

Definition 6. *An $\mathcal{L}_{\mathsf{PA}}$-theory T is a set of $\mathcal{L}_{\mathsf{PA}}$-formulas closed under logical consequence.*

A theory T can be generated by a set of axioms by taking the collection of its logical consequences. Our discussion is indifferent to the distinction between a theory and the axioms that generate it. We will be concerned with Peano Arithmetic and its subtheories.

[1] For a discussion about whether arithmetical truth is a good starting point for modeling the behavior of the truth predicate in natural language, see Fujimoto 2019 [Fuj19].

Definition 7 (Peano Arithmetic PA)**.** *We denote Peano Arithmetic* PA *the universal closure of the following $\mathcal{L}$-formulas:*

(PA1) $$x' \neq 0$$

(PA2) $$x' = y' \to x = y$$

(PA3) $$x + 0 = x$$

(PA4) $$x + y' = (x + y)'$$

(PA5) $$x \times 0 = 0$$

(PA6) $$x \times y' = (x \times y) + x$$

(PA7) $$x \neq 0 \to \exists y(x = y')$$

(PA8) $$x < y \leftrightarrow \exists z(y = z' + x)$$

And the axiom schema of arithmetical induction:

(IND) $$[\Phi(0) \wedge \forall x(\Phi(x) \to \Phi(x'))] \to \forall x \Phi(x)$$

where $\Phi(x)$ is a schematic variable ranging over all $\mathcal{L}_{\mathsf{PA}}$-formulas with x free.

The intended model for PA is the natural numbers $\mathbb{N}$. We also occasionally write the set of natural numbers as ω, especially in contexts where the set of all natural numbers is considered as an ordinal. For all $n \in \mathbb{N}$, the numeral for n, noted as $\bar{n}$ is the symbol 0 succeeded by n successor symbols. The numeral for number 0 is 0, and the numeral for n is $0\underbrace{'\cdots'}_{n}$.

Among the subtheories of PA, the weakest system we will be considering is Q, the induction-free fragment of PA.

Definition 8 (Robinson's Arithmetic Q)**.** *We denote Robinson's Arithmetic as the universal closures of* PA1 *to* PA8*, i.e,. Peano Arithmetic without the induction schema.*

Intermediate between Q and PA, there is a group of theories which can be obtained by restricting the induction schema to formulas of certain complexities in the arithmetical hierarchy.

Definition 9 (Bounded Quantification)**.** *We define bounded quantification $\forall x < n\ \varphi(x)$ and $\exists x < n\ \varphi(x)$ as abbreviations of $\forall x(x < n \to \varphi(x))$ and $\exists x(x < n \wedge \varphi(x))$.*

Definition 10 (Arithmetical Hierarchy)**.** *The arithmetical hierarchy is a syntactical classification of formulas in prenex normal form based on the structure of their quantifiers. $\Sigma_0 = \Pi_0 = \Delta_0$ formulas are the ones with only bounded quantifiers. For $n \geq 1$, Σ_n formulas are of the shape $\exists x \varphi(x)$ where $\varphi(x)$ is a Π_{n-1} formula, and Π_n formulas are of the shape $\forall x \varphi(x)$ where $\varphi(x)$ is a Σ_{n-1} formula.*

A formula φ is provably $\Sigma_n(\Pi_n)$ in a set of axioms T if there exists a $\Sigma_n(\Pi_n)$ formula ψ s.t. $T \vdash \varphi \leftrightarrow \psi$. φ is probably Δ_n in T if it is provably Σ_n and Π_n in T.

Definition 11 ($\mathrm{I}\Sigma_n, \mathrm{I}\Pi_n, \mathrm{I}\Delta_n$)**.** *We define $\mathrm{I}\Sigma_n$ (or $\mathrm{I}\Pi_n, \mathrm{I}\Delta_n$) as* Q *with the induction schema where $\Phi(x)$ ranges over $\mathcal{L}_{\mathsf{PA}}$-formulas with complexity Σ_n (or Π_n, Δ_n).*

$\mathrm{I}\Sigma_n, \mathrm{I}\Pi_n, \mathrm{I}\Delta_n$ are less ideal as starting points for investigating the logical strength of compositional truth, as they contain certain forms of the induction principle, which might interfere with compositional truth. Q is induction-free, but for reasons that we will see later, it is too weak in itself to host a truth theory. A subtheory of PA that is both strong enough for a truth theory and is free of induction is the theory of the positive part of discretely ordered rings.

Definition 12 (PA^-)**.** *Let* PA^- *be the theory of the positive part of discretely ordered rings with a least element, i.e.,* PA^- *is axiomatized by the following* $\mathcal{L}_{\mathsf{PA}}$*-formulas:*

(A1)	$(x + y) + z = x + (y + z)$
(A2)	$x + y = y + x$
(M1)	$(x \times y) \times z = x \times (y \times z)$
(M2)	$x \times y = y \times x$
(AM)	$x \times (y + z) = x \times y + x \times z$
(Z1)	$(x + 0 = x) \wedge (x \times 0 = 0)$
(I1)	$x \times 1 = x$
(O1)	$(x < y \wedge y < z) \rightarrow x < z$
(O2)	$\neg x < x$
(O3)	$x < y \vee x = y \vee y < x$
(O4)	$x < y \rightarrow x + z < y + z$
(O5)	$0 < z \wedge x < y \rightarrow x \times z < y \times z$
(S1)	$x < y \rightarrow \exists z(x + z = y)$
(Z2)	$0 < 1 \wedge (x > 0 \rightarrow x \geq 1)$
(Z3)	$x \geq 0$

Where A stands for addition, M for multiplication, Z for Zero, I for identity, O for ordering, and S for subtraction. The meaning of the axioms should be clear.

The intended model of PA^- is $\mathbb{Z}[X]^+$. $\mathbb{Z}[X]$ is the ring of polynomials with one variable X and coefficients from $\mathbb{Z}$. The order on $\mathbb{Z}[X]$ is the natural ordering obtained by making X infinitely large. More specifically, if $a_0, \ldots, a_n \in \mathbb{Z} \setminus \{0\}$ are coefficients of $x = a_0 + a_1X + a_2X^2 + \cdots + a_nX^n \in \mathbb{Z}[X]$, $x > 0$ iff $a_n > 0$, and for $p, q \in \mathbb{Z}$, $p > q$ iff $p - q > 0$. We define $\mathbb{Z}[X]^+$ as $\{p \in \mathbb{Z}[X] \mid \mathbb{Z}[X]^+ \vDash p > 0\}$. Since PA^- is a system weaker than PA, $\mathbb{N}$ and all non-standard models of PA are models of PA^-[2].

1.1.1 Definability in a Fragment

In adding a truth theory, we will be interested in the arithmetical system's ability to represent different notions, which is captured by the formal notion of (provable) definability and expressibility.

Definition 13 (Definability)**.** *A set of natural numbers (or sequences of natural numbers) A is defined by a formula* $\varphi(\vec{x})$ *if* $n \in A$ *iff* $\mathbb{N} \vDash \varphi(n)$*, i.e.,* $\varphi(n)$ *is true.*

Definition 14 (Provable Definability)**.** *If S is a set of sentences of* $\mathcal{L}_{\mathsf{PA}}$*, and T is a set of expressions, we will say that T is provably definable from S if for some formula* $\varphi(x)$*,* $\varphi(\ulcorner\psi\urcorner)$ *is provable from S if and only if* ψ *belongs to T.*

Definition 15 (Expressible)**.** *A property* φ *of natural numbers is expressible iff the set* $\{n : \varphi(n)\}$ *is definable. The notion of provably expressible is analogous.*

Definition 16. *A set* $X \subseteq \mathbb{N}$ *is* Σ_n *(or* Π_n*) if it is defined by a* Σ_n*-formula (or* Π_n*-formula) with exactly one free variable. An* n*-ary relation is* Σ_n *(or* Π_n*) if it is defined similarly by a formula of the corresponding complexity with* n

[2] For details of the theory, see Kaye [Kay91], Chapter 2.

variables free. A function f is Σ_n (or Π_n) if its graph $\Gamma(f) := \{(x, y) \mid f(x) = y\}$ is Σ_n (or Π_n).

Since most syntactical concepts are defined recursively, primitive recursive functions form a class of functions important for truth theories.

Definition 17 (Primitive Recursion)**.** *Given n-ary function g and $(n+2)$-ary function h, we may define $(n+1)$-ary function f from g and h by primitive recursion:*

$$f(\vec{x}, 0) = g(\vec{x})$$
$$f(\vec{x}, n+1) = h(f(\vec{x}, n), \vec{x}, n)$$

where $\vec{x} = x_1 \dots x_n$. We allow $n = 0$, where $g(\vec{x})$ is a constant.

Definition 18 (Primitive Recursive Functions)**.** *Primitive recursive functions are the smallest class of functions with the following property:*

- *The 0-ary constant zero function $\mathsf{0}$ is primitive recursive.*
- *The unary successor function $S(x) = x + 1$ is primitive recursive.*
- *For any $n, i \in \mathbb{N}$, the projection function $\pi^i_n(x_1, \dots, x_n) = x_i$ is primitive recursive.*
- *For primitive recursive functions g and $h_1 ... h_m$, the function composition $f(\vec{x}) = g(h_1(\vec{x}), \dots, h_m(\vec{x}))$ is primitive recursive.*
- *For n-ary function g and $n+2$-ary function h that are primitive recursive, f obtained from g, h by primitive recursion is primitive recursive.*

1.1.2 Models of Arithmetic

We will also be concerned with the models, especially non-standard models, of arithmetic. The reader is assumed to be familiar with general concepts in model theory. We will use $\mathcal{M}$ to denote the model, and M to denote the domain of the model $\mathcal{M}$.

A theorem important for the proof of the syntactic conservativity is Tarski's elementary chain theorem.

Definition 19 (Expansion, Extension, Elementary extension)**.** *Let $\mathcal{M}$ and $\mathcal{K}$ be models in the same language $\mathcal{L}$.*

- *$\mathcal{M}$ is an extension of $\mathcal{K}$ iff $K \subseteq M$, and the relations and function of $\mathcal{K}$ are relations and function of $\mathcal{M}$ restricted to K.*
- *$\mathcal{M}$ is an expansion of $\mathcal{K}$ iff $\mathcal{M}$ and $\mathcal{K}$ are the same except that $\mathcal{M}$ contains new predicate, relations, functions, or constant symbols.*
- *$\mathcal{M}$ is an elementary extension of $\mathcal{K}$, denoted as $\mathcal{K} < \mathcal{M}$ if and only if $\mathcal{M}$ is an extension of $\mathcal{K}$, and for every formula $\varphi(x_1, \dots, x_n) \in \mathcal{L}$, for all $a_1, \dots, a_n \in K$ $\mathcal{K} \vDash \varphi(a_1, \dots, a_n)$ iff $\mathcal{M} \vDash \varphi(a_1, \dots, a_n)$.*
- *For a class of $\mathcal{L}$ formulas Γ, $\mathcal{M}$ is a Γ-elementary extension of $\mathcal{K}$, denoted as $\mathcal{K} <_\Gamma \mathcal{M}$, if for all $a_1, \dots, a_n \in K$ and $\gamma(x_1, \dots, x_n) \in \Gamma$, $\mathcal{K} \vDash \gamma(a_1, \dots, a_n)$ iff $\mathcal{M} \vDash \gamma(a_1, \dots, a_n)$.*

Definition 20 (Elementary chain)**.** *An elementary chain of models is a family of models $\{\mathcal{M}_n : n \in \mathbb{N}\}$ such that for every $k, n \in \mathbb{N}$, if $k < n$, then $\mathcal{M}_k < \mathcal{M}_n$.*

Definition 21 (Limit Model)**.** *The limit model of an elementary chain $\{\mathcal{M}_n : n \in \mathbb{N}\}$ is defined as the model $\mathcal{M}$ whose domain $M = \bigcup_{n\in\mathbb{N}} M_n$, whose relation and function symbols are the union of all the relation and function symbols of the $\mathcal{M}_n$-s, and has the same constants as in the $\mathcal{M}_n$-s.*

Theorem 22 (Elementary chain theorem)**.** *Given an elementary chain, the limit model of an elementary chain is an elementary extension of every model in the chain.*

We will also be looking into non-standard models of arithmetic. Non-standard models are important because one cannot single out the natural numbers, even internally in a model, as a predicate. This fact is captured by the following lemma.

Lemma 23 (Overspill)**.** *Let $\mathcal{M}$ be a non-standard model of* PA *and $b \in M$, and assume $\varphi(x, y)$ is a formula with x, y free. Then, if $\mathcal{M} \vDash \varphi(n, b)$ holds for every $n \in \omega$, there is a non-standard number $c \in M$ such that $\mathcal{M} \vDash \forall x \leq c\, \varphi(x, b)$.*

Proof. We follow the presentation of Halbach 2010 [Hal10]. Suppose, for a contradiction, that there is no such c. Consider the formula defined as $\psi(x) :\leftrightarrow \forall y < x\, (\varphi(x, b))$. Since $\mathcal{M} \vDash \varphi(0, b)$, $\mathcal{M} \vDash \psi(0)$. Now consider any $a \in M$. If a is a standard number, then a' is also standard, so we have $\mathcal{M} \vDash \psi(a) \to \psi(a')$. Suppose a is a non-standard number, since there is no c such that $\psi(c)$ and the successor of non-standard numbers are always non-standard, $\mathcal{M} \vDash \psi(c) \to \psi(c')$. So $\mathcal{M} \vDash \forall x(\psi(x) \to \psi(x'))$. By induction, $\forall x \psi(x)$, a contradiction. □

The induction schema is necessary for the proof, so overspill only holds for $T \supseteq$ PA. For theories without full induction, one does not have full overspill, but the above argument generalizes to the fragment with induction. For example, $\mathrm{I}\Sigma_n$ has overspill for all Σ_n formulas. Dimitracopoulos [Dim89] shows that Σ_n-induction is equivalent to Σ_n-overspill.

Definition 24 (Type)**.** *Let $T \supseteq \mathsf{PA}^-$ be an arithmetical theory, and $\mathcal{M} \vDash T$. Let $p(\vec{x}) = \{\varphi_i(\vec{x}) \mid i \in I\}$ be any set of formulae sharing common variables $\vec{x} = x_0, \ldots, x_n$.*

- *p is a type over the theory T if $T + \{\varphi(\vec{c}) \mid \varphi(\vec{x}) \in p(\vec{x})\}$ is consistent, where $\vec{c}$ is a tuple of new constants.*
- *p is a type over the model $\mathcal{M}$ if for all finite subsets $I_0 \subseteq I$, $M \vDash \exists \vec{x} \wedge i \in I_0 \varphi_i(\vec{x})$.*
- *p is a realized type if p is a type over $\mathcal{M}$, and there exists $\vec{a} \in M$ such that for all $\varphi \in p$, $\mathcal{M} \vDash \varphi(\vec{a})$.*
- *p is a recursive type if p is a type over $\mathcal{M}$ and p is recursive.*
- *p is a complete type if for every $\varphi(x) \in$ Form, either $\varphi \in p$ or $\neg\varphi \in p$, otherwise p is partial.*
- *Let T be a theory and p be a partial type. Then $p(x)$ is isolated in T if there is a formula $\varphi(x)$ such that $\exists x \varphi(x)$ is consistent with T and $T \vdash \varphi(x) \to \sigma(x)$ for all $\sigma(x) \in p(x)$.*

The case where p has parameters $\vec{v} = v_0, \ldots, v_n$ $p(\vec{x}, \vec{v})$ is handled similarly.

Theorem 25 (Omitting Types Theorem)**.** *Let T be a consistent theory in a countable language. If a partial type $p(x)$ is not isolated in T, then there is a countable model of T which omits $p(x)$.*

Definition 26 (Recursive Saturation)**.** *A model $\mathcal{M}$ is recursively saturated iff every recursive type of $\mathcal{M}$ is realized in $\mathcal{M}$.*

One can replace "recursive" with "primitive recursive" in the definition; the resulting definition is equivalent to recursive saturation.

1.2 Theories of Truth

A truth theory is a set of axioms explicating the behavior of a truth predicate, whose intended extension is the set of all true sentences. To add a truth theory, one usually starts by choosing the theory whose truth predicate one wishes to investigate, known as the "base theory". In our case, this is always a first-order arithmetic theory, as introduced above. One adds the truth predicate by first syntactically expanding the language $\mathcal{L}_{\mathsf{PA}}$ with a one-place predicate T that does not belong to $\mathcal{L}_{\mathsf{PA}}$. The truth axioms are drafted in $\mathcal{L}^{+}_{\mathsf{PA}} = \mathcal{L}_{\mathsf{PA}} \cup \{\mathsf{T}\}$.

Only base theories satisfying certain expressive constraints can hold a truth theory. We note two. First, since truth is a predicate of syntactical objects[3] but the (intended) domain of arithmetical theories is numbers, defining truth as a predicate in $\mathcal{L}_{\mathsf{PA}}$ assumes an encoding of $\mathcal{L}_{\mathsf{PA}}$-formulas, an injective function from $\mathcal{L}_{\mathsf{PA}}$ formulas to natural numbers. Following Gödel, we denote $\ulcorner\varphi\urcorner$ as the encoding of φ. $\mathsf{T}(n)$ is understood as "the sentence φ with $\ulcorner\varphi\urcorner = n$ is true." A more subtle point originates in the treatment of variables and quantifiers. Tarski, in his influential paper *The Concept of Truth in Formalized Languages*, introduced sequences of objects in defining the satisfaction conditions for formulas with free variables and quantifiers. To illustrate, consider the arithmetical formula

(1) $$v_1 < v_2$$

where v_1 and v_2 are variables. Being an open formula, (1) does not yet have a truth value. It has truth values once numbers are assigned to v_1 and v_2. For example, (1) is true when v_1 is assigned 0 and v_2 is assigned 1. The order of assignments is important as (1) is not satisfied by the assignment of v_1 with 1 and v_2 with 0. Since the variables are indexed by natural numbers, the information of an assignment can be easily encoded as an ordered string of objects, where the value on the i-th digit corresponds to the value of variable v_i. An ordered string is known as a sequence. Since the language $\mathcal{L}_{\mathsf{PA}}$ is compositional, the number of variables in a formula can be any natural number. The sequences representing variable assignments can also be of arbitrary finite length. Defining a truth theory of arithmetical theory thus presupposes that the base theory can encode sequences. This requirement is formalized by Pudlák [Pud85].

Definition 27 (Sequential Theory)**.** *A theory T is sequential if T contains Q relativized some formula $N(x)$ (interpreted as natural numbers), and there exists a formula $\beta(x, i, w)$ such that*

(SEQ) $T \vdash \forall w, x, k\, \exists w'\, \forall i, y\, [(N(k) \wedge i \leq k) \rightarrow [\beta(y, i, w') \leftrightarrow ((i < k \wedge \beta(y, i, w) \vee (i = k \wedge y = x))]]$.

(SEQ) states that we can always append a further digit to the sequence encoded by w.

Visser [Vis08] and Jeřábek [Jeř12] show that Q is not sequential; therefore Q in itself is too weak for a truth theory. But since it interprets the sequential theory $\mathrm{I}\Delta_0$, most philosophical literature manually amends Q with the sequentiality axiom Q_{seq}, and treats it as the minimal base for truth. As Jeřábek [Jeř12] shows that PA^{-} is sequential, this thesis assumes PA^{-} as our base theory. We will show later that PA^{-} is sufficiently expressive to encode syntax, thus it meets both conditions for hosting a truth theory.

We consider two kinds of truth theory: a disquotational theory containing only the Tarski biconditionals and a compositional theory with additional compositional axioms. Their exact implementation will depend on several design choices of the language and the theory one adopts, including whether the language is relational or a term language,

[3] In fact, there is wide disagreement on what kind of objects truth-bearers are. To list a few: declarative sentences, declarative sentences in contexts, utterances of declarative sentences, propositions, the contents of thoughts, beliefs, judgments. Since the axiomatic approach to truth (in arithmetical languages) uses the grammatical structure of $\mathcal{L}_{\mathsf{PA}}$ formulas in defining truth, we assume truth bearers are sentences, or at least objects with a certain grammatical structure similar to their syntactical structure.

whether assignments are encoded as finite strings with arbitrary length or of infinite length, the details of numeral encoding, etc. Therefore different authors usually vary in their formulation of axioms, though the underlying idea is the same. We thus only present semi-formally the idea in this preliminary chapter, leaving the formal definitions for later in chapters 3 and 4. We will supplement the semi-formal explanation with occasional remarks on implementation details.

A *disquotational theory of truth* consists of all instances of the T-schema:

(T-schema) $$T(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$$

where φ is a schematic variable ranging over $\mathcal{L}_{\mathsf{PA}}$-sentences. Since allowing the truth predicate to occur within φ immediately gives rise to liar-paradoxes, we restrict φ as a schematic variable ranging over $\mathcal{L}_{\mathsf{PA}}$-sentences *without* the truth predicate. The truth predicate is thus "typed", as opposed to an untyped one that allows self-application of the form $\mathsf{T}(\ulcorner \varphi \urcorner)$, where T occurs in φ. (For a detailed discussion of typed vs. untyped truth, see Halbach [Hal10]). We also restrict the application of the T-schema to closed formulas only. Some (e.g., Cieśliński [Cie17]'s UTB) allow application of truth to open formulas by adding universal quantifiers at the front of the biconditional, so the T-schema for the open formula $\varphi(\vec{v})$ is $\forall \vec{t}(T(\ulcorner \varphi(\vec{t}) \urcorner) \leftrightarrow \varphi(\vec{t})$. Since our focus is on compositionality, we omit the internal variations on disquotational theories.

A *compositional theory of truth* consists of axioms for logical connectives:

(TSent) $$\mathsf{T}(\ulcorner \varphi \urcorner) \rightarrow \varphi \textit{ is a sentence}$$

(TR) $$R(t_0 \dots t_n) \leftrightarrow \mathsf{T}(\ulcorner R(t_0 \dots t_n) \urcorner)$$

(T$\neg$) $$\mathsf{T}(\ulcorner \neg \varphi \urcorner) \leftrightarrow \neg \mathsf{T}(\ulcorner \varphi \urcorner)$$

(T$\vee$) $$\mathsf{T}(\ulcorner \varphi \vee \psi \urcorner) \leftrightarrow \mathsf{T}(\ulcorner \varphi \urcorner) \vee \mathsf{T}(\ulcorner \psi \urcorner)$$

(T$\exists$) $$\mathsf{T}(\ulcorner \exists x \varphi(x) \urcorner) \leftrightarrow \exists z \mathsf{T}(\ulcorner \varphi(\bar{z}) \urcorner)$$

Both the disquotational and the compositional theories can be supplemented with the principle of extended induction:

(Extended IND) $$\Psi(0) \wedge \forall x(\Psi(x) \rightarrow \Psi(x')) \rightarrow \forall x \Psi(x)$$

where Ψ ranges over all $\mathcal{L}^{+}_{\mathsf{PA}}$ formulas with x free. Combining these design choices gives four truth theories:

Definition 28. *Fix an arithmetical base theory $T \supseteq Q$, where we assume T to be sequential.*

- $\mathsf{TT}[T]$ *is T with all Tarskian biconditionals for the language of T.*
- $\mathsf{TT}^{+}[T]$ *is* $\mathsf{TT}[T]$ *with extended induction to formulas $\varphi(x)$ possibly containing the truth predicate and with x free.*
- $\mathsf{CT}[T]$ *is T with all the axioms for compositional truth as described above.*
- $\mathsf{CT}^{+}[T]$ *is* $\mathsf{CT}[T]$ *with extended induction to formulas $\varphi(x)$ possibly containing the truth predicate and with x free.*
- $\mathsf{UTB}^{+}[T]$ *is T with all Tarskian biconditionals with variables for the language of T and extended induction.*

1.2.1 Truth and Satisfaction

Our discussion will also invoke the concept of satisfaction. While truth is a one-place predicate on closed sentences, satisfaction is a relation between assignments and open formulas (a two-place predicate on assignment-formula pairs). The conceptual connection between truth and satisfaction is as follows:

(2) An open formula $\varphi(x_1, \ldots, x_n)$ is satisfied by the variable assignment α iff $\varphi(\bar{a}_1, \ldots, \bar{a}_n)$ is true,

where $\alpha(x_i) = a_i$ for each i.

This conceptual relation signals a connection between a theory of satisfaction and a theory of truth. Just like a theory of truth, a theory of satisfaction is obtained by expanding the language $\mathcal{L}_{\mathsf{PA}} \cup \{\mathsf{S}, \mathsf{F}\}$, where S is a two-place satisfaction predicate on assignment-formula pairs and F is a one-place predicate for formulas. In a term language, another three-place predicate Den is introduced, where the intended meaning of $\mathsf{Den}(\alpha, t, x)$ is that term t denotes object x in assignment α. To simplify matters, we follow Heck [Hec18]'s implementation of Tarski [Tar56]'s idea, where we assume that assignments are infinite sequences of term-object pairs; hence, they not only assign values to variables, but also to terms like constants and numerals, with the condition that every assignment assigns the constant and numeral the appropriate elements.

Just like truth, there is a disquotational and a compositional theory of satisfaction. The *disquotational* theory contains all axioms of the form

(S-Biconditionals) $$\mathsf{S}(\ulcorner \varphi(t_0, \ldots, t_n) \urcorner, \alpha) \leftrightarrow \varphi(a_0 \ldots a_{n_1}) \wedge \bigwedge_{i<n} \alpha(t_i) = a_i$$

where φ and t are schematic variables ranging over $\mathcal{L}_{\mathsf{PA}}$ formulas and its free variables. A *compositional theory of satisfaction* consists of axioms

(SR) $$\mathsf{S}(\ulcorner R(t_0, \ldots, t_n) \urcorner, \alpha) \leftrightarrow R(a_0 \ldots a_{n_1}) \wedge \bigwedge_{i<n} \alpha(t_i) = a_i$$

(S$\neg$) $$\mathsf{S}(\ulcorner \neg\varphi \urcorner, \alpha) \leftrightarrow \neg\mathsf{S}(\ulcorner \varphi \urcorner, \alpha)$$

(S$\vee$) $$\mathsf{S}(\ulcorner \varphi \vee \psi \urcorner, \alpha) \leftrightarrow (\mathsf{S}(\ulcorner \varphi \urcorner, \alpha) \vee \mathsf{S}(\ulcorner \psi \urcorner, \alpha)$$

(S$\exists$) $$\mathsf{S}(\ulcorner \exists x \varphi(x) \urcorner, \alpha) \leftrightarrow \exists \alpha' \overset{x}{\sim} \alpha\, \mathsf{S}(\ulcorner \varphi(x) \urcorner, \alpha')$$

where $\alpha' \overset{x}{\sim} \alpha$ means that α' differs from α only in its assignment to x. There are further axioms for arithmetical constants and denotation for a term language:

(0) $$\mathsf{Den}(\alpha, \ulcorner 0 \urcorner, x) \leftrightarrow x = 0$$

(') $$\mathsf{Den}(\alpha, \ulcorner t' \urcorner, x) \leftrightarrow \exists y(\mathsf{Den}(\alpha, t, y) \wedge x = y')$$

(+) $$\mathsf{Den}(\alpha, \ulcorner t_1 + t_2 \urcorner, x) \leftrightarrow \exists y \exists z(\mathsf{Den}(\alpha, t_1, y) \wedge \mathsf{Den}(\alpha, t_2, x) \wedge x = y + z)$$

($\times$) $$\mathsf{Den}(\alpha, \ulcorner t_1 \times t_2 \urcorner, x) \leftrightarrow \exists y \exists z(\mathsf{Den}(\alpha, t_1, y) \wedge \mathsf{Den}(\alpha, t_2, x) \wedge x = y \times z)$$

In a relational language where one need not care about the assignment of terms, one can also implement assignments as sequences of arbitrary *finite* length that only assign values to the free variables of a formula. In that case, the axioms for

logical constants need modification. The clauses for disjunction and quantifier become:

$$(\mathsf{S}\vee') \qquad \mathsf{S}(\ulcorner\varphi \vee \psi\urcorner, \alpha) \leftrightarrow (\mathsf{S}(\ulcorner\varphi\urcorner, \alpha \restriction \mathsf{FV}(\varphi)) \vee \mathsf{S}(\ulcorner\psi\urcorner, \alpha \restriction \mathsf{FV}(\psi))$$

$$(\mathsf{S}\exists') \qquad \mathsf{S}(\ulcorner\exists x\varphi(x)\urcorner, \alpha) \leftrightarrow \exists\alpha' \supseteq \alpha\, \mathsf{S}(\ulcorner\varphi(x)\urcorner, \alpha')$$

where $\alpha' \supseteq \alpha$ means that α' is an extension of α. The two implementations of assignments are equivalent.

Definition 29. *Fix an arithmetical base theory $T \supseteq \mathsf{Q}$, where we assume T to be sequential*

- $\mathsf{TS}[T]$ *is T with all the Tarskian biconditionals and satisfaction biconditionals for $\mathcal{L}_T$-formulas.*
- $\mathsf{CS}[T]$ *is T with all the axioms for compositional satisfaction as described above.*

A theory of satisfaction opens up an alternative way of defining truth. Those who consider sentences as a special case of formulas, i.e., formulas with no free variables, would naturally see truth as a special case of the more general notion of satisfaction (e.g, see Wolenski [Wolo3]). This is supported by the fact that the above satisfaction axioms are well-defined for cases of $\mathsf{S}(\ulcorner\varphi\urcorner, \alpha)$ where φ is a sentence. Since all assignments are defined to assign each constant (numeral) the corresponding element in the domain, a sentence is always either satisfied by all variable assignments or none, unlike a formula will be satisfied by some variable assignments and unsatisfied by some others. The difference between a sentence and a formula is reduced to a technical one.

The behavior of sentences also suggests a definition of truth: a sentence is true if it is satisfied by all variable assignments (or, equivalently, it is satisfied by some assignment). One can thus define a truth theory by first equipping the base theory with a theory of satisfaction, then defining truth in terms of satisfaction.

$$(\mathrm{T}) \qquad \mathsf{T}(x) \leftrightarrow x \text{ is a sentence} \wedge \forall\alpha\mathsf{S}(x, \alpha).$$

Remark 30. *Truth defined via the compositional theory of truth and truth as defined via the compositional theory of satisfaction are equivalent. Let T be an arithmetical theory of reasonable strength, let $\mathcal{M} \vDash \mathsf{CT}[T] \cup \mathsf{CS}[T]$. For all $\mathcal{L}_{\mathsf{PA}}$-sentences φ, $\mathcal{M} \vDash \mathsf{T}(\ulcorner\varphi\urcorner)$ iff $\mathcal{M} \vDash \forall\alpha(\mathsf{S}(\ulcorner\varphi\urcorner))$.*

Proof. We prove the claim by induction. For simplicity, we suppose the only terms are numerals - the treatment of other terms should be similar. For the base case, let $\varphi = Rt_0 \dots t_n$. Suppose $\mathcal{M} \vDash \mathsf{T}(\varphi)$, which entails that $t_0 \dots t_n$ are all numerals. Therefore $\mathcal{M} \vDash \mathsf{T}(\varphi)$ iff $\mathcal{M} \vDash Rt_0 \dots t_n$ iff $\mathcal{M} \vDash \forall\alpha(\bigwedge_{i<n} t_i = a_i \wedge R(a_0, \dots, a_n))$ and φ is a sentence. - since any assignment α by definition assigns any numeral $\bar{n}$ the corresponding number n.

For the inductive case, we only consider the case where $\varphi = \exists v\psi(v)$. Since φ is a sentence, ψ has no other free variables other than v. Suppose $\mathcal{M} \vDash \mathsf{T}(\ulcorner\varphi\urcorner)$ which implies that $\mathcal{M} \vDash \exists z\mathsf{T}(\ulcorner\psi(\bar{z})\urcorner)$ and $\mathcal{M} \vDash \exists z[\forall\alpha\mathsf{S}(\psi(\bar{z}), \alpha) \wedge \psi(\bar{z})$ is a sentence.]. Since $\psi(\bar{z})$ is a sentence, $\exists x\psi(x)$ must be. Since $\mathcal{M} \vDash \forall\alpha\mathsf{S}(\psi(\bar{z}), \alpha)$, and every variable assignment assigns $\bar{z}$ the number z, $\psi(x)$ is satisfied by any assignment that assigns x z. Therefore
$\mathcal{M} \vDash \forall\alpha\exists\alpha' \overset{x}{\sim} \alpha\, \mathsf{S}(\psi, \alpha')$, where $\alpha'(x) = z$. It follows that $\mathcal{M} \vDash \forall\alpha\mathsf{S}(\exists x\psi(x), \alpha)$.

For the other direction, suppose $\mathcal{M} \vDash \forall\alpha\mathsf{S}(\exists x\psi(x), \alpha)$, therefore $\mathcal{M} \vDash \forall\alpha\exists\alpha' \overset{x}{\sim} \alpha\, \mathsf{S}(\psi(x), \alpha)$, we denote $\alpha'(x) = z$. We must have $\mathcal{M} \vDash \forall\alpha(\psi(\bar{z}), \alpha)$ since the denotation of $\bar{z}$ is z in each α. □

Note that the proof importantly relies on the existence of a numeral for every number n.

This thesis will treat truth not as a generalization of satisfaction but as a different notion, for two reasons. First, as hinted above, the derivation from satisfaction to truth is developed initially not as a result of philosophical reflection, but due to

a technical difficulty that truth refuses recursive characterizations. Suppose we want to design the recursive truth clauses for universally quantified formulas $\forall x\varphi(x)$. In the standard approach, the truth of $\forall x\varphi(x)$ should coordinate in some way with the truth of $\varphi(x)$ via some use of the universal quantifier in the language where the recursive definition is given. Yet $\varphi(x)$, by definition, is an open formula that adopts no truth definition. It only has satisfaction conditions when paired with some variable assignments. Therefore, a natural response is to first define satisfaction recursively (which yields the familiar clause $\forall x\varphi(x)$ is satisfied by assignment α iff $\varphi(x)$ is satisfied by assignment α' that differs from α in x the most), then define truth from satisfaction. However, in the definition of truth axioms, we see that the problem vanishes once we introduce numerals. In the case of arithmetical truth, we can manually close the formula $\varphi(x)$ by replacing x with numerals, so that the notion of truth applies to it, as illustrated in axiom (T$\exists$). Second, in defining truth *in terms of* satisfaction, one risks alluding to a philosophical position that truth is *conceptually* derivative from satisfaction. There is a conceptual merit in distancing one's technical exposition from a debatable philosophical position, unless one wishes to defend it, which we do not.

In any case, the choice in establishing truth directly or via satisfaction affects the upcoming proofs only in minor technical details. We will be discussing the conservativity result for both truth and satisfaction. The intuitive conceptual connection between truth and satisfaction as sketched in (2) will also be preserved in our approach, as we will prove a theorem expressing exactly that.

1.3 Measurements of Logical Strength

Given theories T and B stated in $\mathcal{L}_T$ and $\mathcal{L}_B$ respectively, there are several ways to compare their logical strengths. When $\mathcal{L}_T = \mathcal{L}_B$, a theory is stronger when it can prove more theorems. The idea generalises smoothly to cases where one language contains the other. Supposing $\mathcal{L}_T \supseteq \mathcal{L}_B$, T and B can be compared by evaluating their logical strength in the restricted language $\mathcal{L}_B$. If T, despite being in a richer language, proves nothing more than B, it is, in a sense, not stronger than B on the subject matter of $\mathcal{L}_B$. This leads to the notion of syntactic conservativity.

Definition 31 (Syntactic Conservativity)**.** *Given theories T and T' stated in $\mathcal{L}$ and $\mathcal{L}'$ respectively, and $\mathcal{L} \subseteq \mathcal{L}'$. T' conservatively extends T if for any sentence A in $\mathcal{L}$*

- *whenever $T \vdash A$, then $T' \vdash A$,*
- *whenever $T' \vdash A$ and A is formulated in $\mathcal{L}$, then $T \vdash A$.*

Comparison is more difficult in the case where neither $\mathcal{L}_T \supseteq \mathcal{L}_B$ nor $\mathcal{L}_T \subseteq \mathcal{L}_B$. A method is introduced and examined by Tarski [Tar53], which, intuitively, involves translating one language into the other and comparing what can be proven modulo the translation.

Definition 32 (Relative Interpretability)**.** *Given two theories T (target) and B (base) stated in languages $\mathcal{L}_B$ and $\mathcal{L}_T$ respectively, T is relatively interpretable in B if there exists a relative interpretation of T, consisting of:*

- *a translation of $\mathcal{L}_T$ into $\mathcal{L}_B$,*
- *proofs in B of the translations of the axioms of T.*

Both relative interpretability and syntactic conservativity are syntactical notions. There is a semantic measurement of logical strength measured by the number of models of theories. When we consider the totality of models as exhausting the logical space, and theories as restrictions on the logical space by singling out the models in which the theory is true, the logical strength of a theory is naturally correlated with its ability to restrict the totality of models - the fewer models a theory has, the stronger it is. A tautology is weakest since it casts no restriction whatsoever, and a contradiction is the

strongest because it eliminates every logical possibility. This basic idea underlies many formal approaches to meaning and information states (e.g., inquisitive semantics [CGR18]).

Definition 33 (Semantic Conservativity)**.** *Let $T \subseteq T'$ be two theories and let $\mathcal{L}$ be the language of the theory T. T' is semantically conservative over T if for any model $\mathcal{M} \models T$ there exists an expansion $(\mathcal{M}, P_1, \ldots, P_\alpha)$ to a model of the theory T'.*

Remark 34 (Cieśliński [Cie17])**.** *For theories T_1 and T_2, if T_1 is semantically conservative over T_2 then T_1 is syntactically conservative over T_2.*

Proof. Suppose T_1 is not syntactically conservative over T_2, i.e. there exists φ where $T_2 \vdash \varphi$ but $T_1 \nvdash \varphi$. By completeness of first-order logic there exists $\mathcal{M} \models T_1 \cup \{\neg\varphi\}$. Suppose that T_1 is semantically conservative over T_2, then every model of T_1 can be expanded to a model of T_2, including that of $\mathcal{M}$. But since $T_2 \vdash \varphi$, φ is true in $\mathcal{M}$, a contradiction. □

As we shall see later, the converse does not hold. So semantic conservativity is stronger than syntactic conservativity.

Finally, since semantic conservativity concerns the number of models, it would help sketch a model-theoretic characterization of syntactic conservativity.

Remark 35 (Model-theoretic characterization of syntactic conservativity (Cieśliński [Cie17], p. 16))**.** *Given theories T_1 and T_2 stated in $\mathcal{L}_{T_1}$ and $\mathcal{L}_{T_2}$ respectively, and that $T_1 \subseteq T_2$. T_2 is syntactically conservative over T_1 iff for every model $\mathcal{M}$ of T_1, there exists a model $\mathcal{N}$ such that*

- *For every $\mathcal{L}_{T_1}$-sentence φ, $\mathcal{M} \models \varphi$ iff $\mathcal{N} \models \varphi$,*
- *$\mathcal{N} \models T_2$.*

Proof. For the left-to-right direction, suppose that T_2 is syntactically conservative over T_1. Suppose, for a contradiction, that for every model $\mathcal{M}$ of T_1, if $\mathcal{N} \models T_2$ then there exists some $\mathcal{L}$-sentence φ that is satisfied in only one of $\mathcal{N}, \mathcal{M}$. Therefore $\mathsf{Th}(\mathcal{M}) \cup T_2$ is inconsistent. By compactness, there exists a finite subset $A \subset \mathsf{Th}(\mathcal{M})$ where $A \cup T_2$ is inconsistent. Therefore $T_2 \vdash \neg A$. However, $T_1 \nvdash \neg A$ since $\mathcal{M}$ is a model of T_1. So T_2 is not syntactically conservative over T_1, a contradiction.

For the right-to-left direction, suppose that for every model $\mathcal{M}$ of T_1, there exists a model $\mathcal{N}$ satisfying the conditions. Suppose, for a contradiction, that T_2 is not syntactically conservative over T_1. Therefore there exists φ where $T_2 \vdash \varphi$ but $T_1 \nvdash \varphi$. Therefore there exists a model $\mathcal{M}$ of T_1 where $\mathcal{M} \nvDash \varphi$. Therefore there exists $\mathcal{N} \models T_2$ and $\mathcal{N} \nvDash \varphi$. But this contradicts that $T_2 \vdash \varphi$. □

Chapter 2

Compositional Truth for PA^- is Syntactically Conservative.

In this chapter, we tackle the first question sketched in the introduction: Is compositional truth conservative in PA^-? The answer is positive. We follow Enayat & Visser 2015 [EV15]'s model-theoretic proof that compositional truth without extended induction for Peano Arithmetic $\mathsf{CT}[\mathsf{PA}]$ is conservative over PA, and generalise the result to PA^-. Section 2.1 demonstrates that PA^- carries an arithmetisation of syntax. The conservativeness of satisfaction classes is proved in Section 2.2, and the case of truth will be proved in Section 2.4.

2.1 Some Properties of PA^-

We alluded in the preliminary section that arithmetization of syntax requires a base theory to be sequential and be able to represent the graphs of primitive recursive functions, so that they can capture the recursive definition of syntactical concepts. This is dealt with in Subsections 2.1.1 and 2.1.2, respectively.

2.1.1 Pairing and Sequence

We use round brackets (x, y) to denote pairs constituting individuals x, y, and sharp brackets $\langle x_0, x_2, \ldots, x_n\rangle$ to denote the sequence of objects $x_0, \ldots, x_n$. The conceptual difference between pairs and sequences, being two different kinds of objects, is reflected in the different arithmetical functions used to encode and decode such objects. For an arbitrary PA^--model $\mathcal{M}$ we have $\langle x, y\rangle \neq (x, y)$ for $x, y \in M$ in general.

Proposition 36. *There is a Δ_0 pairing and inverse pairing function in PA^-.*

Proof. Following Jeřábek 2012 [Jeř12], we define $(x, y) = (x + y)^2 + x$. To define the inverse pairing function, we aim to define projection functions that project the first and second items in a pair $\pi_1((x, y)) = x$ and $\pi_2((x, y)) = y$. Observe that $(x + y)^2 \leq (x + y)^2 + x = \langle x, y\rangle < (x + y + 1)^2$. So $n = x + y + 1$ is least number such that $n^2 > w$. We have a Δ_0 function $f(w) = n$ whose graph is defined by $n^2 > w \wedge (n - 1)^2 \leq w$. Define $\pi_1(w) = w - (f(w) - 1)^2$; $\pi_2(w) = (f(w) - 1) - \pi_1(w)$, both are Δ_0. □

The important result that makes the rest of the chapter possible is from Jeřábek 2012 [Jeř12], which shows that PA^-

is sequential. The proof implements Gödel's β-function via the Chinese remainder theorem in PA^-.[1] We omit the proof and refer the readers to [Jeř12]. It suffices for our purpose to show that given a sequence encoding function $\beta(a, b, i) = (s)_i$ where s is the sequence encoded by a, b, its graph $\Gamma(\beta)$ is definable in PA^-.

Proposition 38. *There is a formula $\beta'(x, i, w)$ with x, i, w free expressing the i-th element in sequence w is x in PA^-.*

Proof. We slightly modify the implementation of the β-function in [Jeř12] with an extra parameter denoting the length of the sequence. That is,

$$\beta(x, i, w) :\leftrightarrow \exists u, v, q, l\, [(w = (u, v, l) \wedge i < l) \wedge u = q(1 + (i + 1)v) + x) \wedge x \leq (i + 1)v)]$$

where (x, y, z) abbreviates $((x, y), z)$. To ensure that the β-function is total, it is implemented in PA^- as

$$\beta'(x, i, w) :\leftrightarrow [\beta(x, i, w) \wedge \forall j < \pi_2(w)\, \exists y\, \beta(y, j, w)]$$

Since $u, v \leq w$, and $q \leq u \leq w$, $\beta(x, i, w)$ is equivalent to a Δ_0 formula, also observe that $y \leq w$. So β' is Δ_0. □

This yields a natural characterization of sequence in the language of arithmetic:

Definition 39 (Sequence number)**.** $\mathsf{Seq}(w) :\leftrightarrow \exists u, v, l\, [(w = (u, v, l)) \wedge \forall i < l\, \exists x\, \beta(x, i, w)]$.

For readability, we denote $\ell(w) = \pi_2(w)$ as the length of sequence w, and rewrite $\beta'(x, i, w)$ as $(w)_i = x$. Implementing sequences via the β-function allows a sequence α to be encoded by multiple pairs (v_0, u_0) and (v_1, u_1). Therefore, the identity relation in (codes of) sequences should not be treated as extensional identity in standard first-order logic, but should be understood as forming equivalence classes of numbers that encode the same sequence.

Definition 40 (Identity of Sequences)**.** *Two numbers w_0, w_1 encode the same sequence, denoted as $w_0 \approx w_1$, if $\ell(w_0) = \ell(w_1)$ and $\forall i < \ell(w_0)\, \exists x < w_0\, (x = (w_0)_i \wedge x = (w_1)_i)$.*

This Δ_0 definition of $w_0 \approx w_1$ is only possible because for any PA^--model $\mathcal{M}$ and all $x, i, w \in M$ such that $\mathcal{M} \vDash (w)_i = x$, $\mathcal{M} \vDash x < w$.

PA^- therefore has the machinery to represent proofs as sequences of formulas, and assignments as sequences of variable-object pairs. However, to develop a theory of truth requires more than representing sequences as numbers. The functioning of assignments relies on operations like restricting the domain of an assignment and changing the value of a specific variable, which requires operations on sequences, including concatenation, shortening, splitting, etc. Our implementation of them uses a trick called "shortening of cuts", a well-celebrated practice in studying weak theories of arithmetic due to Robert Solovay. We first introduce the trick in the context of arithmetic.

[1] Gödel's β-function allows one to encode a sequence of natural numbers with arbitrary finite length with a pair of natural numbers. It is defined as follows:
Definition 37 (Gödel's β-function)**.**

(2.1) $\beta : \mathbb{N}^3 \to \mathbb{N}$

(2.2) $\beta(a, b, i) = \mathrm{rem}(1 + (i + 1)b, a)$

It is an application of the Chinese remainder theorem, which states that given $n_0, ..., n_k \in \mathbb{N}$ pairwise co-prime and $a_0, ..., a_k \in \mathbb{N}$, the system $x \equiv a_0 \pmod{n_0}, x \equiv a_1 \pmod{n_1}, ..., x \equiv a_k \pmod{n_k}$ has a solution.

Shortening of Cuts

Consider an arithmetical theory $T \supseteq \mathsf{Q}$ where T does not contain the axiom schema of mathematical induction

$$\varphi(0) \wedge \forall x(\varphi(x) \to \varphi(Sx)) \to \forall x \varphi(x)$$

where φ is a schematic variable for any $\mathcal{L}_{\mathsf{PA}}$-formula with x free. Almost for all theories T there is a formula φ where $T \vdash \varphi(0)$ and $T \vdash \forall x(\varphi(x) \to \varphi(x+1))$, yet $T \nvdash \forall x(\varphi(x))$ (Hájek and Pudlák [HP93], 1993, p. 172).

A well-known example is Q. Consider the formula $x \neq x'$ with x free. Q proves the antecedent of the corresponding induction principle for $x \neq x'$, i.e. $\mathsf{Q} \vdash 0 \neq 0'$ and $\mathsf{Q} \vdash \forall x(x \neq x' \to x' \neq x'')$. Yet it fails to establish the conclusion. It does not prove that no number is its own successor. Intuitively, the source of failure is that in some non-standard PA^--model $\mathcal{M}$, $x \neq x'$ holds for all natural numbers 0, 1, 2, etc. But there is one non-standard number "after" all the standard numbers where $x \neq x'$ does not hold. Formulas that have the status of $x \neq x'$ in Q, i.e. hold for 0 and are closed under successor, are called "inductive." Since PA^- does not have axioms for successor, we replace every occurrence of successor with $+1$.

Definition 41 (Inductive formula)**.** *Consider theory $T \supseteq \mathsf{PA}^-$, a formula $\varphi(x)$ with x free (x is understood as a number variable) is inductive in T if $T \vdash \varphi(0)$, and $T \vdash \forall x(\varphi(x) \to \varphi(x+1))$.*

Inductive formulas are true for all natural numbers. But their behavior on the non-standard numbers could be erratic. In a non-standard model $\mathcal{M}$, an inductive formula $I(x)$ could be true for all non-standard numbers, or true for some non-standard numbers i and j, but there exists some $i < k < j$ where $\neg I(k)$. Shortening of cuts shows that the behavior of $I(x)$ in the non-standard numbers can be handled in a principled way. For any inductive formula $I(x)$ and model $\mathcal{M}$, there is always an initial segment of $\mathcal{M}$ where I is true. This initial segment is known as a cut.

Definition 42 (Cut)**.** *Consider theory $T \supseteq \mathsf{PA}^-$, a formula $\varphi(x)$ with x free is a cut in T if it is inductive in T and in addition downward closed, i.e. $T \vdash \forall x(\varphi(x) \to \forall y(y < x \to \varphi(y)))$. A cut is proper in T if $T \nvdash \forall x \varphi(x)$.*

Theorem 43 (Hájek and Pudlák 1993, p. 368)**.** *Let $I(x)$ be inductive in $T \supseteq \mathsf{PA}^-$. Then there is a cut $J(x)$ in T for which $T \vdash \forall x(J(x) \to I(x))$.*

Proof. Define $J(x) := \forall w < x \, I(w)$. □

To define concatenation and other operations on sequences, we adapt the shortening of cuts to the context of sequences and formulas. The usual notion of cut does not apply directly to sequences since not all numbers encode sequences. We start by reviewing the basic operations on sequences - adjunction, where we append an extra number a to a sequence x, and concatenation, where we append sequence y to the end of x.

Definition 44 (Adjunction)**.** $z \approx x \circ a :\leftrightarrow \mathsf{Seq}(z) \wedge \mathsf{Seq}(x) \wedge$
$\forall i \leq \ell(x) \, [(i < \ell(x) \to (z)_i = (x)_i) \wedge (i = \ell(x) \to (z)_i = a)]$.

Definition 45 (Concatenation)**.** $z \approx x * y :\leftrightarrow \mathsf{Seq}(x) \wedge \mathsf{Seq}(y) \wedge \mathsf{Seq}(z) \wedge$
$\forall i < (\ell(x) + \ell(y)) \, [(i < \ell(x) \to (z)_i = (x)_i) \wedge (\ell(x) \leq i < \ell(x) + \ell(y) \to (z)_i = (y)_{i - \ell(x)})]$.

Note that both $\circ$ and $*$ are relational. Adjunction is comparable to the successor relation in numbers; we thus define the notion of Seq-cut accordingly.

Definition 46 (Seq-cut and Seq-inductive formula)**.** *Let $T \supseteq \mathsf{PA}^-$ be a theory in $\mathcal{L}_{\mathsf{PA}}$. S is Seq-inductive in T if*

- $T \vdash \forall x(S(x) \to \mathsf{Seq}(x))$,

- $T \vdash \forall x \forall a (x \in S \to \exists y (y = x \circ a \wedge y \in S))$.

S is a Seq*-cut in T if in addition $T \vdash \forall s \forall t \, (s \in S \wedge \ell(t) < \ell(s) \to t \in S)$.*

The relation between Seq-cut and cut is as follows:

Remark 47. *Let $T \supseteq \mathsf{PA}^-$ be a theory in $\mathcal{L}_{\mathsf{PA}}$. if S is* Seq*-inductive in T, then*

- *If there exists $s \in S$ where $\ell(s) = n$, then $\forall s (\ell(s) = n \to s \in S)$, and*
- *$\ell(S) := \{n \mid \exists s \in S, \ell(s) = n\}$ is inductive.*

S is an Seq*-cut in T if in addition $\ell(S)$ is downward closed.*

Remark 48 (Every Seq-inductive formula can be shortened to a Seq-cut)**.** *Let $I(x)$ be* Seq*-inductive in $T \supseteq \mathsf{PA}^-$. Then there is a* Seq*-cut $J(x)$ in T for which $T \vdash \forall x (J(x) \to I(x))$.*

Proof. Define $J(x) := \forall w (\ell(w) \leq \ell(x) \to I(w))$. □

Remark 49. *There is a* Seq*-cut* Seq_1 *closed under concatenation.*

Proof. Define $\mathsf{Seq}_1(x) = \mathsf{Seq}(x) \wedge \forall y (\mathsf{Seq}(y) \to y * x \downarrow)$, where $y * x \downarrow$ abbreviates $\exists z (z = y * x)$. By the definition of Seq-inductive formulas, it suffices to show that Seq_1 is closed under concatenation, as adjunction is equivalent to concatenation of sequences with length 1.

Trivially $\emptyset \in \mathsf{Seq}_1$. For all $x_0, x_1 \in \mathsf{Seq}_1$, we want to show that $x_0 * x_1 \downarrow$ and $x_0 * x_1 \in \mathsf{Seq}_1$, i.e. $\mathsf{Seq}(x_0 * x_1) \wedge \forall y (\mathsf{Seq}(y) \to y * (x_0 * x_1) \downarrow)$. For the first conjunct, since $x_0, x_1 \in \mathsf{Seq}_1$, $\mathsf{Seq}(x_0)$ and $\forall y (\mathsf{Seq}(y) \to y * x_1 \downarrow)$, so $x_0 * x_1 \downarrow$. For the second conjunct, consider an arbitrary $y \in \mathsf{Seq}$, we know that $y * (x_0 * x_1) \approx (y * x_0) * x_1$. Therefore $y * (x_0 * x_1)) \downarrow$ iff $((y * x_0) * x_1) \downarrow$. But we know that $(y * x_0) \downarrow$ since $x_0 \in \mathsf{Seq}_1$, and $((y * x_0) * x_1) \downarrow$ since $x_1 \in \mathsf{Seq}_1$. □

Before proceeding to prove further results, we define subsequences and segments of a sequence:

- "The sequence encoded by y is a subsequence of the sequence encoded by x":
  $x \preccurlyeq y := \mathsf{Seq}_1(x) \wedge \mathsf{Seq}_1(y) \wedge \ell(x) \leq \ell(y) \wedge (\exists n < \ell(y) - \ell(x)) \forall m < \ell(x) ((x)_m = (y)_{n+m})$
- "The sequence encoded by y is an initial segment of the sequence encoded by x"
  $y \preccurlyeq_{\mathfrak{i}} x := \mathsf{Seq}_1(x) \wedge \mathsf{Seq}_1(y) \wedge \ell(y) \leq \ell(x) \forall i < \ell(y) ((y)_i = (x)_i)$;
- "The sequence encoded by y is an end segment of the sequence encoded by x"
  $y \preccurlyeq_{\mathfrak{e}} x := \mathsf{Seq}_1(x) \wedge \mathsf{Seq}_1(y) \wedge \forall \ell(x) - \ell(y) < m < \ell(x) ((x)_m = (y)_{m-(\ell(x)-\ell(y))})$
- "y occurs before z in a sequence encoded by x"
  $y <_x z := \exists i < \ell(x) \exists j < \ell(x) ((x)_i = y \wedge (x)_j = z \wedge i < j)$.
- "x occurs in the sequence encoded by y" as $x \in y := \exists i < \ell(y) (x)_i = y$.

Proposition 50. *There exists a* Seq*-cut* Seq_2 *where for α a sequence with length x and $y \leq x$, there are sequences β and γ where β has length y and γ has length $x - y$ such that $\alpha = \beta * \gamma$.*

Proof. Define $\mathsf{EndSeg} = \{((x,k),y) \mid y \preccurlyeq_E x \wedge \ell(y) = k\}$, i.e. $\mathsf{EndSeg}(x,k)$ returns the end segment of x of length k. Similarly, define $\mathsf{InitSeg} = \{((x,k),y) \mid y \preccurlyeq_I x \wedge \ell(y) = k\}$. Let $\mathsf{Seq}_2(x) :\leftrightarrow \mathsf{Seq}_1(x)$
$\wedge \forall k (k \leq \ell(x) \to \exists y \exists z (y = \mathsf{InitSeg}(x,k) \wedge z = \mathsf{EndSeg}(x, \ell(x) - k) \wedge x = y * z)$.

Seq_2 is inductive: vacuously $\mathsf{Seq}_2(\emptyset)$. For any singleton $x = \langle n \rangle$, there exists initial segment $y = \langle n \rangle$ and $z = \emptyset$ where $x = y * z$. To show Seq_2 is closed under concatenation, let $x, y \in \mathsf{Seq}_2$ and $k \leq \ell(x) + \ell(y)$. Note that PA^- proves that $\exists w(w+k = \ell(x) \vee \ell(x)+w = k)$. If the first case hold, there exists $x', x'' \prec x$ where $x' = \mathsf{InitSeg}(x, w), x'' = \mathsf{EndSeg}(x, \ell(x) - w)$ and $x' * x'' = x$. Define $x''' = x'' * y$. It is easy to see that $x * y = x' * (x'' * y) = x' * x'''$. If the second case hold, there exists $y', y'' \prec y$ where $y' = \mathsf{InitSeg}(x, w), y'' = \mathsf{EndSeg}(x, \ell(y) - w)$ and $y' * y'' = y$. Define $y''' = x * y'$. It is easy to see that $x * y = (x * y') * y'' = y''' * y''$.

Seq_2 is closed under taking end segments of arbitrary length: let $x \in \mathsf{Seq}_2$ and $z = \mathsf{EndSeg}(x, y)$ for arbitrary y. We observe that all end segments of z are end segments of x. Since $\forall y(y \leq \ell(x) \to \mathsf{Seq}_1(\mathsf{EndSeg}(x, y))$, and $\forall y(y \leq \ell(z) \to y \leq \ell(x)), \forall y(y \leq \ell(z) \to \mathsf{Seq}_1(\mathsf{EndSeg}(z, y))$. Therefore $\mathsf{Seq}_2(z)$. The case for initial segments is similar. □

To implement the resetting of a variable in an assignment, we need to implement the subtraction of an element a from a sequence s. The easiest way is to define it by primitive recursion on sequences. We thus left the discussion of subtraction after primitive recursion. For readability, we rename Seq_2 as Seq.

2.1.2 Primitive Recursion, Other Results Concerning Expressivity.

The usual procedure of arithmetization of syntax first shows that all syntactical expressions can be expressed within a certain complexity on the arithmetical hierarchy, then proves that all formulas within the complexity are provably definable in PA^-. To generalize the procedure, we first prove some facts about what is expressible in PA^-. This subsection aims to prove two results: (i) PA^- is Σ_1-complete when we relativize the hierarchy to inductive predicates, and (ii) any primitive recursive function is provably definable in PA^-. We first make precise the notion of a relativized arithmetical hierarchy.

Definition 51 (Relativized Arithmetical Hierarchy)**.** *Let X be a set definable in $\mathcal{L}$ and T be the base theory in $\mathcal{L}$. Let $\mathcal{L}^+$ be the language $\mathcal{L}$ plus a new predicate X, and T^+ be T plus a defining axiom $\forall x(\mathsf{X}(x) \leftrightarrow X(x))$. A formula is $\Sigma_0(X)$ in T if it is Σ_0 in T^+, analogously for Δ_n, Σ_n and Π_n.*

The usual notions of Σ_n-completeness (or Π_n-completeness), that if $\mathbb{N} \vDash \varphi$ where φ is a Σ_n formula (or Π_n formula) then $T \vdash \varphi$, generalize to the relativized hierarchy. A theory T is $\Sigma_1(X)$-complete iff for any $\Sigma_1(X)$-formula φ, $\mathbb{N} \vDash \varphi$ entails $T \vdash \varphi$.

Lemma 52. *Let J be an inductive predicate, φ any formula and φ' obtained from φ by replacing all occurrences of $J(t)$ in φ by $0 = 0$, then $\mathbb{N} \vDash \varphi$ iff $\mathbb{N} \vDash \varphi'$.*

Proof. By induction on the complexity of formulas. Base case follows since J is inductive, so for all $\bar{a} \in \mathbb{N}$, $\mathbb{N} \vDash J(\bar{a})$. Inductive cases for propositional connectives are trivial. Let $\varphi \equiv \exists x \psi(x, \bar{a})$, where $a \in \mathbb{N}$ is arbitrary. $\mathbb{N} \vDash \varphi$ iff there exists $x \in \mathbb{N}$ where $\mathbb{N} \vDash \psi(x, \bar{a})$, iff (by induction hypothesis) there exists $x \in \mathbb{N}$ where $\mathbb{N} \vDash \psi'(x, \bar{a})$, iff $\mathbb{N} \vDash \psi'(x, \bar{a})$, iff $\mathbb{N} \vDash \exists x \psi(x, \bar{a})$. □

Lemma 53. *Let J be an inductive predicate, and φ be $\Delta_0(J)$ and φ' obtained from φ by replacing all occurrences of $J(t)$ in φ by $0 = 0$, then $\mathsf{PA}^- \vdash \varphi \leftrightarrow \varphi'$.*

Proof. By induction on the complexity of $\Delta_0(J)$ formulas. Base case: the only non-trivial case is $\varphi \equiv J(n)$ where n is a numeral, and $\varphi' \equiv 0 = 0$. Since PA^- proves J inductive, $\mathsf{PA}^- \vdash J(n)$. Therefore $\mathsf{PA}^- \vdash J(n) \leftrightarrow 0 = 0$. Inductive case: all cases of propositional connectives are trivial. Let $\varphi = (\forall v < k)\psi(v)$, and $\varphi' = (\forall v < k)\psi'(v)$.

Note that by Σ_0 completeness of PA^-, $\mathsf{PA}^- \vdash \forall x(x < k \leftrightarrow x = 0 \vee x = 1 \vee ... \vee x = k-1)$, therefore $\mathsf{PA}^- \vdash (\forall v < k)\psi(v) \leftrightarrow \psi(0) \wedge \psi(1) \wedge ... \wedge \psi(k-1)$ by elementary logic. $\mathsf{PA}^- \vdash (\forall v < k)\psi(v)$ iff $\mathsf{PA}^- \vdash \psi(0) \vee \psi(1) \vee ... \vee \psi(k-1)$ iff (by induction hypothesis) $\mathsf{PA}^- \vdash \psi'(0) \vee \psi'(1) \vee ... \vee \psi'(k-1)$ iff $\mathsf{PA}^- \vdash (\forall v < k)\psi'(v)$'. □

Theorem 54. PA^- *is* $\Sigma_1(J)$*-complete for any inductive predicate* J*.*

Proof. Consider $\varphi \equiv \exists x\psi(x)$ a $\Sigma_1(J)$-formula, therefore $\psi(x)$ is $\Delta_0(J)$. Suppose $\mathbb{N} \vDash \exists x\psi(x)$, there is an $n \in \mathbb{N}$ where $\mathbb{N} \vDash \psi(n)$, so $\mathbb{N} \vDash \psi'(n)$. By Δ_0-completeness of PA^- [Kay91], $\mathsf{PA}^- \vdash \psi'(n)$; by Lemma 53 $\mathsf{PA}^- \vdash \psi(n)$, therefore $\mathsf{PA}^- \vdash \exists x\psi(x)$. □

We then show that all primitive recursive functions are expressible in PA^-, i.e., for every primitive recursive function f, there exists a Δ_1 formula $\varphi(\vec{x})$ such that $n \in \Gamma(f)$ iff $\mathbb{N} \vDash \varphi(n)$. We adopt a simplification of the proof that general recursive functions are Δ_1-definable, presented in the preface of Hájek & Pudlák 1993 [HP93].

Lemma 55. *Every* $\Sigma_1(X)$ *function is* $\Delta_1(X)$*.*

Proof. Let $f(\vec{x})$ be a $\Sigma_1(X)$ function. Therefore, there exists a $\Sigma_1(X)$-formula $\varphi(\vec{x}, y)$ where $(\vec{x}, y) \in \Gamma(f)$ iff $\mathbb{N} \vDash \varphi(\vec{x}, y)$. We define the anti-extension of $\Gamma(f)$ by $\exists z(\varphi(\vec{x}, z) \wedge y \neq z)$. □

Lemma 56 (PA^-)**.** *The graphs of initial functions are definable by* $\Sigma_1(\mathsf{Seq})$*-formulas:*

- *Constant Zero function is defined by* $\mathsf{0}$*.*
- *Successor function* $S(x)$ *is defined by* $\varphi(x, y) := y = x + 1$*.*
- *Given* $n, i \in \mathbb{N}$ *as parameters, projection function* $\pi^i_n(\vec{x}, y) := \mathsf{Seq}(\vec{x}) \wedge \ell(x) = n \wedge x_i = y$*.*

Lemma 57 (PA^-)**.** *Functions whose graphs are definable by* $\Sigma_1(\mathsf{Seq})$*-formulas are closed under function composition.*

Proof. It suffices to consider the function composition of two functions f and g. Suppose they are defined by $\Sigma_1(X)$ formulas φ and ψ respectively. Define $\exists y(\varphi(\vec{x}, y) \wedge \psi(y, z))$. □

Lemma 58 (PA^-)**.** *Functions whose graphs are definable by* $\Sigma_1(\mathsf{Seq})$*-formulas are closed under primitive recursion.*

Proof. The idea is that for any $(\vec{x}, n, y)$ where $f(\vec{x}, n) = y$, there is a corresponding computation tree consisting $f(\vec{x}, 0), f(\vec{x}, 1), \ldots, f(\vec{x}, n)$, since to compute $f(\vec{x}, n)$ requires computing each of the previous values. Therefore $f(\vec{x}, n) = y$ iff there exists a computation sequence terminating at $f(\vec{x}, n) = y$.

Consider g whose graph is defined by φ and h defined by ψ. Let

$$f(\vec{x}, 0) = g(\vec{x})$$
$$f(\vec{x}, n+1) = h(f(\vec{x}, n), \vec{x}, n)$$

We define the computation sequence of f of length n with input $\vec{x}$ as

$$\begin{aligned}\mathsf{CompSeq}_{f,n}(\vec{x}, z) :\leftrightarrow \mathsf{Seq}(z) \wedge \ell(z) = n + 1 \\ \wedge \exists ab \leq z\, (\pi_2(z_0) = a \wedge \varphi(\vec{x}, a)) \\ \wedge \forall i < \ell(z)\, \forall j < i\, (i = j + 1 \to \psi(\pi_2(z_j), \pi_1(\pi_1(z_j)), j, \pi_2(z_i))))\end{aligned}$$

Therefore $f(\vec{x}, n) = y :\leftrightarrow \exists z\, (\mathsf{CompSeq}_{f,n}(\vec{x}, z) \wedge \pi_2(z_n) = y)$ is $\Sigma_1(\mathsf{Seq})$. □

Corollary 59. *The graphs of primitive recursive functions are $\Delta_1(\mathsf{Seq})$-definable.*

Corollary 60. *By Theorem 54, PA^- decides every input-output pair $(x, f(x))$ of every primitive recursive function f.*

It is important to bear in mind that Corollary 60 only shows that for every primitive recursive f, there exists some φ where $\mathbb{N} \vDash y = f(x)$ iff $\mathsf{PA}^- \vdash \varphi(x, y)$. $f(x)$ might be undefined when x is non-standard. Therefore one cannot assume f, as defined by $\varphi(x, y)$, is provably total in PA^-. With this in mind, we turn to the definition of subtractive sequences.

Definition 61 (Subtraction of an element from a sequence)**.** *Given $\mathcal{M} \vDash \mathsf{PA}^-$ and $x \in M$, we define subtraction of x from a sequence $s \in \mathsf{Seq}_2^{\mathcal{M}}$, denoted as $s \setminus x$, by the following primitive recursive function:*

$$\begin{aligned}\emptyset \setminus x &= \emptyset \\ \langle y \rangle \setminus x &= \emptyset && \text{if } y = x \\ \langle y \rangle \setminus x &= \langle y \rangle && \text{if } y \neq x \\ (s_1 * s_2) \setminus x &= (s_1 \setminus x) * (s_2 \setminus x)\end{aligned}$$

By Lemma 60, the graph of this function can be defined by some $\Delta_1(\mathsf{Seq}_2)$ formula in PA^-.

Proposition 62 (Subtractive Sequences)**.** *There exists a cut $\mathsf{Seq}_3 \subseteq \mathsf{Seq}$ closed under subtraction.*

Proof. Define $\mathsf{Seq}_3(s) := \mathsf{Seq}(s) \wedge \forall s' \forall a\, (\ell(s') \leq \ell(s) \to s' \setminus a \downarrow \wedge \mathsf{Seq}(s' \setminus a))$.

It is easy to see that $\mathsf{Seq}_3(\emptyset)$. Suppose $\mathsf{Seq}_3(x)$ and $\mathsf{Seq}_3(y)$. Let s' be such that $\ell(s') < \ell(x * y)$ and $a \in M$. Observe that since Seq is a cut, there must exist s_x where $\ell(s_x) \leq \ell(x)$ and s_y where $\ell(s_y) \leq \ell(y)$ such that $s_x * s_y = s'$. Therefore $(s' \setminus a) = (s_x \setminus a) * (s_y \setminus a)$. Since $\mathsf{Seq}_3(x)$ and $\mathsf{Seq}_3(y)$, $s_x \setminus a \downarrow$ and $s_y \setminus a \downarrow$. Therefore $s' \setminus a \downarrow$. Since $s_x \setminus a \in \mathsf{Seq}$ and $s_y \setminus a \in \mathsf{Seq}$, $s_y \setminus a \in \mathsf{Seq}$.

To show Seq_3 closed under subtraction, consider $\mathsf{Seq}_3(s)$ and $s \setminus a$ for arbitrary a. We wish to show that $\mathsf{Seq}_3(s \setminus a)$. We know that $\mathsf{Seq}(s) \wedge \forall s' \forall a (\ell(s') \leq \ell(s) \to s' \setminus a \downarrow \wedge \mathsf{Seq}(s' \setminus a))$. Since $\ell(s) \leq \ell(s)$, $s \setminus a \downarrow$ and $\mathsf{Seq}(s \setminus a)$. Consider arbitrary s' where $\ell(s') \leq \ell(s \setminus a)$. Observe that $\ell(s \setminus a) \leq \ell(s)$, therefore $\ell(s') \leq \ell(s)$. So by $\mathsf{Seq}_3(s)$, $s' \setminus b \downarrow$ and $\mathsf{Seq}_2(s' \setminus b)$ for arbitrary b. □

We rename Seq_3 as Seq.

2.2 Arithmetisation of Syntax

We start by assigning distinct numbers to each vocabulary in $\mathcal{L}_{\mathsf{PA}}$[2]:

$$0, \mathsf{v}, \mathsf{c}, \mathsf{A}, \mathsf{M}, \neg, \vee, \exists$$

We then follow the convention of Feferman 1991 [Fef91] to encode variables and formulas as pairs. For variables v_x, we encode it as

$$\ulcorner \mathsf{v}_x \urcorner = (\mathsf{v}, x)$$

where v is a constant number. Atomic formulas are encoded as a predicate-variable pair.

$$\ulcorner x + y = z \urcorner = (\mathsf{A}, (x, y, z)) \qquad \ulcorner x \times y = z \urcorner = (\mathsf{M}, (x, y, z))$$

Finally, complex formulas are encoded as the pair of their main connective and immediate subformulas:

$$\ulcorner \neg\varphi \urcorner = (\neg, \varphi) \qquad \ulcorner (\varphi \vee \psi) \urcorner = (\vee, (\varphi, \psi)) \qquad \ulcorner \exists v \varphi \urcorner = (\exists, (v, \varphi))$$

We thus define $\mathsf{Var}(x)$, $\mathsf{Atom}(x)$, $\mathsf{Form}(x)$, $x \lhd y$ to be the formulas expressing x is (the code of) a variable, x is (the code of) an atomic formula, x is (the code of) a formula, x is the code of an immediate subformula of the formula encoded by y, respectively.

- $\mathsf{Var}(x) := \exists z < x \; x = (\mathsf{v}, z)$.
- $\mathsf{Atom}(x) := \exists a, b, c < x \; (\mathsf{Var}(a) \wedge \mathsf{Var}(b) \wedge \mathsf{Var}(c) \wedge (x = (\mathsf{A}, (a, b, c))) \vee x = (\mathsf{M}, (a, b, c)))$.

To ensure that the notion of formula is extensionally adequate in PA^-, we use the trick of simultaneously defining the extension and the anti-extension of the set of all formulas.

- $\mathsf{antiAtom}(x) := \neg\mathsf{Atom}(x) \wedge \neg\exists y < x (x = (\neg, y))$
  $\wedge \neg\exists yz < x (x = (\wedge, (y, z)) \wedge \neg\exists v y (\mathsf{Var}(v) \wedge (\forall, (v, y))))$
- $\mathsf{FormSeq}(x) = \mathsf{Seq}(x) \wedge \forall y < x (y \in x \rightarrow \mathsf{Atom}(y)$
  $\vee \exists z (y = (\neg, z) \wedge z <_x y)$
  $\vee \exists a \exists b (y = (\vee, (a, b)) \wedge a <_x y \wedge b <_x y)$
  $\vee \exists v \exists a (y = (\exists, (v, a)) \wedge a <_x y \wedge \mathsf{Var}(v)))$
- $\mathsf{antiFormSeq}(x) = \mathsf{Seq}(x) \wedge \forall y < x (y \in x \rightarrow \mathsf{antiAtom}(y)$
  $\vee \exists z (y = (\neg, z) \wedge z <_x y)$
  $\vee \exists a \exists b (y = (\vee, (a, b)) \wedge a <_x y \wedge b <_x y)$
  $\vee \exists v \exists a (y = (\exists, (v, a)) \wedge a <_x y \wedge \mathsf{Var}(v)))$
- $\mathsf{Form}(x) = \exists y (\mathsf{FormSeq}(y) \wedge x \in y)$.
- $\mathsf{antiForm}(x) = \exists y (\mathsf{antiFormSeq}(y) \wedge x \in y)$.

And finally,

[2] We omit "′" since PA^- has no successor axioms.

- $x \lhd y := \mathsf{Form}(y) \wedge ((y = (\neg, x) \vee \exists z < y(y = (\vee, (x, z))) \vee \exists z < y(y = (\vee, (z, x)))$
$\vee \exists v(\mathsf{Var}(v) \wedge y = (\exists, (v, x))))$.

We then define $y \in \mathsf{FV}(x)$, $\mathsf{Asn}(\alpha)$, $\mathsf{Asn}(\alpha, x)$, $x \in \mathsf{Dom}(y)$, $\mathsf{Sent}(x)$ to be the formulas expressing the following respectively: y is among the free variables of x; α is a code of an assignment; α is the code of an assignment for x; x is in the domain of the assignment α; and finally, x is a code of an sentence. We follow the natural choice of encoding assignments as a finite sequence of pairs $\langle v, a\rangle$ where $v \in \mathsf{Var}$ and a is any object in the model.

- $\mathsf{FVSeq}(x) = \mathsf{Seq}(x) \wedge \forall i < \ell(x)(\exists y \exists v (x)_i = (y, v) \wedge$
$(\mathsf{Atom}(y) \wedge (\ell(v) = 3 \wedge \mathsf{Var}(v_1) \wedge \mathsf{Var}(v_2) \wedge \mathsf{Var}(v_3))$
$\vee \exists j < i(\pi_1((x)_i) \approx (\neg, \pi_1((x)_j)) \wedge \pi_2((x)_i) = \pi_2((x)_j))$
$\vee \exists k, j < i(\pi_1((x)_i) \approx (\vee, (\pi_1((x)_j), \pi_1((x)_k))) \wedge \pi_2((x)_i) = \pi_2((x)_j) * \pi_2((x)_k))$
$\vee \exists j < i \exists v(\mathsf{Var}(v) \wedge \pi_1((x)_i) \approx (\exists, (v, \pi_1((x)_j))) \wedge \pi_2((x)_i) = \pi_2((x)_j) \setminus v)))$
- $y \in \mathsf{FV}(x) := \exists z \exists v \mathsf{FVSeq}(z) \wedge (x, v) \in z \wedge y \in v$.
- $\mathsf{Asn}(\alpha) := \forall i < \ell(\alpha)\, \exists v \exists a\, ((\alpha)_i = \langle v, a\rangle \wedge \mathsf{Var}(v)) \wedge \forall v \forall a \forall b\, (\langle v, a\rangle \in \alpha \wedge \langle v, b\rangle \in \alpha \to a = b)$.
- $y \in \mathsf{Dom}(\alpha) := \mathsf{Asn}(\alpha) \wedge \exists i \pi_1((\alpha)_i) = x$.
- $\mathsf{Asn}(\alpha, x) := (\mathsf{Form}(x) \wedge \mathsf{Asn}(\alpha)) \wedge \forall y (y \in \mathsf{Dom}(\alpha) \leftrightarrow y \in \mathsf{FV}(x))$.
- $\mathsf{Sent}(x) := \mathsf{Form}(x) \wedge \forall v < x\, (\mathsf{Var}(v) \to v \notin \mathsf{FV}(x))$.

It remains to be checked that these definitions of syntactical concepts function as expected. A minimal requirement is that they should be provably definable in PA^-, i.e., PA^- decides positive and negative instances of it. By Theorem 54 and Theorem 60, it suffices to show that these concepts are primitive recursive, or that their extension and anti-extension are expressed by a $\Sigma_1(X)$ formula for some inductive X. We observe that $\mathsf{Var}(x)$, $\mathsf{Atom}(x)$, $\mathsf{Asn}(\alpha)$ and $y \in \mathsf{Dom}(\alpha)$ are all $\Delta_1(\mathsf{Seq})$.

To show that the extension and anti-extension of $\mathsf{Form}(x)$ and $y \in \mathsf{FV}(x)$ are $\Sigma_1(X)$ for some inductive X, we first show that for any sequence x, either $\mathsf{Form}(x)$ or $\mathsf{antiForm}(x)$, but not both. We extend the method of cuts to formulas.

Definition 63 (Form-cut and Form-inductive formula)**.** *Let $T \supseteq \mathsf{PA}^-$ be a theory in $\mathcal{L}_{\mathsf{PA}}$. $F \subseteq \mathsf{Form}$ is Form-inductive in T if*

- $T \vdash \forall x(\mathsf{Atom}(x) \to F(x))$,
- $T \vdash \forall x(F(x) \to F((\neg, x))$,
- $T \vdash \forall x \forall y(F(x) \wedge F(y) \to F((\vee, (x, y)))$,
- $T \vdash \forall x \forall v(F(x) \wedge \mathsf{Var}(v) \to F((\exists, (v, x)))$.

F is a Form-cut in T if in addition F is closed under subformula relation: $T \vdash \forall x(F(x) \to \forall y(\mathsf{FormSeq}(y) \wedge x \in y \to \forall z(z \in y \to F(z))))$

Remark 64 (Every Form-inductive formula can be shortened to a Form-cut)**.** *Let $I(x)$ be Form-inductive in $T \supseteq Q$. Then there is a Form-cut $J(x)$ in T for which $T \vdash \forall x(J(x) \to I(x))$.*

Proposition 65. *There exists a cut $\mathsf{Form}_2(x)$ where all formulas as defined by Form_1 are such that either $\mathsf{Form}(x)$ or $\mathsf{antiForm}(x)$, and there is no sequence x that is both $\mathsf{Form}(x)$ and $\mathsf{antiForm}(x)$.*

Proof. Define $\mathsf{Form}_1(x) = \mathsf{Form}(x) \vee \mathsf{antiForm}(x)$. We prove $\mathsf{Form}_2(x)$ inductive by mimicking the argument on structural induction on sequences as defined before. For atomic formula x, observe that either $\mathsf{Atom}(x)$ or $\mathsf{antiAtom}(x)$ by definition. Now suppose we have $\mathsf{Form}_1(y)$ for some y, we consider x such that $y \lhd x$. For simplicity we only consider the case where $x = (\exists, (a, y))$ with $\mathsf{Var}(a)$. Since $\mathsf{Form}_1(y)$, $\mathsf{Form}(y) \vee \mathsf{antiForm}(y)$. If $\mathsf{Form}(y)$, then there exists $\mathsf{FormSeq}(z)$ where $y \in z$. We define $z' = z * x$. It is easy to see that $\mathsf{FormSeq}(z)$, so $\mathsf{Form}(x)$. The case for $\mathsf{antiForm}(y)$ is similar, so either $\mathsf{Form}(x)$ or $\mathsf{antiForm}(x)$. We can further shorten the cut to that $\mathsf{Form}_2 = \mathsf{Form}_1 \wedge \neg(\mathsf{Form}(x) \wedge \mathsf{antiForm}(x))$ by a similar argument. □

Given Form_2, we can define $\mathsf{Form}_3(x) :\leftrightarrow \mathsf{Form}_2(x) \wedge \mathsf{Form}(x)$ and $\mathsf{antiForm}_2(x) :\leftrightarrow \mathsf{Form}_2(x) \wedge \mathsf{antiForm}(x)$, which defines a partition of Form_2. PA^- thus decides the extension and anti-extension of Form with respect to a cut.

Corollary 66. $\mathsf{Form}(x)$ *and* $\mathsf{antiForm}(x)$ *are provably definable in* PA^-.

We then define a complexity measure on formulas:

Definition 67 (Complexity of formulas)**.** *We define a complexity measure on formulas* $c(\varphi)$ *where*

- *if* $\varphi = (R, (t_1, \ldots, t_n))$, $c(\varphi) = 0$,
- *if* $\varphi = (\neg, \psi)$, $c(\varphi) = c(\psi) + 1$,
- *if* $\varphi = (\vee, (\psi, \chi))$, $c(\varphi) = c(\psi) + c(\chi) + 1$,
- *if* $\varphi = (\exists, (v, \psi))$, $c(\varphi) = c(\psi) + 1$.

Since $c(x)$ *is a primitive recursive definition, its graph is expressible in* PA^-.

Remark 68. *For all* $x, y \in \mathsf{Form}_3$, $x \lhd y$ *entails* $c(x) < c(y)$.

Remark 69. *The exists a* Form*-cut* Form_4 *such that* $\mathsf{PA}^- \vdash \forall x(\mathsf{Form}_4(x) \to c(x)\downarrow)$.

Proof. Define $\mathsf{Form}_4(x) :\leftrightarrow \mathsf{Form}_3(x) \wedge c(x)\downarrow$. For atomic x, by definition $c(x) = 0$, so $c(x)\downarrow$. Suppose $x, y \in \mathsf{Form}_4$, $(\vee, (x, y)) \in \mathsf{Form}_4$ since $c((\vee, (x, y))) = c(x) + c(y) + 1$, and $c(x)\downarrow, c(y)\downarrow$. The other cases are similar. □

For readability, we rename Form_4 as Form.

Proposition 70. $y \in \mathsf{FV}(x)$ *is provably definable in* PA^-.

Proof. $y \in \mathsf{FV}(x)$ is $\Sigma_1(\mathsf{Seq})$. $y \notin \mathsf{FV}(x) := \exists z \exists v < x \mathsf{FVSeq}(z) \wedge (x, v) \in z \wedge y \notin v$ is also $\Sigma_1(\mathsf{Seq})$. □

Corollary 71. *All concepts above are definable, and their basic properties are verifiable in* PA^-.

Provable definability in PA^- ensures that the PA^- definitions of expressions are extensionally adequate, and that it accurately decides which actual sequences of symbols are formulas and which are not. However, extensional adequacy is insufficient for the proper functioning of these definitions in proving theorems about the expressions they are tailored to express. An illustration of the problem is in Halbach 2010 [Hal10]. Suppose we have a "natural" definition of closed terms $\mathrm{ClTerm}(x)$ (which roughly says that x is a term and x contains no free variable symbols), from which we define the "unnatural" definition

$$\mathrm{ClTerm}^*(x) := \mathrm{ClTerm}(x) \wedge \neg B(f(x^\circ), \ulcorner 0 = 1 \urcorner)$$

where $f(x^\circ)$ returns the n-th proof in PA, and $B(x,y)$ iff x is a proof of y. Therefore $\mathrm{ClTerm}^*(x)$ iff x is a closed term and x's proof in PA is not a proof of inconsistency. Since PA is consistent, $\mathrm{ClTerm}^*(x)$ and $\mathrm{ClTerm}(x)$ are extensionally equivalent. Yet $\mathsf{PA} \vdash \forall x \mathrm{ClTerm}(x)$ and $\mathsf{PA} \nvdash \forall x \mathrm{ClTerm}^*(x)$ for Gödelian reasons. PA's inability to prove that $\forall x \mathrm{ClTerm}^*(x)$ may have undesirable effects downstream. Thus, we have to check that PA^- proves all proposed definitions have the desirable "natural" properties of the syntactical expressions they are designed to capture.

Remark 72. *Formulas as specified in* Form *satisfy the usual inductive definition, that if* $x = \ulcorner \varphi \urcorner \in \mathsf{Form}$*, then*

$$\begin{aligned}\mathsf{Atom}(x) \lor \exists y \leq x(\mathsf{Form}(y) \land x = \ulcorner \neg y \urcorner) \\ \lor \exists y \leq x \exists z \leq x(\mathsf{Form}(y) \land \mathsf{Form}(z) \land x = \ulcorner y \to z \urcorner) \\ \lor \exists ... \\ \lor \exists y \leq x \exists v \leq x(\mathsf{Form}(y) \land \mathsf{Var}(v) \land x = \ulcorner \exists v y \urcorner)\end{aligned}$$

One advantage of Feferman's encoding is that we get the unique reading of formulas for free:

Remark 73. *Formulas are uniquely determined by their immediate subformulas and their main connective, i.e., for any formula* $x \in \mathsf{Form}$*, exactly one of the four cases holds:*

1. x *is the code of an atomic formula,*
2. *there exists unique* φ *whose code* $\ulcorner \varphi \urcorner \in \mathsf{Form}$*, where* $x = \ulcorner \neg \varphi \urcorner$*,*
3. *there exists unique* φ, ψ *whose codes* $\ulcorner \varphi \urcorner, \ulcorner \psi \urcorner \in \mathsf{Form}$*, where* $x = \ulcorner (\varphi \lor \psi) \urcorner$*,*
4. *there exists unique* φ *and* v *with* $\ulcorner \varphi \urcorner \in \mathsf{Form}$ *and* $\ulcorner v \urcorner \in \mathsf{Var}$*, where* $x = \ulcorner \exists v \varphi \urcorner$*.*

Remark 74. *The immediate subformula relation* $x \lhd y$ *is irreflexive and asymmetric.*

Proof. It suffices to note that $x \lhd y$ entails $x < y$ by remark 73, and that PA^- proves that $<$ is irreflexive and asymmetric. Consequently, there is no loop involving $x, y_1, ...y_n$ with $n \geq 0$ where $x \lhd y_1 \lhd \cdots \lhd y_n \lhd x$. □

We also check that the usual operations involved in the semantic definition of quantifiers, i.e., resetting a specific variable's value, restricting the variable's domain, are definable.

Remark 75 (Deleting a variable from an assignment)**.** *Consider* $\mathcal{M} \vDash \mathsf{PA}^-$*,* $\varphi \in \mathsf{Form}^{\mathcal{M}}$ *and* α *such that* $\mathsf{Asn}(\alpha, \varphi)$*. For any* $v \in \mathsf{Dom}(\alpha)$*, there exists* $\alpha' \subset \alpha$ *where* $\mathsf{Asn}(\alpha')$*, for all* $t \in \mathsf{FV}(\varphi) \setminus \{v\}$ $\alpha'(t) = \alpha(t)$*, and* $v \notin \mathsf{Dom}(\alpha')$*.*

Proof. $\mathsf{Asn}(\alpha, \varphi)$ entails $\mathsf{Seq}(\alpha)$. By Lemma 62, for any $n \in M$ there exists $s \in \mathsf{Seq}$ where $s \approx \alpha \setminus n$. We know that there exists $n \in M$ where $n \in \alpha$ and $\pi_1(t) = a$, let $\alpha' = \alpha \setminus n$. □

Remark 76 (Resetting a variable in an assignment)**.** *Consider* $\mathcal{M} \vDash \mathsf{PA}^-$*,* $\varphi \in \mathsf{Form}^{\mathcal{M}}$ *and* α *such that* $\mathsf{Asn}(\alpha, \varphi)$*. For any* $v \in \mathsf{Var}^{\mathcal{M}}$ *and* $a \in M$*, there exists* $\alpha[v:a]$ *where* $\mathsf{Asn}(\alpha[v:a], \varphi)$ *and* $\alpha[v:a]$ *differs from* α *only in that* $\alpha[v:a](v) = a$*.*

Proof. By Remark 75, there exists $\alpha' \subset \alpha$ where $v \notin \mathsf{Dom}(\alpha)$. Define $\alpha[v:a]$ as $\alpha * (v, a)$. □

Remark 77 (Restricting the domain of an assignment)**.** *Consider* $\mathcal{M} \vDash \mathsf{PA}^-$*,* $\varphi \in \mathsf{Form}^{\mathcal{M}}$ *and* $\mathsf{Asn}(\alpha)$ *where* $\mathsf{Dom}(\alpha) \supseteq \mathsf{FV}(\varphi)$*. There exists* $\alpha \restriction \mathsf{FV}(\varphi)$ *where* $\mathsf{Dom}(\alpha \restriction \mathsf{FV}(\varphi)) = \mathsf{FV}(\varphi)$ *and* $\forall x \in \mathsf{FV}(\varphi)$*,* $\alpha \restriction \mathsf{FV}(\varphi)(x) = \alpha(x)$*.*

Proof. Remark 75 implies that Asn is closed under subsets. So $\alpha \restriction \mathsf{FV}(\varphi)$ exists. □

2.3 Conservativity of Satisfaction in PA^-

The standard model-theoretic treatment of satisfaction for a theory T formulated in $\mathcal{L}_T$ is via adding a binary satisfaction predicate $\mathsf{S}(x, y)$ and a unary formula predicate $\mathsf{F}(x)$ into the language. Intuitively, $\mathsf{S}(x, y)$ says that the formula coded by x is satisfied by the assignment coded by y, and $\mathsf{F}(x)$ says that x encodes a $\mathcal{L}_T$-formula. The behavior of the satisfaction predicate is studied via the theory T^{FS} (read as "T with full satisfaction") formulated in the enriched language $\mathcal{L}_T \cup \{\mathsf{S}, \mathsf{F}\}$. We focus on the case where T is PA^-.

Definition 78 (Satisfaction Predicate)**.** $\mathsf{PA}^{-\mathsf{FS}} := \mathsf{PA}^- \cup \mathsf{Tarski}(\mathsf{S}, \mathsf{Form})$*, where* $\mathsf{Tarski}(\mathsf{S}, \mathsf{Form})$ *refers to the universal generalizations of the following formulas in* $\mathcal{L}_T \cup \{\mathsf{S}, \mathsf{F}\}$*. Note that* R *is a meta-variable that ranges over relations in* $\mathcal{L}_{\mathsf{PA}}$*, i.e.* $\{\mathsf{A}, \mathsf{M}\}$*.* $t_0 \dots t_n$ *are also meta-variables.*

- $\mathsf{tarski}_0(\mathsf{S}, \mathsf{F}) := (\mathsf{F}(x) \to \mathsf{Form}(x)) \wedge (\mathsf{S}(x, \alpha) \to (\mathsf{F}(x) \wedge \mathsf{Asn}(\alpha, x))) \wedge (y \lhd x \wedge \mathsf{F}(x) \to \mathsf{F}(y))$.
- $\mathsf{tarski}_{1,R}(\mathsf{S}, \mathsf{F}) :=(\mathsf{F}(x) \wedge (x = (R, (t_0 \dots t_{n-1}))) \wedge \mathsf{Asn}(\alpha, x) \wedge \bigwedge_{i<n} \alpha(t_i) = a_i)$
  $\to (\mathsf{S}(x, \alpha) \leftrightarrow R(a_0 \dots a_{n_1}))).$
- $\mathsf{tarski}_2(\mathsf{S}, \mathsf{F}) := (\mathsf{F}(x) \wedge x = (\neg, y) \wedge \mathsf{Asn}(\alpha, x)) \to (\mathsf{S}(x, \alpha) \leftrightarrow \neg\mathsf{S}(y, \alpha))$.
- $\mathsf{tarski}_3(\mathsf{S}, \mathsf{F}) :=(\mathsf{F}(x) \wedge x = (\vee, (y_1, y_2)) \wedge \mathsf{Asn}(\alpha, x))$
  $\to (\mathsf{S}(x, \alpha) \leftrightarrow (\mathsf{S}(y_1, \alpha \restriction \mathsf{FV}(y_1)) \vee \mathsf{S}(y_2, \alpha \restriction \mathsf{FV}(y_2)))))$.
- $\mathsf{tarski}_4(\mathsf{S}, \mathsf{F}) := (\mathsf{F}(x) \wedge x = (\exists, (t, y)) \wedge \mathsf{Asn}(\alpha, x)) \to (\mathsf{S}(x, \alpha) \leftrightarrow \exists \alpha' \supseteq \alpha\, \mathsf{S}(y, \alpha'))$.

Moreover, one usually requires that the satisfaction predicate be defined for all formulas in the model.

Definition 79 (Satisfaction Classes and Full Satisfaction Classes)**.** *Consider a* PA^-*-model* $\mathcal{M}$*,* $F \subseteq M$ *and* S *a binary relation on* M

1. S *is an* F*-satisfaction class if* $(\mathcal{M}, S, F) \vDash \mathsf{Tarski}(\mathsf{S}, \mathsf{F})$.
2. S *is a full satisfaction class if* S *is a* F*-satisfaction class for* $F = \mathsf{Form}^{\mathcal{M}}$.

In addition, we call a satisfaction class *standard* if F only contains the codes of formulas encoded by standard numbers. That is, we consider $\omega_{\mathcal{M}}$ the well-founded initial segment of $\mathcal{M}$ that is isomorphic to the ordinal ω. F is the set of standard $\mathcal{L}_{\mathsf{PA}}$-formula of $\mathcal{M}$ if $F = \mathsf{Form}^{\mathcal{M}} \cap \omega_{\mathcal{M}}$.

With these definitions, we proceed to prove the conservativeness of compositional satisfaction for PA^-. Following [EV15], this is done by first proving the following lemma.

Lemma 80 (PA^-)**.** *Let* $\mathcal{N}_0 \vDash \mathsf{PA}^-$*,* $F_1 := \mathsf{Form}^{\mathcal{N}_0}$*,* $F_0 \subseteq F_1$*, and suppose* S_0 *is an* F_0*-satisfaction class. Then there is an elementary extension* $\mathcal{N}_1$ *of* $\mathcal{N}_0$ *that carries an* F_1*-satisfaction class* $S_1 \supseteq S_0$ *and* $(c, \alpha) \in S_0$ *whenever* $c \in F_0$*,* $\alpha \in N_0$ *and* $(c, \alpha) \in S_1$*.*

For clarity, we establish the proof in steps. The intuitive idea of the proof is to stipulate a set of axioms for the satisfaction class for the elementary extension $\mathcal{N}_1$ of $\mathcal{N}_0$. The content of the set of axioms should exactly match the requirements for the satisfaction class as stipulated in the lemma. The axioms are formulated in an enriched language:

Definition 81 ($\mathcal{L}^+_{\mathsf{PA}}(\mathcal{N}_0)$)**.** *Let* $\mathcal{L}^+_{\mathsf{PA}}(\mathcal{N}_0)$ *as* $\mathcal{L}_{\mathsf{PA}}$ *enriched by constant symbols for each member in* N_0*, and* U_c *for each* $c \in \mathsf{Form}^{\mathcal{N}_0}$.

The intended interpretations for U_c is $\{\alpha \in A_c : S_1(c, \alpha)\}$, where $A_c := \{\alpha : \mathcal{N}_1 \vDash \mathsf{Asn}(\alpha, c)\}$. That is, $\mathsf{U}_c(\alpha)$ iff α is an assignment of c in $\mathcal{N}_1$, and α behaves in a Tarskian way.

Definition 82 ($\mathsf{Th}^+(\mathcal{N}_0)$)**.** *Let* $\mathsf{Th}^+(\mathcal{N}_0) := \mathsf{Th}(\mathcal{N}_0, a)_{a\in N_0} \cup \Theta \cup \Gamma$, *where*

$$\Theta := \{\theta_c : c \in F_1\}$$

with each of θ_c *as follows*

- *If* $R \in \mathcal{L}_{\mathsf{PA}}$ *and* $\mathcal{N}_0 \vDash c = \ulcorner R(t_0, ..., t_{n-1})\urcorner$, *then* $\theta_c := \forall\alpha(\mathsf{U}_c(\alpha) \leftrightarrow \mathsf{Asn}(\alpha, c) \wedge R(\alpha(t_0), ...\alpha(t_{n-1})))$.
- *If* $\mathcal{N}_0 \vDash c = \ulcorner\neg d\urcorner$, *then* $\theta_c = \forall\alpha(\mathsf{U}_c \leftrightarrow \mathsf{Asn}(\alpha, c) \wedge \neg\mathsf{U}_d(\alpha))$.
- *If* $\mathcal{N}_0 \vDash c = \ulcorner d_1 \vee d_2\urcorner$, *then* $\theta_c = \forall\alpha(\mathsf{U}_c \leftrightarrow \mathsf{Asn}(\alpha, c) \wedge (\mathsf{U}_{d_1}(\alpha \restriction \mathsf{FV}(d_1)) \vee \mathsf{U}_{d_2}(\alpha \restriction \mathsf{FV}(d_2))))$.
- *If* $\mathcal{N}_0 \vDash c = \ulcorner\exists v_a b\urcorner$, *then* $\theta_c = \forall\alpha(\mathsf{U}_c \leftrightarrow \mathsf{Asn}(\alpha, c) \wedge \exists\alpha' \supseteq \alpha\, \mathsf{U}_b(\alpha') \wedge \mathsf{Asn}(\alpha', b))$.

And

$$\Gamma := \{\mathsf{U}_c(\alpha) : c \in F_0, (c, \alpha) \in S_0\} \cup \{\neg\mathsf{U}_c(\alpha) : c \in F_0, (c, \alpha) \notin S_0\}.$$

The intuitive interpretation of Θ is that each of U_c contains assignments α for c that follow the Tarskian definitions for assignments for formulas. And Γ ensures that the other requirement for the F_1-satisfaction class S_1, namely, that $(c, \alpha) \in S_0$ whenever $c \in F_0$, $\alpha \in N_0$ and $(c, \alpha) \in S_1$, is satisfied.

Remark 83. $\mathsf{Th}^+(\mathcal{N}_0)$ *is syntactically well-defined in* PA^-.

Proof. By Remark 77, $\mathsf{U}_x(\alpha \restriction \mathsf{FV}(x))$ is well-defined. □

Before proving Lemma 80, we first show that $\mathsf{Th}^+(\mathcal{N}_0)$ is consistent.

Lemma 84 (PA^-)**.** $\mathsf{Th}^+(\mathcal{N}_0)$ *has a model.*

Proof. By compactness, it suffices to show that arbitrary finite subsets $T_0 \subseteq \mathsf{Th}^+(\mathcal{N}_0)$ are interpretable in PA^-. Since $\mathsf{Th}(\mathcal{N}_0, a)_{a\in N_0}$ is known to be consistent, we only have to check the T_0s where $T_0 \cap (\Theta \cup \Gamma) \neq \emptyset$. This is done by showing that each of the U_cs has an extension that satisfies the axioms stipulated in $\Theta \cup \Gamma$.

Let $C = c \in \mathsf{Form}^{\mathcal{N}_0}$, we aim to construct subsets $\{U_c : c \in C\}$ of N_0, where the following conditions hold when U_c is interpreted by U_c:

1. U_c behaves in a Tarskian way: $(\mathcal{N}_0, U_c)_{c\in C} \vDash \{\theta_c : c \in C\}$,
2. The behavior of U_c respects that of $\mathcal{N}_0$: for $c \in C \cap F_0$, $U_c = \{\alpha \in N_0 : (c, \alpha) \in S_0\}$.

The construction of U_c is done by induction on the complexity of formulas. First, define $\lhd^*$ on C where

$$c \lhd^* d \text{ iff } (c \lhd d)^{\mathcal{N}_0} \text{ and } \theta_d \in T_0 \cap \Theta.$$

Since C is finite and $\lhd^*$ is loop-free by Remark 74, $(C, \lhd^*)$ is well-founded. We thus define rank_C for $c \in C$ as

$$\mathsf{rank}_C(c) := \sup\{\mathsf{rank}_C(d) + 1 : d \in C \text{ and } d \lhd^* c\}$$

and

$$C_i := \{c \in C : \mathsf{rank}_C(c) \leq i\}.$$

Observe that $\mathsf{rank}_C(c) = 0$ precisely when $\theta_c \notin T_0 \cap \Theta$, moreover, if $c \in C_k$, all codes d of immediate subformulas of the formula coded by c are in C_{k-1}. This allows us to define the extension of U_c for each $c \in C$ by the following recursive definition:

- If $c \in C_0$, then $U_c := \begin{cases} \{\alpha : (c, \alpha) \in S_0\}, \text{ if } c \in F_0; \\ U_c := \emptyset \text{ if } c \notin F_0. \end{cases}$
- If $c \in C_{i+1} \setminus C_i$ and $c = \neg d$, then $U_c := \{\alpha \in A_c : \alpha \notin U_d\}$.
- If $c \in C_{i+1} \setminus C_i$ and $c = a \vee b$, then $U_c := \{\alpha \in A_c : \alpha \restriction \mathsf{FV}(a) \in U_a \text{ or } \alpha \restriction \mathsf{FV}(b) \in U_b\}$.
- If $c \in C_{i+1} \setminus C_i$ and $c = \exists v_a b$, then $U_c := \{\alpha \in A_c : \exists \alpha' \in N (\alpha \subseteq \alpha' \text{ and } \alpha' \in U_b)\}$.

Notice that the last clause is well-defined since by Remark 76, for all α and φ such that $\mathsf{Asn}(\alpha, \varphi)$, for any $v \in \mathsf{Dom}(\alpha)$ arbitrary $a \in M$, $\alpha[v : a]$ is well-defined.

We prove by induction to show that conditions 1) and 2) are satisfied when U_cs are interpreted by U_c. For the base case $c \in C_0$, 1) is vacuously satisfied and 2) is satisfied by definition. For the inductive case, consider for example $c = a \vee b$, and suppose that U_a and U_b satisfies both 1) and 2). $U_c := \{\alpha \in A_c : \alpha \restriction \mathsf{FV}(a) \in U_a \text{ or } \alpha \restriction \mathsf{FV}(b) \in U_b\}$, so $(\mathcal{N}_0, U_c)_{c \in C} \vDash \theta_c$. 2) is satisfied because S_0 is an N_0 satisfaction class. Therefore every finite subset of $\mathsf{Th}^+(\mathcal{N}_0)$ has a model, so $\mathsf{Th}^+(\mathcal{N}_0)$ has a model. □

We finish the proof of Lemma 80 by showing that the model of $\mathsf{Th}^+(\mathcal{N}_0)$ is an elementary extension of $\mathcal{N}_0$ with an F_1-satisfaction class S_1 with the properties stipulated in Lemma 80.

Proof. Recall that $\mathsf{Th}^+(\mathcal{N}_0) := \mathsf{Th}(\mathcal{N}_0, a)_{a \in N_0} \cup \Theta \cup \Gamma$. So any model of $\mathsf{Th}^+(\mathcal{N}_0)$ is of the shape $(\mathcal{N}_1, U_c)_{c \in F_1}$, where $\mathcal{N}_1 \vDash \mathsf{Th}(\mathcal{N}_0, a)_{a \in N_0}$. So $\mathcal{N}_1$ is an elementary extension of $\mathcal{N}_0$. Let S_1 be the binary relation defined on N_1 via

$$S_1(c, \alpha) \text{ iff } \alpha \in U_c.$$

Since condition 1) in Lemma 84 is met, S_1 is an F_1 satisfaction class. Also $S_1 \supseteq S_0$ and $(c, \alpha) \in S_0$ whenever $c \in F_0$, $\alpha \in N_0$ and $(c, \alpha) \in S_1$ by condition 2. □

Lemma 85. *Let M_0 be a model of PA^- of any cardinality.*

1. *If S_0 is an F_0-satisfaction class on M_0, then there is an elementary extension M of M_0 that carries a full satisfaction class that extends S_0.*
2. *There is an elementary extension M of M_0 that carries a full satisfaction class.*

Proof. By Lemma 80, there is an elementary extension $\mathcal{M}_1$ of $\mathcal{M}_0$ that carries a full satisfaction class that extends S_0, where $F_1 := \mathsf{Form}^{\mathcal{M}_0}$. Carrying out the argument for $i : i \in \omega$ yields sequences $\langle \mathcal{M}_i : i \in \omega \rangle$ and $\langle S_i : i \in \omega \rangle$ where

- $\mathcal{M}_{i+1}$ elementary extends $\mathcal{M}_i$,
- S_{i+1} is an F_{i+1} satisfaction class on $\mathcal{M}_{i+1}$ with $F_{i+1} := \mathsf{Form}^{\mathcal{M}_i}$,
- $S_i = S_{i+1} \cap \{(c, \alpha) : c \in F_i, \mathcal{M}_i \vDash \mathsf{Asn}(\alpha, c)\}$.

Let $\mathcal{M} := \bigcup_{i\in\omega} \mathcal{M}_i$ and $S := \bigcup_{i\in\omega} S_i$. By Tarski's elementary chain theorem, $\mathcal{M}$ elementarily extends $\mathcal{M}_0$. S is a full satisfaction class on $\mathcal{M}$. □

Corollary 86. $\mathsf{PA}^{-\mathsf{FS}}$ *is a conservative extension of* PA^-.

Proof. Suppose not, i.e. there exists a formula φ where $\mathsf{PA}^{-\mathsf{FS}} \vdash \varphi$ but $\mathsf{PA}^- \nvdash \varphi$. Therefore $\mathsf{PA}^- \cup \{\neg\varphi\}$ is consistent, there exists $\mathcal{M}_0 \vDash \mathsf{PA}^- \cup \{\neg\varphi\}$. By Theorem 85, there is an elementary extension $\mathcal{M}_1$ of $\mathcal{M}_0$ that carries a full satisfaction class. Therefore $\mathcal{M}_1 \vDash \varphi$. Since $\mathcal{M}_1$ is an elementary extension of $\mathcal{M}_0$, $\mathcal{M}_1 \vDash \varphi$, contradiction. □

2.4 Conservativity of Truth in PA^-

We then turn to showing the conservativity of truth in PA^-. As mentioned in the preliminary section, we will define truth by structural recursion on closed sentences directly, instead of treating it as a special case of satisfaction. The alleged difficulty that recursion cannot handle the quantifier case $\exists x\varphi(x)$ — as $\varphi(x)$, being an open formula, cannot be true — is resolved by substituting the variables in the sentence $\varphi(x)$ directly with a numeral:

$$x = (\exists, (v, y)) \to (\mathsf{T}(x) \leftrightarrow \exists z \mathsf{T}(y(\bar{z})))$$

where $\bar{z}$ is the numeral for the number z. Numerals are usually implemented as sequences of symbols, where the number the numeral denotes is usually a function of the length of the sequence, e.g., repeated successor operation on constant $\mathsf{0}$. Therefore the numeral for number 0 is $\mathsf{0}$, and the numeral for n is $\mathsf{0}\underbrace{'\cdots'}_{n}$.

However, implementing numerals as sequences does not coordinate well with the shortening of cuts. For an arbitrary model $\mathcal{M}$ and a Seq-cut $I \subseteq M$, since the cut sets a constraint on the length of sequences, it is not guaranteed that for any element $x \in I$, the corresponding numeral $\bar{x}$ is in I. So the truth clause for quantifiers might miss out on elements in the domain. So the quantifier cases should be implemented in PA^- in some other ways. A method compatible with the shortening of cuts is domain constants[3].

Definition 87. *Let* $\mathcal{M} \vDash \mathsf{PA}^-$. *We introduce* $\mathsf{c}_x \notin \mathcal{L}_{\mathsf{PA}}$ *as domain constants for* $\mathcal{M}$. *The intended interpretation is that for arbitrary* $\mathcal{M} \vDash \mathsf{PA}^-$ *and* $x \in M$. $(\mathsf{c}_x)^{\mathcal{M}} = x$. *Define* $\ulcorner \mathsf{c}_x \urcorner = (\mathsf{c}, x)$ *where* c *is a constant number distinct from* v, *etc.*

Substitution of domain constants is equivalent to substitution of numerals: for $n \in \mathbb{N}$ and $\varphi(v)$ with v free, $\mathcal{M} \vDash \varphi(\bar{n})$ iff $\mathcal{M} \vDash \varphi(\mathsf{c}_n)$. They are also compatible with any cut. Consider a Seq-cut I in PA^-. For any model $\mathcal{M}$ and $x \in I^{\mathcal{M}}$, there exists $s \in I$ where for some $i \leq \ell(I)$, $(s)_i = \mathsf{c}_x$. We adapt the notion of formula to include domain constants, which gives

- $\mathsf{Const}(x) := \exists z < x\, x = (\mathsf{c}, z)$.
- $\mathsf{Term}(x) := \mathsf{Var}(x) \vee \mathsf{Const}(x)$.
- $\mathsf{Atom}(x) := \exists a, b, c < x\, (\mathsf{Term}(a) \wedge \mathsf{Term}(b) \wedge \mathsf{Term}(c) \wedge (x = (\mathsf{A}, (a, b, c))) \vee x = (\mathsf{M}, (a, b, c)))$.

It is clear that this work, since domain constants are in our treatment of the syntax, is fully analogous to variables.

[3]This illuminates another difference between pairs and sequences. The sequences available will be constrained given a cut, but since the elements of sequences may be anything, we still have access to all the pairs as elements of the sequence.

Definition 88 (Truth class)**.** $\mathsf{PA}^{-\mathsf{T}} := \mathsf{PA}^- \cup \mathsf{Tarski}(\mathsf{T})$*, where* $\mathsf{Tarski}(\mathsf{T})$ *refers to the universal generalizations of the following formulas in* $\mathcal{L}_T \cup \{\mathsf{T}\}$*, where* $\mathsf{Sent}(x)$ *is defined above as expressing "*x *is a* $\mathcal{L}_{\mathsf{PA}}$*-formula with no free variables."*

- $\mathsf{tarski}_0(\mathsf{T}) := \mathsf{T}(x) \to \mathsf{Sent}(x)$.
- $\mathsf{tarski}_{1,R}(x) := x = (R, (\mathsf{c}_{y_0}, \ldots, c_{y_{n-1}})) \to (R(y_0, \ldots, y_{n-1}) \leftrightarrow T(x))$.
- $\mathsf{tarski}_2(x) := x = (\neg, y) \to (\mathsf{T}(x) \leftrightarrow \neg\mathsf{T}(y))$.
- $\mathsf{tarski}_3(x) := x = (\vee, (y_1, y_2)) \to (\mathsf{T}(x) \leftrightarrow \mathsf{T}(y_1) \vee \mathsf{T}(y_2))$.
- $\mathsf{tarski}_4(x) := x = (\exists, (v, y)) \wedge \mathsf{Var}(v) \to (\mathsf{T}(x) \leftrightarrow \exists z \mathsf{T}(y(\mathsf{c}_z)))$.

Enayat & Visser do not prove the conservativity of truth classes directly, but rely on an equivalence between truth and certain satisfaction classes. The idea is as follows. Recall that in the preliminary section, we introduced the conceptual connection between satisfaction and truth. The pair (φ, α) is in the extension of satisfaction, iff φ, with all its free variables replaced by the corresponding constants for values they are assigned, is true. Therefore, the information in a truth class can be represented by a satisfaction class, if whenever the sentence φ is in T, all formula-assignment pairs ψ, α, where one can obtain φ by a suitable substitution of variables from ψ, α, is in S.

Note the emphasis on "all formula-assignment pairs". The transformation of replacing variables by constants is not a bijection: different formulas that are the same except for their free variables ($Rvvu$ and $Ruww$), when paired with suitable assignments ($\alpha(v) = \alpha(u) = \beta(u) = \beta(w) = a$), would be mapped to the same closed sentence ($Raaa$). Only some well-behaved satisfaction classes represent truth classes, namely, those where either all formula-assignment pairs mapping to the same sentence all fall into the extension of the satisfaction predicate, or all of them fall into the *anti*-extension of the satisfaction predicate. We dub these well-behaved satisfaction classes "extensional" — in the sense that they ignore the intensional difference that sentences have different free variables, *a fortiori*, different meanings.

One might wonder why the existing axioms for satisfaction classes are insufficient to enforce extensionality. After all, it seems that for all atomic formulas $(R, (t_0, \ldots, t_n))$, the axiom

$$(\mathsf{F}(x) \wedge x = (R, (t_0 \ldots t_{n-1})) \wedge \mathsf{Asn}(\alpha, x) \wedge \bigwedge_{i<n} \alpha(t_i) = a_i) \to (\mathsf{S}(x, \alpha) \leftrightarrow R(a_0 \ldots a_{n_1})))$$

ensures extensionality for all atomic formulas, and axioms tarski_2 to tarski_4 carry the property forward to all formulas. The culprits are non-standard numbers and formulas.

Example 89. *Let* $\mathcal{M}$ *be a non-standard model of* PA^-*, and* c *a non-standard number. Consider the formulas*

$$\varphi_1 = ((((x_0 = x_0) \wedge x_1 = x_1) \wedge x_2 = x_2) \wedge \cdots \wedge x_c = x_c)$$
$$\varphi_2 = ((((x_c = x_c) \wedge x_{c-1} = x_{c-1}) \wedge x_{c-2} = x_{c-2}) \wedge \cdots \wedge x_0 = x_0)$$

where all the x_i *are variables. Let* α *be the assignment where* $\alpha(x_i) = a \in M$ *for all* i*. Then there exists a full satisfaction class* S *on* $\mathcal{M}$ *where* $(\varphi_1, \alpha) \in S$ *but* $(\varphi_2, \alpha) \notin S$*.*

Thus, one has to manually restrict attention to extensional satisfaction classes. We start by articulating the idea of substitution and replacing the free variables in a formula with the corresponding constants.

Definition 90 (Substitution)**.** *A substitution for a formula ψ of $\mathcal{L}_{\mathsf{PA}}$ is a function*

$$\sigma : \mathsf{FV}(\psi) \to \mathsf{Var}$$

We require σ to preserve substutivity: for $x \in \mathsf{FV}(\psi)$, x is not in the scope of any quantifier that binds $\sigma(x)$.

Notation-wise, define $\psi * \sigma$ as the formula obtained from ψ by applying the substitution σ.

Definition 91 (Replacing variables with domain constants)**.** *Let $\mathcal{M} \vDash \mathsf{PA}^-$, we define $\mathsf{c} : m \mapsto \mathsf{c}_m$ for every $m \in M$ as the injection that maps every element in the domain to the corresponding domain constant. For a formula $\varphi \in \mathsf{Form}^{\mathcal{M}}$ and $\mathcal{M}$-internal assignment α such that $\mathcal{M} \vDash \mathsf{Asn}(\alpha, \varphi)$, the $\mathcal{L}_{\mathsf{PA}}$-sentence $\varphi(\mathsf{c} \circ \alpha)$ is obtained by replacing each occurrence of free variables $v \in \mathsf{FV}(\varphi)$ with the domain constant c_a where $\alpha(v) = a$.*

Definition 92 (Extensional Satisfaction Classes)**.** *To define extensional satisfaction classes, we first define an equivalence relation on formula-assignment pairs (φ, α) where α is a variable assignment for φ. We define $(\varphi_1, \alpha_1) \sim (\varphi_2, \alpha_2)$ iff there is a formula-assignment pair (ψ, β) and substitutions σ_1, σ_2, where*

$$\varphi_1 = \psi * \sigma_1,\ \beta = \alpha_1 \circ \sigma_1;$$
$$\varphi_2 = \psi * \sigma_2,\ \beta = \alpha_2 \circ \sigma_2.$$

where $\circ$ denotes function composition.

An F-satisfaction class S is extensional if for all φ_1 and φ_2 in F, $\mathcal{M} \vDash (\varphi_1, \alpha_1) \sim (\varphi_2, \alpha_2)$ implies $(\varphi_1, \alpha_1) \in S$ iff $(\varphi_2, \alpha_2) \in S$.

Note that PA^- proves that $(\varphi_1, \alpha_1) \sim (\varphi_2, \alpha_2)$ is an equivalence relation (in some cut). We also observe that

Remark 93. *$\sim$ has the following property when interpreted in PA^-:*

- *if $((\neg, \varphi_0), \alpha_0) \sim ((\neg, \varphi_1), \alpha_1)$, then $(\varphi_0, \alpha_0) \sim (\varphi_1, \alpha_1)$.*
- *if $(\vee, (\varphi_0, \varphi_1)), \alpha) \sim ((\vee, (\varphi_0', \varphi_1')), \alpha')$, then $(\varphi_0, \alpha \restriction \mathsf{FV}(\varphi_0)) \sim (\varphi_0', \alpha' \restriction \mathsf{FV}(\varphi_0'))$ and $(\varphi_1, \alpha \restriction \mathsf{FV}(\varphi_1)) \sim (\varphi_1', \alpha' \restriction \mathsf{FV}(\varphi_1'))$.*
- *if $\varphi = (\exists, (t, \psi))$ and $\varphi' = (\exists, (t', \psi'))$ and $(\varphi, \alpha) \sim (\varphi', \alpha')$, then $t = t'$ and for some e, $(\varphi, \alpha[t : e]) \sim (\varphi', \alpha'[t : e])$.*
- *if $(\varphi, \alpha) \sim (\psi, \beta)$, then $c(\varphi) = c(\psi)$. (where c is defined in definition 67)*

The full proof of the remark will be left to the appendix. Given the precise definition of extensionality, the idea that extensional satisfaction classes correspond to truth classes can be made precise as the following proposition:

Proposition 94. *Suppose $\mathcal{M} \vDash \mathsf{PA}^-$, T is a full truth class on $\mathcal{M}$, and S is an extensional full satisfaction class on $\mathcal{M}$. Define*

- *$S(T)$ is an extensional satisfaction class on $\mathcal{M}$, where $S(T)$ is defined as the collection of ordered pairs (φ, α) such that $\varphi(\mathsf{c} \circ \alpha) \in T$.*
- *$T(S)$ is a truth class on $\mathcal{M}$, where $T(S)$ is defined as the collection of $\varphi \in \mathcal{L}_{\mathsf{PA}}^+$ such that for some $\psi \in \mathcal{L}_{\mathsf{PA}}^+$ and some assignment α for ψ, $\varphi = \psi(c \circ \alpha)$ and $(\psi, \alpha) \in S$.*
- *$S(T(S)) = S$, and $T(S(T)) = T$.*

whose proof is also left to the appendix.

Given Proposition 94, the conservativeness of compositional truth for PA^- follows from the conservativeness of extensional satisfaction for PA^-. The idea is to extend the proof of Lemma 80 to stipulate the theory constructed to be extensional by adding a new set of axioms.

Definition 95. *Let φ_0 and φ_1 be $\mathcal{L}_{\mathsf{PA}}$ formulas. We write $\varphi_0 \approx \varphi_1$ if there is a formula ψ, and substitutions σ_0 and σ_1 where $\varphi_i = \psi * \sigma_i$ for $i = 0, 1$.*

Definition 96 (Externally defined subformula-closed set)**.** *Let $\mathcal{M}$ be a PA^--model. Let $c \in \mathsf{Form}^{\mathcal{M}}$. Let $\mathbf{TC}_{\mathcal{M}}(c)$ be the externally defined subformula-closed set with respect to c, i.e.*

$$\mathbf{TC}_{\mathcal{M}}(c) := \bigcup_{n<\omega} \mathbf{TC}_{\mathcal{M}}(n)$$

where $\mathbf{TC}_{\mathcal{M}}(c, 0) := \{c\}$ and $\mathbf{TC}_{\mathcal{M}}(c, n+1) := \{x \in M : x \lhd^{\mathcal{M}} d \text{ for some } d \in \mathbf{TC}_{\mathcal{M}}(c, n)\}$.

Remark 97. *The following holds for $\sim$, $\approx$ and $\mathbf{TC}_{\mathcal{M}}$:*

- *$(\varphi_0, \alpha_0) \sim (\varphi_1, \alpha_1)$ implies $\varphi_0 \approx \varphi_1$.*
- *if $d \in \mathbf{TC}_{\mathcal{M}}(c)$ and $d \neq c$, then $c \not\approx d$.*
- *$\approx$ preserves principal connectives: $(\neg, c) \approx (\neg, d)$ implies $c \approx d$, $(\vee, (c_1, c_2)) \approx (\vee, (d_1, d_2))$ implies $c_1 \approx d_1$ and $c_2 \approx d_2$, finally $(\exists, (t, c)) \approx (\exists, (t', d))$ implies $t = t'$ and $c \approx d$.*
- *if $\varphi \approx \psi$, $c(\varphi) = c(\psi)$. (where $c(\varphi)$ is the complexity of formula defined in Definition 67)*

Proof. i) is obvious, ii) and iv) follow from trivial inductions. iii) follow from Remark 93 and i). □

We then proceed to prove the following important lemma for truth:

Lemma 98. *Let $\mathcal{N}_0 \vDash \mathsf{PA}^-$, $F_1 := \mathsf{Form}^{\mathcal{N}_0}$, $F_0 \subseteq F_1$, and suppose S_0 is an extensional F_0-satisfaction class. Then there is an elementary extension $\mathcal{N}_1$ of $\mathcal{N}_0$ that carries an extensional F_1-satisfaction class $S_1 \supseteq S_0$ and $(c, \alpha) \in S_0$ whenever $c \in F_0$, $\alpha \in N_0$ and $(c, \alpha) \in S_1$.*

The proof strategy is similar to proving the conservativeness of satisfaction classes, with the extra axioms for stipulating that the satisfaction class is extensional. Again, we start by defining $\mathsf{Th}^+(\mathcal{N}_0)$.

Definition 99. *Let $\mathsf{Th}^+(\mathcal{N}_0) := \mathsf{Th}(\mathcal{N}_0, a)_{a \in N_0} \cup \Theta \cup \Gamma$, and Θ and Γ be exactly as in Definition 82. Further define $\Delta := \{\delta_{cc'} : c, c' \in F_1\}$, where*

$$\delta_{cc'} := \forall\alpha\forall\alpha'((c, \alpha) \sim (c', \alpha') \to (\mathsf{U}_c(\alpha) \leftrightarrow \mathsf{U}_{c'}(\alpha'))).$$

$\delta_{cc'}$ ensures any $\mathcal{M} \vDash \mathsf{Th}^+(\mathcal{N}_0)$ carries a satisfaction class S where if $(c, \alpha) \sim (c', \alpha')$, then $(c, \alpha) \in S$ iff $(c', \alpha') \in S$.

Lemma 100. *$\mathsf{Th}^+(\mathcal{N}_0)$ has a model.*

Proof. We show that every finite subset of $\mathsf{Th}^+(\mathcal{N}_0)$ is interpretable in some $\mathcal{N}$. Let C be the collection of $c \in F_1$ such that c appears in T_0. We define $\lhd^*$ and $\mathsf{rank}_C(c)$ exactly as in Lemma 80. We construct models for arbitrary finite subsets $T_0 \subseteq \mathsf{Th}^+(\mathcal{N}_0)$. To ensure extensionality, we first extend C to finite $\bar{C}$ where whenever $c \approx c'$ and $d \lhd^* c$ with $c, c', d \in \bar{C}$, then there is some $d' \in \bar{C}$ where $d' \lhd^* c'$ with $d' \approx d$.

Define $d' \lhd^\circ c$ iff $d' \lhd^* c' \approx c$ for some $c' \in C$. Observe that $\lhd^\circ$ is loop-free in PA^-. Suppose, for a contradiction, that there exists φ, ψ where $\varphi \lhd^\circ \psi$ and $\psi \lhd^\circ \varphi$. There is φ', ψ' where $\varphi \lhd^* \psi'$, $\psi \lhd^* \varphi'$ and $\varphi \approx \varphi'$ and $\psi \approx \psi'$. By Remark 97, $c(\varphi) = c(\varphi')$ and $c(\psi) = c(\psi')$. By Remark 68 and the definition of $\lhd^*$, $c(\varphi) < c(\psi')$ and $c(\psi) < c(\varphi')$, a contradiction. The well-foundedness of $(\lhd^\circ, C)$ leads to the definition of rank°_C for $c \in C$ as

$$\mathsf{rank}^\circ_C(c) := \sup\{\mathsf{rank}^\circ_C(d) + 1 : d \in C \text{ and } d \lhd^\circ c\}$$

which then allows us to define $D_i := \{c \in C : \mathsf{rank}^\circ_C(c) = i\}$. Let $n = \max\{\mathsf{rank}^\circ_C(c) : c \in C\}$; we can recursively construct $E_n, E_{n-1}, \ldots, E_0$ as follows:

- $E_n := D_n$,
- $E_{n-(i+1)} := D_{n-(i+1)} \cup \{d : d \lhd^{\mathcal{N}_0} c \text{ for some } c \in E_{n-i}\}$.

Finally, let $\bar{C} := E_n \cup \cdots \cup E_0$. By replacing C with $\bar{C}$, we observe that if c and c' are both in C with $c \approx c'$, then $\mathsf{rank}_C(c) = \mathsf{rank}_C(c')$.

Similar to the proof of Lemma 80, we construct $\{U_c : c \in C\}$ where the following conditions hold when U_c is interpreted by U_c:

1. U_c behaves in a Tarskian way: $(\mathcal{N}_0, U_c)_{c \in C} \vDash \{\theta_c : c \in C\}$,
2. The behavior of U_c respects that of $\mathcal{N}_0$: for $c \in C \cap F_0$, $U_c = \{\alpha \in N_0 : (c, \alpha) \in S_0\}$,
3. S is extensional: $(\mathcal{N}_0, U_c)_{c \in C} \vDash \{\delta_{cc'} : c, c' \in C\}$.

The proof that the first two conditions hold is exactly like before. To establish 3), we use induction on $\mathsf{rank}_C(c)$ to show that $\forall c \in C\, P(c)$, where

$$P(c) := \forall c' \in C, (\mathcal{N}_0, U_c)_{c \in C} \vDash \forall\alpha\forall\alpha'((c, \alpha) \sim (c', \alpha') \to (\mathsf{U}_c(\alpha) \leftrightarrow \mathsf{U}_{c'}(\alpha'))).$$

For the base case, consider c such that $\mathsf{rank}_C(c) = 0$ and for some c', α, α' $(c, \alpha) \sim (c', \alpha')$. By remark 97, $c \approx c'$, so $\mathsf{rank}_C(c) = \mathsf{rank}_C(c') = 0$. Recall the definition for U_c for $c \in C_0$, where if $c \in F_0$ then $U_c = \{\alpha : (c, \alpha) \in S_0\}$, and $U_c = \emptyset$ otherwise. Since S_0 is an extensional satisfaction class,$(c, \alpha) \in S_0$ iff $(c', \alpha') \in S_0$, so $U_c(\alpha)$ iff $U_c(\alpha')$.

For the inductive case, suppose $P(c)$ holds for $\mathsf{rank}_C(c) = k$. Consider the case of c where $\mathsf{rank}_C(c) = k + 1$. For simplicity, we only consider the case where $c = (\exists, (t, d))$. Suppose for some c', α', $(c, \alpha) \sim (c', \alpha')$. Therefore $c \approx c'$, and that c' must be of the shape $(\exists, (t, d'))$ where $d \approx d'$. Since C is such that for all $c, c' \in C$ where $c \approx c'$, then $\mathsf{rank}_C(c) = \mathsf{rank}_C(c')$. $\mathsf{rank}_C(d) = \mathsf{rank}_C(d') = k$. Applying the inductive hypothesis to d and d' gives $\forall\alpha\forall\alpha'((d, \alpha) \sim (d', \alpha') \to (\mathsf{U}_d(\alpha) \leftrightarrow \mathsf{U}_{d'}(\alpha')))$. We also know that if $\alpha \in U_d$, then $\alpha[t : e] \in U_c$ for some e. By Remark 93, $(c, \alpha[t : e]) \sim (c', \alpha'[t : e])$. Since S is extensional, $\alpha[t : e] \in U_c$ iff $\alpha'[t : e] \in U_{c'}$. The proof of the other direction is similar.

Therefore, we have a set of U_cs that satisfy all three conditions, so $\mathsf{Th}^+(\mathcal{N}_0)$ is consistent. □

The rest proceeds exactly as in the proof of Lemma 80.

Theorem 101. *Let $\mathcal{M}_0 \vDash \mathsf{PA}^-$. There is an elementary extension $\mathcal{M}$ of $\mathcal{M}_0$ that carries a full extensional satisfaction class.*

Corollary 102. *Every model of PA^- has an elementary extension that carries a full truth class.*

Corollary 103. $\mathsf{PA}^{-\mathsf{FS}}$ *is a conservative extension of* PA^{-}.

The conservativeness of compositional truth for PA^{-} should not be surprising, as it is known that compositional truth without extended induction for PA is conservative over PA.

Chapter 3

Compositional Truth for PA^- is not Semantically Conservative.

3.1 Motivating Remarks

In the previous chapter, we proved that PA^- with compositional truth $\mathsf{CT}[\mathsf{PA}^-]$ is syntactically conservative over PA^-, i.e., adding compositional truth does not allow one to prove new theorems in the arithmetical language. This chapter turns to the second question drafted in the introduction:

> Are there any other notions of conservativeness? What is the behavior of compositional truth in PA^- with respect to these other notions?

A natural counterpart of syntactic conservativity is semantic conservativity — the property that every model of the base theory can be expanded into a model of the expanded theory. Since semantic conservativity is underdiscussed in the philosophical literature, we will start by motivating why it is important for axiomatic theories of truth and deflationism, postponing the formal setup for studying non-standard models of arithmetic to later sections.

The notion of semantic conservativity is related to the existence of non-standard models. By the Löwenheim-Skolem theorem, any consistent set of first-order axioms with an infinite model has more than one model (modulo isomorphism). That is, besides the "standard model" that describes the subject matter the axioms intend to capture, there are so-called "non-standard models" that are not isomorphic to the standard model, but satisfy all axioms in the theory. Non-standard models usually contain more items than the standard model, known as "non-standard elements". In arithmetic, the standard model is $\mathbb{N}$, and all non-standard models contain elements $c > \mathbb{N}$ that are, in a sense, larger than all natural numbers. Observe that semantic conservativity is nontrivial only if we acknowledge the existence of non-standard models, as every set of axioms consistent with the standard model will automatically be semantically conservative if the only model we acknowledge as a model of the theory is the standard one.

Motivations for semantic conservativity thus relate to arguments against restricting one's attention only to the intended model of a theory. A prominent one is due to Halbach [Hal99] in the context of deflationism, via a two-step argument. First, Halbach argues that the deflationary understanding of truth is only compatible with the T-schemas being understood as axioms. The deflationary notion of truth is absolute, meaning that they do not define the notion of truth of an object language in a meta-language. Therefore, unlike Tarski's original proposal, the deflationist uses the T-schema to stipulate the content of truth-in-$\mathcal{L}$ in $\mathcal{L}$. By the undefinability theorem, the T-schema cannot form a definition of truth. Neither can its instances stipulate the extension of the truth predicate in $\mathcal{L}$, since they are not sentences in the

metalanguage. The only option left, as argued by Halbach, is to treat them as axiomatisations of a primitive notion of truth.

But then the T-schema in itself cannot fix the standard model, and sentences φ that are true in the standard model $\mathbb{N} \vDash \varphi$ but are false in some non-standard model cannot be viewed as consequences of deflationary truth. More importantly for our purpose, Halbach argues that the deflationist cannot restrict their attention to the standard model $\mathbb{N}$ in any way. For example, they cannot say that deflationism is applied to an interpreted language whose intended domain is just the natural numbers $\mathbb{N}$. This is because outlining the content of $\mathbb{N}$ necessarily involves outlining what is true about it, which then relies on a pre-theoretical notion of truth that the deflationists are trying to explicate, rendering their justification circular. The point is best illustrated in technical terms. Suppose φ is a theorem of deflationist truth if and only if $(\mathbb{N}, T) \vDash \varphi$. Here $(\mathbb{N}, T) \vDash \varphi$ is a statement in the metalanguage, where T fixes the extension of the truth predicate in the object language. The problem is that deflationary truth does not allow one to decide the exact content of T; therefore, deflationists have no way to decide whether φ is a consequence of arithmetical truth.

Regardless of one's opinion on the cogency of Halbach's step 1, Halbach's step 2 generalises to any axiomatic theory of truth, whether it is deflationary in spirit or not. Generally, it is a merit for an axiomatic approach of anything to distance itself from the standard model of that subject matter, because neither the standard model $\mathcal{M}$, nor all of its logical consequences $\mathsf{Th}(\mathcal{M})$ can be specified by first-order axiomatisations. Therefore, the arithmetical standard model $\mathbb{N}$, albeit assumed to capture what is the case in the mathematical reality, is in a sense "unknown" from a first-order axiomatic perspective. The investigations of truth should not depend on the assumption that we have full knowledge of $\mathbb{N}$ or $\mathsf{Th}(\mathbb{N})$. The usual approach, as described by Kaye [Kay16], is to stipulate a set of axioms that is reasonably accepted about $\mathbb{N}$ and investigate its consequences, including the non-standard models it gives rise to. In arithmetic, this is usually PA. A PA^- theorem φ is assertible/true not because it is true in $\mathbb{N}$, but is a proof-theoretical consequence of PA.

In the case of truth, one should also start by stipulating what is accepted about truth (e.g., T-schema), and investigate the (possibly non-standard) models this gives rise to. The notion of semantic conservativity seems to be just the right notion for this occasion. In the rest of the chapter, we will see that $\mathsf{CT}[\mathsf{PA}^-]$ is not semantically conservative over PA^-. We start by showing the argument for semantic nonconservativity of compositional truth in PA in Section 3.2, which is generalized to PA^- in Section 3.3.

3.2 Results for PA

Most published arguments for the semantic nonconservativity of $\mathsf{CT}[\mathsf{PA}]$ (e.g. Halbach [Hal10], Cieśliński [Cie17]) rely on the following theorem:

Theorem 104 (Lachlan's Theorem)**.** *Let $\mathcal{M} \vDash \mathsf{PA}$ be non-standard and let S be a partial non-standard satisfaction class for $\mathcal{M}$, then $\mathcal{M}$ is recursively saturated.*

where a partial satisfaction class is defined in Definition 85.[1] On the other hand, some non-standard models of PA are

[1] As noted by Kaye [Kay16], Lachlan's theorem bridges semantic conservativity with syntactic conservativity, albeit in a language different from $\mathcal{L}_{\mathsf{PA}}$. According to Lachlan's theorem, any non-standard model that carries a satisfaction class is recursively saturated. Therefore, every model of $\mathcal{M} \vDash \mathsf{PA}^{\mathsf{FS}}$ either is standard, or is recursively saturated. Let "if the model is non-standard, then it is recursively saturated" be expressed by the following statement:

$$\varphi :\leftrightarrow \forall a(\bigwedge_{n\in\mathbb{N}} \exists x \bigwedge_{i<n} \theta_i(x,a) \rightarrow \forall y \exists x \bigwedge_{i\in\mathbb{N}} (i < y \rightarrow \theta_0(x,a)))$$

where $(\theta_i)_{i\in\mathbb{N}}$ is a schematic variable over recursive sequences of $\mathcal{L}_{\mathsf{PA}}$ formulas. Since some models are not recursively saturated, $\mathsf{PA}^{\mathsf{FS}} \vdash \varphi$ but $\mathsf{PA} \nvdash \varphi$. Thus $\mathsf{PA}^{\mathsf{FS}}$ is not conservative over PA in an infinite language. Since there seems to be no reason why one should restrict attention to

not recursively saturated:

Lemma 105. *Any consistent extension of* PA *carries a model that is not recursively saturated. In particular, there are non-standard models of Peano arithmetic that are not recursively saturated.*

For a proof of the lemma, see Halbach [Hal10], p. 75. Since a non recursively saturated model cannot expand to a recursively saturated one, we have:

Corollary 106. CS[PA] *is not semantically conservative over* PA.

Theorem 107 (Cieśliński 2017 [Cie17], Corollary 7.0.6)**.** *All non-standard models of* PA *expandable to models of* CT[PA] *are recursively saturated. So* CT[PA] *is not semantically conservative over* PA.

Proof. Recall the definition of $\varphi(\mathsf{c} \circ \alpha)$ which replaces every $v \in \mathsf{FV}(\varphi)$ with the domain constant c_a where $\alpha(v) = a$. Let $\mathcal{M}$ be a non-standard model of PA where $(\mathcal{M}, T) \vDash \mathsf{CT}[\mathsf{PA}]$. To show that $\mathcal{M}$ is expandable to a model of $\mathsf{PA}^{\mathsf{FS}}$, we define a satisfaction class as follows:

$$S = \{(\alpha, \varphi) \mid \mathsf{Form}(\varphi) \wedge \mathsf{Asn}(\alpha, \varphi) \wedge \varphi(\mathsf{c} \circ \alpha) \in \mathsf{T}\}$$

Which is just $\mathsf{S}(\mathsf{T})$ defined above, which is shown to be a satisfaction class by induction. By Lachlan's theorem, all non-standard models of CS[PA] are recursively saturated. □

To generalise Theorem 107 to PA^{-}, we note several important lemmas that the proof relies on.

Lemma 108 (Kaye [Kay91], Lemma 11.3)**.** *Let $f : \mathbb{N} \to \mathbb{N}$ be recursive, and let $\mathcal{M} \vDash \mathsf{PA}$ be non-standard. then there is a $b \in M$ such that $\mathcal{M} \vDash (b)_n = f(n)$ for all $n \in \mathbb{N}$. Moreover, if the image of f, $\mathsf{Im}(f) = \{f(n) \mid n \in \mathbb{N}\}$ is a recursive set, then there exists $c \in M$ such that $\mathsf{Im}(f) = \{n \in \mathbb{N} \mid \mathcal{M} \vDash \exists x < c\,(b)_x = n\}$.*

Proof. Suppose $f : \mathbb{N} \to \mathbb{N}$ is a recursive function, then it is Σ_1-definable, i.e., there exists a Δ_0-formula θ with three free variables such that for all $n, m \in \mathbb{N}$, $\mathbb{N} \vDash \exists z \theta(n, m, z)$ iff $f(n) = m$. Since PA is Σ_1-complete and sequential, for any $\mathcal{M} \vDash \mathsf{PA}$, $\mathcal{M} \vDash \exists b \forall x, y, z < i(\theta(x, y, z) \to (b)_x = y)$ for all $i \in \mathbb{N}$. By overspill, there exists $b \in M$ and $a \in M \setminus \mathbb{N}$ where $\mathcal{M} \vDash \forall x, y, z < a(\theta(x, y, z) \to (b)_x = y)$. Suppose $f(n) = m$, therefore $\exists z \in \mathbb{N}$, $\mathcal{M} \vDash \theta(m, n, z)$, so $\mathcal{M} \vDash (b)_n = m$. Conversely, suppose there exists some c such that $f(n) = c$, and $\mathcal{M} \vDash (b)_n = m$ for some $m \in M$. We see that $c = m$.

Suppose $\mathsf{Im}(f)$ is recursive. Then there is a Δ_0 formula $\psi(y, w)$ such that for all $n \in \mathbb{N}$, $n \in \mathsf{Im}(f)$ iff $\mathbb{N} \vDash \forall w \psi(n, w)$. Therefore clearly for $i \in \mathbb{N}$, $\mathcal{M} \vDash \forall x < i \forall w < i\, \psi((b)_i, w)$. By overspill, there exists $c \in M \setminus \mathbb{N}$ where $\mathcal{M} \vDash \forall x < c \forall w < c\, \psi((b)_x, w)$. Suppose $n \in \mathsf{Im}(f)$. Then $n = (b)_i$ for some $i \in \mathbb{N}$. Since $c > \mathbb{N}$, $\mathcal{M} \vDash \exists x < c\,(b)_x = n$. Conversely, suppose that $\mathcal{M} \vDash \exists x < c\,(b)_x = n$. We also know that $\mathcal{M} \vDash \forall x < c \forall w < c\, \psi((b)_x, w)$, therefore $\mathcal{M} \vDash \forall w < c\, \psi((b)_x, w)$, and in particular $\mathcal{M} \vDash \psi(n, w)$ for all $w \in \mathbb{N}$. Since ψ is a Δ_0 formula and $\mathbb{N} <_{\Delta_0} \mathcal{M}$, $\mathbb{N} \vDash \forall w \psi(n, w)$, $n \in \mathsf{Im}(f)$. □

Lemma 109. *Let $\mathcal{M} \vDash \mathsf{PA}$ be non-standard, S be a non-standard satisfaction class, and $\varphi(x_0, \ldots, x_n)$ be a standard formula with only $x_0, \ldots, x_n$ free. Then, for all $\alpha \in \mathsf{Asn}^{\mathcal{M}}$,*
$(\mathcal{M}, S, F) \vDash \mathsf{Form}(\varphi) \wedge \mathsf{Asn}(\varphi, \alpha) \wedge \mathsf{S}(\varphi, \alpha) \to \varphi(\alpha(x_0), \ldots, \alpha(x_n))$.

syntactic conservativity of the finite language, this adds another reason why one should consider semantic conservativity in evaluating truth theories.

The philosophical significance of the infinite language is debatable. Those who believe that deflationism cannot use semantic arguments might reply that in the infinite language, one has all the resources to implement model-theoretical concepts, thus it is incompatible with deflationism.

Proof. By induction on the complexity of formulas. Base case: $\varphi = (R, (t_1, \ldots, t_n))$ where t_is are meta-variables referring to variables in $\mathcal{M}$. Suppose $(M, S, F) \vDash \mathsf{Form}(\varphi) \wedge \mathsf{Asn}(\varphi, \alpha) \wedge \mathsf{S}(\varphi, \alpha)$. By definition of satisfaction class, in particular $\mathsf{tarski}_{1,R}$, $R(\alpha(x_0), \ldots, \alpha(x_n))$. so $\mathcal{M} \vDash \varphi(\alpha(x_0), \ldots, \alpha(x_n))$.

Inductive case: $\varphi = (\exists, (x, \varphi_1))$. By inductive hypothesis, for all $\alpha \in \mathsf{Asn}^{\mathcal{M}}$, $(M, S, F) \vDash \mathsf{Form}(\varphi_1) \wedge \mathsf{Asn}(\varphi_1, \alpha) \wedge \mathsf{S}(\varphi_1, \alpha) \to \varphi_1(\alpha(x_0), \ldots, \alpha(x_n))$. Suppose $(M, S, F) \vDash \mathsf{Form}(\varphi) \wedge \mathsf{Asn}(\varphi, \alpha) \wedge \mathsf{S}(\varphi, \alpha)$, by the definition of satisfaction classes, $(M, S, F) \vDash \exists \alpha' \supseteq \alpha \mathsf{S}(\varphi_1, \alpha')$. So $(M, S, F) \vDash \varphi_1(\alpha'(x), \alpha'(x_0) \ldots, \alpha'(x_n))$, $(M, S, F) \vDash \exists x \varphi_1(\alpha'(x_0), \ldots, \alpha'(x_n))$, since $\alpha(x_i) = \alpha'(x_i)$, $(M, S, F) \vDash \exists x \varphi_1(\alpha(x_0), \ldots, \alpha(x_n))$. □

We see the inductive proof easily generalizes to PA^- by applying shortening of cuts to the arithmetized statement of Lemma 109.

3.3 Lachlan's Theorem for PA^-

This section generalises Lachlan's theorem to PA^-. The proofs of the supporting lemmas are given in Subsection 3.3.1, and Lachlan's theorem in Subsection 3.3.2.

3.3.1 Overspill and Other Lemmas

As illustrated in Lemma 23, overspill describes the behavior of non-standard models. In the context of PA, non-standard models refer to any models of PA that are not isomorphic to $\mathbb{N}$. But this is less clear in the case of weak arithmetic. For example, recall that the intended model for PA^-, $\mathbb{Z}[X]^+$, is non-standard when we interpret the operations $+, \times, 0$ as arithmetical (i.e. as about $\mathbb{N}$), but standard when the language is interpreted as the language for discretely ordered commutative rings. Since Lachlan's theorem is about non-standard models, we shall first fix its definition in the context of PA^-.

We notice that $\mathbb{N}$ is a model of PA^-, and any PA^- model contains the standard natural numbers as an initial segment.

Lemma 110. *Let $N(x)$ be such that every $x \in N$ is hereditarily either odd or even. That is,*
$N(x) :\leftrightarrow \forall y \leq x[(\exists z\, y = 2 \times z) \vee (\exists z\, y = 2 \times z + 1)]$ *(Here 2 and 1 abbreviate $\mathsf{0}''$ and $\mathsf{0}'$.) Then $N(x)$ defines a cut on any $\mathcal{M} \vDash \mathsf{PA}^-$.*

Proof. It suffices to prove that $N(x)$ is inductive: $\mathsf{PA}^- \vdash N(\mathsf{0})$ since $\mathsf{PA}^- \vdash \mathsf{0} \times 2 = \mathsf{0}$. Suppose $\mathsf{PA}^- \vdash N(k)$. If $\mathsf{PA}^- \vdash k = 2 \times z$, then $\mathsf{PA}^- \vdash k' = 2 \times z + 1$. If $\mathsf{PA}^- \vdash k = 2 \times z + 1$, then $\mathsf{PA}^- \vdash k' = 2 \times (z + 1)$. □

This means that we can still interpret the connectives and constants as arithmetical.

Definition 111 (Non-standard models for PA^-)**.** *A model $\mathcal{M} \vDash \mathsf{PA}^-$ is a weakly non-standard model if all definable cuts $I \subseteq M$ have a non-standard element $c > \mathbb{N}$. The model $\mathcal{M} \vDash \mathsf{PA}^-$ is a strongly non-standard model if there exists an element $c > \mathbb{N}$ such that for all definable cuts $I \subseteq M$, $c \in I$.* (*Any $c \in M$ is non-standard if $c > \mathbb{N}$.*)

The reason for introducing the distinction between weakly and strongly non-standard models is that weakly non-standard models are insufficient to enforce recursive saturation:

Remark 112. *All weakly non-standard but not strongly non-standard models are not recursively saturated.*

Proof. Let $\mathcal{M}$ be a weakly non-standard but not strongly non-standard model. By definition, the intersection of all definable cuts in $\mathcal{M}$ is $\mathbb{N}$. This is because any cut contains $\mathbb{N}$ by definition of cut, and by the fact that $\mathcal{M}$ is not strongly non-standard, no non-standard c is in all cuts. We can enumerate all cuts in $\mathcal{M}$ by an effective procedure as follows:

1. First enumerate all $\mathcal{L}_{\mathsf{PA}}$ formulas $\varphi_i(x),\ i \in \mathbb{N}$ with single free variable x.
2. Given $\varphi_i(x)$, we transform it syntactically into $\psi_i :\leftrightarrow (\mathsf{cut}(\varphi_i) \wedge \varphi_i(x)) \vee (\neg\mathsf{cut}(\varphi_i) \wedge x = x)$, where $\mathsf{cut}(\varphi)$ abbreviates the formula $\varphi(0) \wedge \forall x(\varphi(x) \to \varphi(x')) \wedge \forall x \forall y(\varphi(x) \wedge y < x \to \varphi(y))$.

Finally, define recursive type $p(x) = \{\psi_n(x) \wedge x \geq \bar{n} \,|\, n \in \mathbb{N}, x \in I_n\}$. Since $\bigcap_{i\in\mathbb{N}}\{x \geq \bar{n} \,|\, x \in I_n\} = \emptyset$, no element $n \in M$ instantiates $p(x)$. □

The proof also establishes that the distinction is meaningful. There exists a strongly non-standard model. Suppose there is not. Consider arbitrary $\mathcal{M} \vDash \mathsf{PA}^-$. We denote $X_i := \{\mathcal{M} \vDash \psi_i(x), x \in M\}$, therefore $\bigcap_{i\in\mathbb{N}} X_i = \mathbb{N}$. But then $\{\psi_i, i \in \mathbb{N}\}$ axiomatize $\mathbb{N}$, which is impossible. However, whether there is a weakly non-standard model that is not strongly non-standard remains unknown.

Lemma 113 (Overspill, PA^-)**.** *Let $\mathcal{M}$ be a weakly non-standard model of PA^- and $b \in M$, and assume $\varphi(x, y)$ is a formula with x, y free. Then, if $\mathcal{M} \vDash \varphi(n, b)$ holds for every $n \in \omega$, there is a non-standard number $c \in M$ such that $\mathcal{M} \vDash \forall x \leq c\, \varphi(x, b)$.*

Proof. Suppose, for a contradiction, that there is no such c. Consider the formula defined as $\psi(x) :\leftrightarrow \forall y \leq x\, \varphi(y, b)$. We show that $\psi(x)$ is inductive: since $\mathcal{M} \vDash \varphi(0, b)$, $\mathcal{M} \vDash \psi(0)$. Now consider any $a \in M$. If a is a standard number, then a' is also standard, so we have $\mathcal{M} \vDash \psi(a) \to \psi(a')$. If a is a non-standard number, by assumption $\mathcal{M} \nvDash \psi(a)$, so $\mathcal{M} \vDash \psi(a) \to \psi(a')$. We also observe that $\psi(x)$ is downward closed, so $\psi(x)$ is a cut. This contradicts the definition of a weakly non-standard model, where there must exist a non-standard c where $\psi(c)$. □

Overspill for strongly non-standard models automatically follows. Moreover, there is a stronger overspill principle for strongly non-standard models:

Lemma 114 (Overspill for strongly non-standard models, PA^-)**.** *Let $\mathcal{M}$ be a strongly non-standard model of PA^- and $b \in M$, and assume $\varphi(x, y)$ is a formula with x, y free. Then, if $\mathcal{M} \vDash \varphi(n, b)$ holds for every $n \in \omega$, there is a non-standard number $c \in M$ such that $\mathcal{M} \vDash \forall x \leq c\, \varphi(x, b)$, and c is such that for all definable cuts $I \subseteq M$, $c \in I$.*

Proof. Similar to the proof of Lemma 113. □

We then turn to generalizing the lemmas needed to prove Lachlan's theorem. First, we need to represent the graph of primitive recursive functions up to non-standard numbers using sequences. We first check that total functions are defined up to some non-standard number in some non-standard models.

Lemma 115. *Let f be a total function in $\mathbb{N}$, and $\mathcal{M}$ be a weakly non-standard model of PA^-. There is a non-standard element $a \in M$, such that for all $b \leq a$, $f(b)$ is defined in $\mathcal{M}$.*

Proof. Consider the set $X = \{x \in M, \forall y \leq x\, f(y)\ \downarrow\}$. X is either closed under $+1$ or not. If it is, then X is inductive, there is a definable cut I in which f is total. Since $\mathcal{M}$ is weakly non-standard, pick a to be any non-standard element in I. If it is not, then there exists a $c \in M$ where $\forall y \leq c\, f(y) \downarrow$ and $F(c + 1) \uparrow$. We take $a := c$, note that c must be non-standard. □

Lemma 116 (Encoding of recursive functions, PA^-)**.** *Let $f : \mathbb{N} \to \mathbb{N}$ be recursive and $\mathcal{M} \vDash \mathsf{PA}^-$ be weakly non-standard. Then there is a $b \in M$ such that $\mathcal{M} \vDash (b)_n = f(n)$ for all $n \in \mathbb{N}$. Moreover, if the image of f, $\mathsf{Im}(f) = \{f(n) \,|\, n \in \mathbb{N}\}$ is a recursive set, then there exists $c \in M$ such that $\mathsf{Im}(f) = \{n \in \mathbb{N} \,|\, \mathcal{M} \vDash \exists x < c\, (b)_x = n\}$.*

Proof. Note that PA^- is Σ_1-complete and sequential. Overspill holds for weakly non-standard models of PA^-. So the proof is the same as for Lemma 108. □

We will also define a new recursive series of formulas from an existing recursive series. The following two lemmas state that this operation is well-defined.

Lemma 117 (PA^-)**.** *Let $\mathcal{M} \vDash \mathsf{PA}^-$ be a weakly non-standard model. Consider recursive function $f : i \mapsto \theta_i$ where there exists $b, c \in M$ such that $\mathcal{M} \vDash \forall x < c((b)_x = f(x) \wedge \mathsf{Form}((b)_x))$, and formulas $\langle \varphi_i \rangle_{i \in \mathbb{N}}$ obtained by recursively defined syntactical operations from $\langle \theta_i \rangle_{i \in \mathbb{N}}$. There exists recursive function $g : i \mapsto \varphi_i$ where there exists $b', c' \in M$ such that $\mathcal{M} \vDash \forall x < c'(b')_x = g(x) \wedge \mathsf{Form}((b')_x)$ for all $n \in \mathbb{N}$.*

Proof. Suppose the syntactical operation is defined by a recursive function $h : \mathsf{Form}^{\mathcal{M}} \mapsto \mathsf{Form}^{\mathcal{M}}$. Since $\mathcal{M} \vDash \mathsf{Form}((b)_n)$ for all $n \in \mathbb{N}$, $g = h \circ f$ is well-defined. It is also recursive, as recursive functions are closed under composition. By Lemma 116, there exists $c \in M$ such that $\mathcal{M} \vDash (c)_n = h(f(n))$ for all $n \in \mathbb{N}$. Since $h : \mathsf{Form}^{\mathcal{M}} \mapsto \mathsf{Form}^{\mathcal{M}}$, $\mathcal{M} \vDash \mathsf{Form}(h(f(n)))$ for all $n \in \mathbb{N}$. By overspill, there exists $c' > \mathbb{N}$ where $\mathcal{M} \vDash \forall x \leq c'((b')_x = h(f(x)) \wedge \mathsf{Form}((b')_x))$. □

It is worth noting that we have no idea whether $c \leq c'$ or the other way round. We have control of the size of c' in strongly non-standard models:

Lemma 118 (PA^-)**.** *Let $\mathcal{M} \vDash \mathsf{PA}^-$ be a strongly non-standard model. Consider recursive function $f : i \mapsto \theta_i$ where there exists $b, c \in M$ such that $\mathcal{M} \vDash \forall x < c((b)_x = f(x) \wedge \mathsf{Form}((b)_x))$, and formulas $\langle \varphi_i \rangle_{i \in \mathbb{N}}$ obtained by recursively defined syntactical operations from $\langle \theta_i \rangle_{i \in \mathbb{N}}$. There exists recursive function $g : i \mapsto \varphi_i$ and $b', c' \in M$ such that $\mathcal{M} \vDash \forall x < c'(b')_x = g(x) \wedge \mathsf{Form}((b')_x)$ and $\mathcal{M} \vDash \forall x < c'((b)_x = f(x) \wedge \mathsf{Form}((b)_x))$.*

Proof. Immediate, given strong overspill. □

3.3.2 Lachlan's Theorem

In this subsection, we give the proof of Lachlan's theorem in PA^-. We follow Kaye [Kay91]'s proof but focus on full satisfaction classes only.

Theorem 119 (Lachlan's Theorem, PA^-)**.** *Let $\mathcal{M} \vDash \mathsf{PA}^-$ be strongly non-standard and let S be a non-standard full satisfaction class for M, then M is recursively saturated.*

Proof. Suppose, for a contradiction, that $\mathcal{M} \vDash \mathsf{PA}^-$ is a strongly non-standard model and not recursively saturated, i.e., there exists a recursive type $p(v)$ that is not realized in $\mathcal{M}$. Since $p(v)$ is recursive, there exists a primitive recursive function $f : i \mapsto \varphi_i(v)$ where $p(v) = \{\varphi_i(v) \mid i \in \mathbb{N}\}$. By Lemma 116, there exists non-standard $a, d \in M$ where $\mathcal{M} \vDash \forall x < d\,(a)_x = f(x) = \varphi_x$. We observe that for all $i \in \mathbb{N}$, $\mathcal{M} \vDash \mathsf{Form}((a)_i)$, therefore we obtain by strong overspill $e \in M$ where $\mathcal{M} \vDash \forall x < e\,((a)_x = \varphi_x \wedge \mathsf{Form}((a)_x))$. We denote $A_i = \{x \in M \mid \mathcal{M} \vDash \varphi_i(x)\}$.

Ultimately, we want to define a partition using A_i. But we first have to ensure that $A_0 = M \supseteq A_1 \supseteq A_2 \supseteq \ldots$ and $A_i \neq A_{i+1}$. This can be done by defining $\varphi'_0 = (x = x)$, $\varphi'_{i+1} = \varphi_i \wedge \varphi'_i \wedge \exists z < x\, \varphi'_i(x)$, by a primitive recursive syntactical transformation function $s : \{\varphi_i \mid i \in \mathbb{N}\} \mapsto \{\varphi'_i \mid i \in \mathbb{N}\}$. By Lemma 118, there exists $a', f \in M$ where $\mathcal{M} \vDash \forall x < f\,((a')_x = \psi'_x \wedge \mathsf{Form}((a')_x))$. Since in strongly non-standard models, there exists $\mathfrak{d}$ where $\mathfrak{d} \in I$ for every I, and we choose exactly such $\mathfrak{d}$ in the proof of both Lemma 118 and Lemma 114, we thus may assume that $e = f = \mathfrak{d}$. We then replace every $A_i, i \in \mathbb{N}$ with sets defined by $\varphi'_i, i \in \mathbb{N}$.

Therefore $B_0 = \emptyset$ and $B_{i+1} = A_i \setminus A_{i+1}$ forms a partition of M. $\{B_i \mid i \in \mathbb{N}\}$ is defined by $\{\theta_i \mid i \in \mathbb{N}\}$ where $\theta_0 = \neg(x = x)$ and $\theta_i = \varphi_i \wedge \neg\varphi_{i+1}$. To implement B with the satisfaction predicate, again by Lemma 118 and Lemma 114, there exists $b \in M$ where b codes the sequence $\theta_0(v), \theta_1(v), ...$, i.e. for all $i \in \mathbb{N}$, $(b)_i = \theta_i(v)$. Moreover, $\mathcal{M} \vDash \forall i \leq k' \, \mathsf{Form}((b)_i)$. This allows us to define B_i by

$$B_i = \{x \in M \mid (\mathcal{M}, S, F) \vDash \theta_i(x)\}$$

for all i. Recall that $\langle v, x\rangle$ refers to the variable assignment that assigns to the variable v the object x. It is easy to see that this definition coincides with the B_is defined above, so they define a partition of M.

We then define $\{C_i, i \in \mathbb{N}\}$ from $\{B_i, i \in \mathbb{N}\}$. The idea is as follows:

$$C_0 = \emptyset;$$
$$C_{i+1} = \begin{cases} B_1 & \text{if } C_i = \emptyset; \\ B_{j+1} & \text{if } j \text{ is least such that } B_j \cap C_i \neq \emptyset; \\ \emptyset & \text{if } C_i \neq \emptyset \text{ but no such } j \text{ exists.} \end{cases}$$

where i might be possibly non-standard. Formally, we define formulas $\{\gamma_i, i \in \mathbb{N}\}$ with the intention that $C_i = \{x \in M \mid (\mathcal{M}, S, F) \vDash \mathsf{S}(\gamma_i, \langle v, x\rangle)\}$. Let $\gamma_0(v) = \neg(v = v)$. Assume we have γ_i, to define γ_{i+1}, we first define

$$\begin{aligned} \delta_1 &= \neg\exists y \gamma_i(y) \\ \delta_2 &= \exists y(\gamma_i(y) \wedge \theta_1(y)) \\ &\vdots \\ \delta_{j+1} &= \exists y(\gamma_i(y) \wedge \theta_j(y)) \end{aligned}$$

Intuitively, δ_1 expresses that C_i is empty, and each of δ_i where $i > 1$ expresses that $C_i \cap B_{i-1}$ is nonempty. We then put

$$\begin{aligned} \gamma_{i+1}(v) :\leftrightarrow\ & (\delta_1 \wedge \theta_1(v)) \vee \\ & (\neg\delta_1 \wedge (\delta_2 \wedge \theta_2(v)) \vee \\ & (\neg\delta_2 \wedge (\delta_3 \wedge \theta_3(v)) \vee \\ & \vdots \\ & (\neg\delta_{\mathfrak{d}-1} \wedge (\delta_{\mathfrak{d}} \wedge \theta_{\mathfrak{d}}(v)) \vee (\neg\delta_{\mathfrak{d}} \wedge \neg(v = v)) \ldots))) \end{aligned}$$

Since both syntactical operations $F, G : \mathsf{Form}^{\mathcal{M}} \mapsto \mathsf{Form}^{\mathcal{M}}$ corresponding to the formation of γ_1 and γ_{i+1} can be written as primitive recursive functions, by Lemma 118 there exists c such that $\mathcal{M} \vDash \forall x < \mathfrak{d} \, (c)_x = G(i, b, \mathfrak{d}) \wedge \mathsf{Form}((c)_x)$. Moreover, suppose that a is the upper bound on the size of formulas handled by the satisfaction predicate, i.e., the size of the cut $\mathsf{Form}^{\mathcal{M}}$. Since $\mathsf{Form}((b)_i) \wedge \mathsf{Form}(G(i, b, \mathfrak{d})) \wedge (b)_i < a \wedge G(i, b, v) < a$ is a cut, $\mathfrak{d}$ is such that $\forall x < \mathfrak{d}$ satisfies the condition above.

It remains to be checked that our definitions of C_i are adequate. It is sufficient to check

1. for all $i \leq v$, $C_i = B_j$ for some $j \in \mathbb{N}$.

2. if $C_i = B_j$ then $C_{i+1} = B_{j+1}$.

3. $C_i \neq \emptyset$ for all $i > 0$.

We prove the three conditions simultaneously by a case distinction on the nature of C_i. The proof in Kaye [Kay91] involves only principles of elementary logic and definitions of satisfaction class ($\mathsf{tarski}_{1,R}$ to tarski_4 in Chapter 3), hence automatically generalizes to PA^-. We present the proof for comprehensiveness.

- $C_i = \emptyset$. $C_i = B_0$, we show that $C_{i+1} = B_1$. By definition, $x \in C_{i+1}$ iff $(\mathcal{M}, S, F) \vDash \mathsf{S}(\gamma_{i+1}(x), \langle v, x\rangle)$. Since $C_i = \emptyset$, $(\mathcal{M}, S, F) \vDash \neg\exists x\mathsf{S}(\gamma_i(x), \langle v, x\rangle)$. By the Tarskian clause for $\exists$, $(\mathcal{M}, S, F) \vDash \forall x\neg\mathsf{S}(\exists y\gamma_i(y), \langle v, x\rangle)$. By the Tarskian clause for $\neg$, $(\mathcal{M}, S, F) \vDash \forall x\mathsf{S}(\neg\exists y\gamma_i(y), \langle v, x\rangle)$. Recall that $\delta_1 = \neg\exists y\gamma_i(y)$, so $(\mathcal{M}, S, F) \vDash \forall x\mathsf{S}(\delta_1, \langle v, x\rangle)$. $(\mathcal{M}, S, F) \vDash \neg\exists x\mathsf{S}(\gamma_i(x), \langle v, x\rangle)$ entails $(\mathcal{M}, S, F) \vDash \mathsf{S}(\delta_1, \langle v, x\rangle)$, for any $x \in M$.

  This allows us to simplify each of the δ_k defined above. More specifically, observe that each of δ_k entails $\neg\delta_1$. Hence $(\mathcal{M}, S, F) \vDash \mathsf{S}(\gamma_{i+1}(x), \langle v, x\rangle)$ iff

  $$(\mathcal{M}, S, F) \vDash \mathsf{S}((\delta_1 \wedge \theta_1(x)) \vee (\neg\delta_1 \wedge \psi(x)), \langle v, x\rangle), \tag{3.1}$$

  where ψ is some subformula of γ_{i+1}. By definition of satisfaction class, in particular, Tarskian clauses for $\vee$, $\wedge$ and $\neg$, we can rewrite 3.1 above as

  $$(\mathcal{M}, S, F) \vDash (\mathsf{S}(\delta_1, \langle v, x\rangle) \wedge \mathsf{S}(\theta_1(x), \langle v, x\rangle)) \vee (\neg\mathsf{S}(\delta_1, \langle v, x\rangle) \wedge \mathsf{S}(\psi(x), \langle v, x\rangle)).$$

  Recall that $(\mathcal{M}, S, F) \vDash \forall x\mathsf{S}(\delta_1, \langle v, x\rangle)$. So $x \in C_{i+1}$ iff $(\mathcal{M}, S, F) \vDash \mathsf{S}(\theta_1(x), \langle v, x\rangle)$, i.e., $x \in B_1$.

- $C_i \neq \emptyset$. Since $B_i = \{x \in M \mid (\mathcal{M}, S, F) \vDash \mathsf{S}(\theta_i(v), \langle v, x\rangle)\}$ defines a partition of M, there exists $j \in \mathbb{N}$ such that $C_i \cap B_j \neq \emptyset$. Choose the least such j. We show that $C_{i+1} = B_{j+1}$. Since $C_i \neq \emptyset$, $(\mathcal{M}, S, F) \vDash \exists x\mathsf{S}(\gamma_i, \langle v, x\rangle)$, therefore $(\mathcal{M}, S, F) \vDash \mathsf{S}(\exists x\gamma_i, \langle v, x\rangle)$ for every x, and $(\mathcal{M}, S, F) \vDash \mathsf{S}(\neg\neg\exists x\gamma_i, \langle v, x\rangle)$ for every x. Therefore $(\mathcal{M}, S, F) \vDash \forall x\mathsf{S}(\neg\delta_1, \langle v, x\rangle)$. Since $C_i \cap B_k = \emptyset$ for all $1 \leq k < j$, we have $(\mathcal{M}, S, F) \vDash \forall x\mathsf{S}(\neg\delta_{k+1}, \langle v, x\rangle)$ for all k. Since $C_i \cap B_j \neq \emptyset$, $(\mathcal{M}, S, F) \vDash \forall x\mathsf{S}(\delta_{j+1}, \langle v, x\rangle)$. Therefore we observe that C_{i+1} is expressed by the following formula:

  $$\begin{aligned}
  &((\delta_1 \wedge \theta_1(v))\vee\\
  &(\neg\delta_1 \wedge (\delta_2 \wedge \theta_2(v))\vee\\
  &(\neg\delta_2 \wedge (\delta_3 \wedge \theta_3(v))\vee\\
  &(\neg\delta_3 \wedge (\dots\\
  &\vdots\\
  &(\neg\delta_j \wedge (\delta_{j+1} \wedge \theta_{j+1}(v))\vee\\
  &(\neg\delta_{j+1} \dots)))))))
  \end{aligned}$$

  which, by the Tarskian clauses for $\vee, \wedge, \neg$, is equivalent to that $(\mathcal{M}, S, F) \vDash \mathsf{S}(\theta_{j_1}, \langle v, x\rangle)$. Therefore $C_{i+1} = B_{j+1}$.

The first and second conditions are proved by an inductive argument. The base case is satisfied because $C_0 = B_0$. For the inductive case, if $C_i = B_j \neq \emptyset$, then the intersections of C_i with each of $B_0, \dots, B_{j-1}$ are all empty. Hence $C_{i+1} = B_{j+1}$. (Here we need to apply the shortening of cuts again, but i is still bounded by non-standard number k since it appears in any cut.) To show the third condition holds, consider any C_{i+1}. Suppose $C_i = \emptyset = B_0$, then

$C_i = B_1$ by the argument above. Suppose otherwise that $C_i \neq \emptyset$. Therefore $C_i = B_{j+1}$ where j is the least number such that $C_{i-1} \cap B_j \neq \emptyset$. Since $\{B_i, i \in \mathbb{N}\}$ forms a partition of M and $B_0 = \emptyset$, $\{B_i, i \in \mathbb{N}^+\}$ forms a partition of M. So such positive j must exist, thus $C_{i+1} = B_{j+1}$, which by definition is not empty.

Finally, we define $f(i)$ as the unique $j \in \mathbb{N}$ where $C_i = B_j$, we thus obtain an infinite descending sequence of natural numbers $f(i) > f(i-1) > f(i-2) > \ldots$, where i is non-standard. A contradiction. □

One extra step is needed to show that $\mathsf{CT}[\mathsf{PA}^-]$ is not semantically conservative to PA^-: we have to show that there is at least one model of PA^- that cannot be extended to a model of $\mathsf{CT}[\mathsf{PA}^-]$.

Lemma 120 (Łełyk)**.** *For each countable sequential theory T, there is a countable model of T which is not recursively saturated*[2]*.*

Proof. Suppose otherwise. Consider $\mathcal{M}$ a model of T, then every recursive type over M is realized, in particular, the following type

$$p(x) := \{\varphi \in x \leftrightarrow \varphi, \forall \varphi \in \mathsf{Sent}\}$$

is realised. By the omitting types theorem, $p(x)$ is isolated. There exists a formula $G(x)$ where $T' :\leftrightarrow T + \exists x G(x)$ is consistent and $T \vdash G(x) \rightarrow p(x)$. This allows one to define $Tr(\varphi) := \exists x(G(x) \wedge \varphi \in x)$, which is a truth predicate for T', contradicting Tarski's undefinability theorem. □

Lemma 121. *If Γ is a consistent extension of PA^-, then there is a countable strongly non-standard model of Γ that is not recursively saturated.*

Proof. Recall the enumeration of cuts in the proof of Lemma 112, where all cuts are enumerated by $\{\psi_i, i \in \mathbb{N}\}$. We stipulate the existence of a non-standard element c that occurs in all cuts by first expanding PA^- with a constant c, then add the following axioms:

- $\Psi = \{\psi_n(\mathsf{c}), n \in \mathbb{N}\}$.
- $\Phi = \{\mathsf{c} > \mathsf{c}_n, n \in \mathbb{N}\}$.

Any finite subset P of $\Psi \cup \Phi$ is consistent with PA^-. Suppose k is the largest number such that $\mathsf{c} > \mathsf{c}_k \in P$. Therefore a PA^- model interpreting c as the standard number $k+1$ satisfies P. By Lemma 120, there is a countable model $\mathcal{M}$ of $\mathsf{PA}^- \cup \Psi \cup \Phi$ that is not recursively saturated. Since $\mathcal{M} \vDash \Psi \cup \Phi$, $c^{\mathcal{M}} > \mathbb{N}$, and is in any cut. So $\mathcal{M}$ is strongly non-standard. □

Corollary 122. $\mathsf{CT}[\mathsf{PA}^-]$ *is not semantically conservative over* PA^-.

Besides showing that $\mathsf{CT}[\mathsf{PA}^-]$ is not semantically conservative over PA^-, Lachlan's theorem outlines a measure of the strength of compositional truth, that compositionality itself is sufficient to enforce recursive saturation.

Corollary 123. *Given base theory PA^-, the model-theoretic strength of truth theories CT^- is that $\mathfrak{PA}^- \supset \mathfrak{RS} \supseteq \mathfrak{CT}^-$, where $\mathfrak{PA}^-$ is the class of all models of PA^-, $\mathfrak{RS}$ the class of recursively saturated models of PA and $\mathfrak{CT}^-$ denotes the class of models that carry a compositional truth without induction.*

In fact, with lemma 120, we have a much stronger result.

[2] This proof was suggested to us by Mateusz Łełyk in private correspondence.

Corollary 124. *Given T a computably enumerable consistent extension of* PA^-*, we have* $\mathfrak{T} \supset \mathfrak{RS} \supseteq \mathfrak{CT}^-$*, where* $\mathfrak{T}$ *is the class of all models of* T*, and* $\mathfrak{RS}$, $\mathfrak{CT}^-$ *are similar.*

This is in line with the results for PA, as summarized in Łełyk and Wcisło [ŁW17]:

Theorem 125. *Given base theory* PA*, the model-theoretic strength of truth theories* TB, UTB *and* CT^- *can be linearly ordered as:*

$$\mathfrak{PA} \supset \mathfrak{TB} \supset \mathfrak{RS} \supset \mathfrak{UTB} \supseteq \mathfrak{CT}^-$$

Where $\mathfrak{PA}$ *is the class of all models of* PA*,* $\mathfrak{TB}$ *denotes the class of all* PA *models that can be expanded to* $\mathsf{TT}^+[\mathsf{PA}]$*, and* $\mathfrak{UTB}$ *denotes the class of models that carries a uniform biconditional truth class with induction.*

where TB and UTB, since they contain induction, are unavailable for PA^-.

3.4 The Strength of Compositional Truth

I close the technical exposition with a short philosophical remark. Besides contributing to the non-conservativity result, Lachlan's theorem bounds the strength of compositional satisfaction: compositionality is sufficient to enforce recursive saturation. By Theorem 125, since compositional truth is recursively saturated but extended induction is not, recursive saturation is a mathematical concept that draws a line between the strength of two theories. (We are also fairly confident that some models of $\mathsf{TT}[\mathsf{PA}^-]$ are not recursively saturated, thus in both PA^- and PA recursive saturation distinguishes compositional from disquotational truth theories. We cannot give the proof due to time constraints, and refer the readers to the proof for PA in Cieśliński [Cie17], p. 100.)

These mathematical facts nicely accompany some facts about the interpretability of compositional and disquotational truth, as listed by Heck [Hec18]. To understand these results, we first introduce Heck [Hec18]'s technical setting, which is slightly different from ours. Recall, in the introduction, that the induction schema in the object theory enables both syntactical and arithmetical induction. A way to disentangle the two roles is to operate with "disentangled syntax" – a framework with an object language $\mathcal{L}$ and a disjoint language of syntax $\mathcal{S}$, usually treated as a copy of $\mathcal{L}$ but in a different font. The object theory T developed in $\mathcal{L}$ and the theory of syntax U in $\mathcal{S}$ can be of any strength stronger than PA^- (or $\mathsf{Q}_{\mathsf{Seq}}$, depending on one's taste about minimal theories.) The variables in T and U, and the domains these variables range over, are also distinct. One thus has a multi-sorted framework with at least three sorts: the first concerns variables ranging over the domain of the object theory, the second concerns variables ranging over the domain of the theory of syntax, while the third concerns variables over variable assignments, which map each variable (syntactic object) and object in the domain of the object theory. We use italic alphabets $v, x, y, z\ldots$ to denote variables in the syntax theory, upright alphabets **v,x,y,z**... for variables in the object theory, and $\alpha, \beta\ldots$ for assignment variables.

The setup leads to the following results:

Base: $U + T$	No Extended Induction	Extended Induction
Add Biconditionals	Locally Interpretable	Locally Interpretable
Add Satisfaction Axioms	Locally Interpretable	Unclear
Add Compositional Axioms	Not Interpretable	Not Interpretable

where U is either $\mathrm{I}\Sigma_n$ or PA. We see that compositional truth contributes more to the strength of the resulting theory than extended induction. Heck gives an exact characterisation of the logical strength of compositionality:

Fact 126. *Let T be a finitely axiomatized theory in $\mathcal{L}$. Then* $\mathsf{CT}_{\mathcal{L}}[\mathsf{Q}] + T$ *is mutually interpretable with* $\mathsf{Q} + \mathsf{Con}(T)$.

Here Q is the theory of syntax, which is as weak as possible, and T the object theory. Therefore, even if the syntactical and arithmetical roles of the induction principle are disentangled, and the truth theory developed in the syntax theory has no way to affect the truth theory, compositional truth still has significant logical strength in the sense that it is not interpretable in T.

The question is whether there is a similar semantic story. We first note that mathematical facts concerning semantic conservativity provide no insight into whether extended induction or compositional truth is stronger:

Fact 127 (Cieśliński [Cie17] Theorem 6.0.13, Corollary 7.0.6, Wicsło [Wci17] Proposition 89.)**.** *Let* PA *be the base theory,*

- $\mathsf{TT}[\mathsf{PA}]$ *is semantically conservative over* PA.
- $\mathsf{TT}^{+}[\mathsf{PA}]$, $\mathsf{CT}[\mathsf{PA}]$, $\mathsf{CT}^{+}[\mathsf{PA}]$ *are not semantically conservative over* PA.

The situation is exactly the opposite of the facts concerning syntactical conservativity, where only the strongest theory is not syntactically conservative over PA. We need a non-ad-hoc measure of logical strength that distinguishes one from the other, just like interpretability in the case of syntactical conservativity. The obvious candidate is recursive saturation. Since this notion is not yet widely known among philosophers, we list several historical remarks to argue briefly that it is indeed a natural measure of logical strength.

The notion of a recursively saturated model has multiple origins. It originates from the study of admissible sets with urelements and admissible fragments of $\mathcal{L}_{\omega,\omega}$ (Barwise & Schlipf [BS76], p. 531), but is entangled with questions in model theory from the early days where the concept was invented. Recursive saturation, and saturation of certain types, more generally, helps answer questions of the expandability of models of arithmetic to models of the stronger theories. An early attempt is made by Robinson [Rob63], where he gives a non-diagonal proof of Tarski's undefinability theorem using the unrealizability of a partial type in a given model. This is followed by Ehrenfeucht and Kreisel [EK66], where they

> "[...] gave an example of nonexpandability by means of an argument closely allied to that cited above of Robinson: A truth definition for arithmetic entails the existence of much larger elements than would necessarily exist in a model not having such a truth definition." (Smoryński [Smo81], p.259)

Lachlan's theorem has a similar explanation. In simple terms, recursive saturation stipulates the existence of large elements — a model is recursively saturated if for every recursive type there is a large element in the model that satisfies the stipulations of the corresponding recursive type. Therefore, Lachlan's theorem states that compositional truth entails the existence of much larger elements than would necessarily exist in models without such a compositional truth definition. Since recursive saturation is closely connected to the investigations of model-theoretic behavior of truth definitions, we see no reason for overlooking related mathematical facts. Summarising facts stated in Corollary 124 and Theorem 125, we have:

Base: PA	No Extended Induction	Extended Induction
Add Biconditionals	Not Recursively Saturated	Not Recursively Saturated
Add Compositional Axioms	Recursively Saturated	Recursively Saturated

Base: $T \supseteq \mathsf{PA}^-$	No Extended Induction	Extended Induction
Add Biconditionals	Not Recursively Saturated	Unclear
Add Compositional Axioms	Recursively Saturated	Recursively Saturated

Compositional truth still contributes more strength than extended induction viewed from a semantic perspective. However, unlike Fact 126, which gives a "syntactic" characterization of the strength of CT, we don't have a parallel semantic characterization. It is not yet known if every recursively saturated model of PA^- is expandable to a model of $\mathsf{CT}[\mathsf{PA}^-]$.

Conclusion, Future Directions

Let's take stock. In this thesis, we proved two of the main results on the syntactic and semantic conservativity of PA^-. On the syntactical side, we first show PA^-'s ability to encode syntax. PA^- carries an arithmetization of syntax, with all the desirable properties provable in a cut. Then we proved that compositional satisfaction and compositional truth without extended induction $\mathsf{CS}[\mathsf{PA}^-]$, $\mathsf{CT}[\mathsf{PA}^-]$ are both syntactically conservative over PA^-. We also prove, in the Appendix, that truth is equivalent to extensional satisfaction. On the semantical side, we proved Lachlan's theorem for PA^-: every strongly non-standard model of PA^- that carries a compositional satisfaction class is recursively saturated. Our proof strategy also explicates the landscape of non-standard models for weak arithmetic. Together with Łełyk's result that each countable sequential theory has a countable model that is not recursively saturated, $\mathsf{CT}[\mathsf{PA}^-]$ is not semantically conservative for PA^-. Hence the landscape for PA transfers smoothly to PA^-.

We also make some scattered philosophical remarks on the way. We motivated the importance of model-theoretical investigations for studying truth theories, which philosophers usually ignore. Our technical results regarding recursive saturation nicely strengthen Richard Heck's project of disentangling the contribution of compositional truth and extended induction.

The following technical questions remain open:

- Is there a semantic characterization of the strength of compositional truth in PA^-? Does every recursively saturated model of PA^- carry a compositional truth class?
- What is the exact landscape of non-standard models for weak arithmetic? In particular, is there a weakly non-standard model that is not strongly non-standard?

And, regrettably, we didn't have time to address the following philosophical question:

- How does semantic conservativity relate to the existing debate on the conservativity argument? In particular, how can deflationism make sense of the result that $\mathsf{TT}^+[T]$ is not semantically conservative?

We hope to address these questions on a future occasion.

Bibliography

[ASW23] Bradley Armour-Garb, Daniel Stoljar, and James Woodbridge. "Deflationism About Truth". In: *The Stanford Encyclopedia of Philosophy*. Edited by Edward N. Zalta and Uri Nodelman. Summer 2023. Metaphysics Research Lab, Stanford University, 2023 (cited on page 2).

[BS76] Jon Barwise and John Schlipf. "An introduction to recursively saturated and resplendent models". In: *The Journal of Symbolic Logic* 41.2 (1976), pages 531–536. https://doi.org/10.2307/2272253 (cited on page 47).

[CGR18] Ivano Ciardelli, Jeroen Groenendijk, and Floris Roelofsen. *Inquisitive Semantics*. Edited by J. A. G. Groenendijk and Floris Roelofsen. Oxford, England: Oxford University Press, 2018 (cited on page 16).

[Cie17] Cezary Cieśliński. *The Epistemic Lightness of Truth: Deflationism and its Logic*. Cambridge: Cambridge University Press, 2017 (cited on pages 12, 16, 38, 39, 46, 47).

[Dim89] C. Dimitracopoulos. "Overspill and Fragments of Arithmetic". In: *Archive for Mathematical Logic* 28.3 (1989), pages 173–179. https://doi.org/10.1007/bf01622877 (cited on page 10).

[EK66] A. Ehrenfeucht and G. Kreisel. "Strong Models of Arithmetic". In: *Bull. de l'Acad. Polonaise des Sciences* XV (1966), pages 107–110 (cited on page 47).

[EV15] Ali Enayat and Albert Visser. "New Constructions of Satisfaction Classes". In: *Unifying the Philosophy of Truth*. Edited by Theodora Achourioti, Henri Galinon, José Martínez Fernández, and Kentaro Fujimoto. Dordrecht: Springer Netherlands, 2015, pages 321–335. ISBN: 978-94-017-9673-6. https://doi.org/10.1007/978-94-017-9673-6_16. https://doi.org/10.1007/978-94-017-9673-6_16 (cited on pages a, 5, 17, 28).

[Fef91] Solomon Feferman. "Reflecting on Incompleteness". In: *Journal of Symbolic Logic* 56.1 (1991), pages 1–49. https://doi.org/10.2307/2274902 (cited on pages 3, 24).

[Fuj19] Kentaro Fujimoto. "Deflationism Beyond Arithmetic". In: *Synthese* 196.3 (2019), pages 1045–1069. https://doi.org/10.1007/s11229-017-1495-8 (cited on page 6).

[Gup93] Anil Gupta. "A Critique of Deflationism". In: *Philosophical Topics* 21.1 (1993), pages 57–81. https://doi.org/10.5840/philtopics199321218 (cited on page 3).

[Hal10] Volker Halbach. *Axiomatic Theories of Truth*. Cambridge, England: Cambridge University Press, 2010 (cited on pages 10, 12, 26, 38, 39).

[Hal99] Volker Halbach. "Disquotationalism and Infinite Conjunctions". In: *Mind* 108.429 (1999), pages 1–22. https://doi.org/10.1093/mind/108.429.1 (cited on page 37).

[Hec18] Richard Heck. "The Logical Strength of Compositional Principles". In: *Notre Dame Journal of Formal Logic* 59.1 (2018), pages 1–33. https://doi.org/10.1215/00294527-2017-0011 (cited on pages a, 3, 4, 13, 46).

[HP93] Petr Hájek and Pavel Pudlák. *Metamathematics of First-Order Arithmetic*. Springer Verlag, 1993 (cited on pages 19, 22).

[Jeř12] Emil Jeřábek. "Sequence encoding without induction". In: *Mathematical Logic Quarterly* 58.3 (Apr. 2012), pages 244–248. ISSN: 1521-3870. https://doi.org/10.1002/malq.201200013. http://dx.doi.org/10.1002/malq.201200013 (cited on pages 5, 11, 17, 18).

[Kay16] Richard Kaye. *Reflections on nonstandard satisfaction*. en. June 2016 (cited on page 38).

[Kay91] Richard Kaye. *Models of Peano Arithmetic*. Clarendon Press, 1991 (cited on pages a, 5, 8, 22, 39, 42, 44).

[Ket99] Jeffrey Ketland. "Deflationism and Tarski's Paradise". In: *Mind* 108.429 (1999), pages 69–94. https://doi.org/10.1093/mind/108.429.69 (cited on page 2).

[Lac81] Alistair H. Lachlan. "Full Satisfaction Classes and Recursive Saturation". In: *Canadian Mathematical Bulletin* 24.1 (1981), pages 295–97 (cited on pages a, 5).

[LN13] Graham E. Leigh and Carlo Nicolai. "Axiomatic Truth, Syntax and Metatheoretic Reasoning". In: *Review of Symbolic Logic* 6.4 (2013), pages 613–636. https://doi.org/10.1017/s1755020313000233 (cited on page 4).

[ŁW17] Mateusz Łełyk and Bartosz Wcisło. "Models of Weak Theories of Truth". In: *Archive for Mathematical Logic* 56.5 (2017), pages 453–474. https://doi.org/10.1007/s00153-017-0531-1 (cited on page 46).

[Pud85] Pavel Pudlák. "Cuts, Consistency Statements and Interpretations". In: *The Journal of Symbolic Logic* 50.2 (1985), pages 423–441. ISSN: 00224812. http://www.jstor.org/stable/2274231 (visited on 05/07/2025) (cited on page 11).

[Rob63] Abraham Robinson. "On Languages Which are Based on Non-Standard Arithmetic". In: *Nagoya Mathematical Journal* 22 (1963), pages 83–117. https://doi.org/10.1017/S0027763000011065 (cited on page 47).

[Sha98] Stewart Shapiro. "Proof and Truth". In: *Journal of Philosophy* 95.10 (1998), pages 493–521. https://doi.org/10.5840/jphil199895102 (cited on pages 2, 3).

[Smo81] C. Smoryński. "Recursively Saturated Nonstandard Models of Arithmetic". In: *The Journal of Symbolic Logic* 46.2 (1981), pages 259–286. ISSN: 00224812. http://www.jstor.org/stable/2273620 (visited on 09/04/2025) (cited on page 47).

[Tar53] Alfred Tarski. "I: A General Method in Proofs of Undecidability". In: *Undecidable Theories*. Edited by Alfred Tarski. Volume 13. Studies in Logic and the Foundations of Mathematics. Elsevier, 1953, pages 1–34. https://doi.org/https://doi.org/10.1016/S0049-237X(09)70292-7. https://www.sciencedirect.com/science/article/pii/S0049237X09702927 (cited on page 15).

[Tar56] Alfred Tarski. "The Concept of Truth in Formalized Languages". In: *Logic, Semantics, Metamathematics*. Edited by Alfred Tarski. Clarendon Press, 1956, pages 152–278 (cited on pages 2, 3, 13).

[Vis08] Albert Visser. "Pairs, Sets and Sequences in First-Order Theories". In: *Archive for Mathematical Logic* 47.4 (2008), pages 299–326. https://doi.org/10.1007/s00153-008-0087-1 (cited on page 11).

[Wax17] Daniel Waxman. "Deflationism, Arithmetic, and the Argument From Conservativeness". In: *Mind* 126.502 (2017), pages 429–463. https://doi.org/10.1093/mind/fzv182 (cited on page 5).

[Wci17] Bartosz Wcisło. "Understanding the Strength of the Compositional Truth". PhD thesis. Uniwersytet Warszawski, 2017 (cited on pages 5, 47).

[Wol03] Jan Woleński. "Truth and Satisfaction by the Empty Sequence". In: *Philosophical Dimensions of Logic and Science*. Edited by A. Rojszczak, J. Cachro, and G. Kurczewski. Kluwer Academic Publishers, 2003, pages 267–276 (cited on page 14).

Appendix: Truth is Equivalent to Extensional Satisfaction

In this appendix, we give full proofs of Remark 93 and Proposition 94, and some other results surrounding the two remarks.

Truth is Equivalent to Extensional Satisfaction in PA.

Remark 128. PA *proves that* $(\varphi_1, \alpha_1) \sim (\varphi_2, \alpha_2)$ *is an equivalence relation.*

Proof. Consider an arbitrary model of $\mathcal{M} \vDash \mathsf{PA}$. It is easy to see that $(\varphi_1, \alpha_1) \sim (\varphi_2, \alpha_2)$ is reflexive and symmetric. To see that it is transitive, suppose that $\mathcal{M} \vDash (\varphi_1, \alpha_1) \sim (\varphi_2, \alpha_2) \sim (\varphi_3, \alpha_3)$. There exist χ_1, γ_1, and $\sigma_1, \sigma_2 : \mathsf{FV}(\chi_1) \to \mathsf{Var}$ where

$$\varphi_1 = \chi_1 * \sigma_1,\ \gamma_1 = \alpha_1 \circ \sigma_1; \tag{3.2}$$

$$\varphi_2 = \chi_1 * \sigma_2,\ \gamma_1 = \alpha_2 \circ \sigma_2, \tag{3.3}$$

and $\chi_2, \gamma_2, \sigma_2', \sigma_3' : \mathsf{FV}(\chi_2) \to \mathsf{Var}$ where

$$\varphi_2 = \chi_2 * \sigma_2',\ \gamma_2 = \alpha_2 \circ \sigma_2'; \tag{3.4}$$

$$\varphi_3 = \chi_2 * \sigma_3',\ \gamma_2 = \alpha_3 \circ \sigma_3'. \tag{3.5}$$

We show that $\mathcal{M} \vDash (\chi_1, \gamma_1) \sim (\chi_2, \gamma_2)$. First, obtain χ from φ_2 by making sure that for any two free occurrences of variables v_1 and v_2, $v_1 \neq v_2$. We denote the substitution function as σ. (This is an abuse of notation since we are substituting occurrences of variables but not variables.) Define $\tau_1 = (\sigma_2 \circ \sigma)^{-1}$ and $\tau_1 = (\sigma_2' \circ \sigma)^{-1}$. Both are well defined in $\mathcal{M}$. Modify variable assignments γ of χ from φ_2 accordingly. Therefore we have

$$\chi_1 = \chi * \tau_1,\ \gamma = \gamma_1 \circ \tau_1; \tag{3.6}$$

$$\chi_2 = \chi * \tau_2,\ \gamma = \gamma_2 \circ \tau_2. \tag{3.7}$$

It is easy to check that $\chi, \gamma, (\sigma_1 \circ \tau_1), (\sigma_3' \circ \tau_2)$ are the witnesses of $\mathcal{M} \vDash (\varphi_1, \alpha_1) \sim (\varphi_3, \alpha_3)$. □

Proposition 129. *Suppose* $\mathcal{M} \vDash \mathsf{PA}$*,* T *is a full truth class on* $\mathcal{M}$*, and* S *is an extensional full satisfaction class on* $\mathcal{M}$*. Define*

- *$S(T)$ is an extensional satisfaction class on $\mathcal{M}$, where $S(T)$ is defined as the collection of ordered pairs (φ, α) such that $\varphi(\mathsf{c} \circ \alpha) \in T$.*
- *$T(S)$ is a truth class on $\mathcal{M}$, where $T(S)$ is defined as the collection of $\varphi \in \mathcal{L}^+_{\mathsf{PA}}$ such that for some $\psi \in \mathcal{L}^+_{\mathsf{PA}}$ and some assignment α for ψ, $\varphi = \psi(c \circ \alpha)$ and $(\psi, \alpha) \in S$.*
- *$S(T(S)) = S$, and $T(S(T)) = T$.*

The routine but laborious proof for PA is omitted in the original paper; we first retrieve it.

Lemma 130 (PA). *Consider $\mathcal{M} \vDash \mathsf{PA}$. For all $\varphi, \psi \in \mathsf{Form}^{\mathcal{M}}$, $\alpha, \beta \in \mathsf{Asn}^{\mathcal{M}}$, $\mathcal{M} \vDash \varphi(\mathsf{c} \circ \alpha) = \psi(\mathsf{c} \circ \beta)$ if and only if $\mathcal{M} \vDash (\varphi, \alpha) \sim (\psi, \beta)$.*

Proof. Our proof is by induction on φ. We treat ψ, α, β as parameters. Base case: $\varphi = (R, (t_1, \ldots, t_n))$. Here t_is are meta-variables referring to variables in $\mathcal{M}$. Suppose we have $\psi \in \mathsf{Form}^{\mathcal{M}}$ and $\beta \in \mathsf{Asn}^{\mathcal{M}}$ where $\mathcal{M} \vDash \varphi(\mathsf{c} \circ \alpha) = \psi(\mathsf{c} \circ \beta)$. By unique reading and since substitution preserves the structure of formulas, ψ must be of the shape $(R, (t'_1, \ldots, t'_n))$. We define $\chi := (R, (s_1, \ldots, s_n))$ where $(s_1, \ldots, s_n)$ are pairwise distinct. Let $\sigma_\varphi, \sigma_\psi : \mathsf{FV}(\chi) \to \mathsf{Var}$ be such that $\sigma_\varphi(s_i) = t_i$ and $\sigma_\psi(s_i) = t'_i$. Let $\gamma(s_i) = \alpha(t_i)$. Since $\mathcal{M} \vDash \varphi(\mathsf{c} \circ \alpha) = \psi(\mathsf{c} \circ \beta)$, for all $i < n$, $\mathcal{M} \vDash \alpha(t_i) = \beta(t'_i)$. Therefore $\mathcal{M} \vDash \gamma(s_i) = \alpha(\sigma_\varphi(t_i)) = \beta(\sigma_\psi(t'_i))$, so $\gamma = \alpha \circ \sigma_\varphi$ and $\gamma = \beta \circ \sigma_\psi$.

Since everything in the proof concerns objects inside a model $\mathcal{M}$, in the following cases we omit "$\mathcal{M} \vDash$" before formulas. For the inductive case:

- $\varphi = (\neg, \varphi_1)$. Suppose there is a $\psi \in \mathsf{Form}^{\mathcal{M}}$ and $\beta \in \mathsf{Asn}^{\mathcal{M}}$ where $\mathcal{M} \vDash \varphi(\mathsf{c} \circ \alpha) = \psi(\mathsf{c} \circ \beta)$. Again, ψ must be of the shape $(\neg, \psi_1)$. So $(\neg, (\varphi_1(\mathsf{c} \circ \alpha))) = (\neg, (\psi_1(\mathsf{c} \circ \beta)))$, $\varphi_1(\mathsf{c} \circ \alpha) = \psi_1(\mathsf{c} \circ \beta)$. By the inductive hypothesis, $(\varphi_1, \alpha) \sim (\psi_1, \beta)$. There exists χ_1, γ and $\sigma_\varphi, \sigma_\psi : \mathsf{FV}(\chi_1) \to \mathsf{Var}$ such that $\varphi_1 = \chi_1 * \sigma_\varphi$; $\psi_1 = \chi_1 * \sigma_\psi$ and $\alpha = \gamma \circ \sigma_\varphi$; $\beta = \gamma \circ \sigma_\psi$. We define $\chi = (\neg, \chi_1)$. It is easy to see that $\varphi = (\neg, \varphi_1) = (\neg, (\chi_1 * \sigma_\varphi)) = (\neg, \chi_1) * \sigma_\varphi = \chi * \sigma_\varphi$, and similarly for ψ. Since $\mathsf{FV}((\neg, \varphi_1)) = \mathsf{FV}(\varphi_1)$, α is also an assignment for φ, similarly β is an assignment for ψ. Therefore there exist $\chi, \gamma, \sigma_\varphi, \sigma_\psi$ witnessing $(\varphi, \alpha) \sim (\psi, \beta)$.
- $\varphi = (\vee, (\varphi_1, \varphi_2))$. Suppose there is a $\psi \in \mathsf{Form}^{\mathcal{M}}$ and $\beta \in \mathsf{Asn}^{\mathcal{M}}$ where $\mathcal{M} \vDash \varphi(\mathsf{c} \circ \alpha) = \psi(\mathsf{c} \circ \beta)$. $(\vee, (\varphi_1, \varphi_2))(\mathsf{c} \circ \alpha) = (\vee, (\psi_1, \psi_2))(\mathsf{c} \circ \beta)$, thus $\varphi_1(\mathsf{c} \circ \alpha) = \psi_1(\mathsf{c} \circ \beta)$ and $\varphi_2(\mathsf{c} \circ \alpha) = \psi_2(\mathsf{c} \circ \beta)$. For clarity, we can further restrict α and β to $\mathsf{FV}(\varphi_i)$ and $\mathsf{FV}(\psi_i)$, $i = 1, 2$. We know that $(\varphi_1, \alpha \restriction \mathsf{FV}(\varphi_1)) \sim (\psi_1, \beta \restriction \mathsf{FV}(\psi_1))$ and $(\varphi_2, \alpha \restriction \mathsf{FV}(\varphi_2)) \sim (\psi_2, \beta \restriction \mathsf{FV}(\psi_2))$. There exist χ_1, γ_1, and $\sigma_{\varphi_1}, \sigma_{\psi_1} : \mathsf{FV}(\chi_1) \to \mathsf{Var}$ where

  (3.8) $$\varphi_1 = \chi_1 * \sigma_{\varphi_1},\ \gamma_1 = (\alpha \restriction \mathsf{FV}(\varphi_1)) \circ \sigma_{\varphi_1};$$

  (3.9) $$\psi_1 = \chi_1 * \sigma_{\psi_1},\ \gamma_1 = (\beta \restriction \mathsf{FV}(\psi_1)) \circ \sigma_{\psi_1},$$

  and $\chi_2, \gamma_2, \sigma_{\varphi_2}, \sigma_{\psi_2} : \mathsf{FV}(\chi_2) \to \mathsf{Var}$ where

  (3.10) $$\varphi_2 = \chi_2 * \sigma_{\varphi_2},\ \gamma_2 = (\alpha \restriction \mathsf{FV}(\varphi_2)) \circ \sigma_{\varphi_2};$$

  (3.11) $$\psi_2 = \chi_2 * \sigma_{\psi_2},\ \gamma_2 = (\beta \restriction \mathsf{FV}(\psi_2)) \circ \sigma_{\psi_2}.$$

  There may be overlapping variables in χ_1 and χ_2 but σ_φ and σ_ψ map them to different variables. But there always exists χ'_1, χ'_2, and σ's with $\mathsf{FV}(\chi_1) \cap \mathsf{FV}(\chi_2) = \emptyset$ that satisfy (1.4-7). When $v_1 \in \mathsf{FV}(\chi_1)$ and $v_2 \in \mathsf{FV}(\chi_2)$ is such that $v_1 = v_2$, α-convert v_2 to fresh v_3. Define $\sigma'(v_3) = \sigma(v_2)$ and $\gamma'(v_3) = \gamma(v_2)$ accordingly. It is easy

to check that $\varphi_i = \chi_i' * \sigma'_{\varphi_i}$, $\gamma_i' = (\alpha \restriction \mathsf{FV}(\varphi_1)) \circ \sigma'_{\varphi_1}$, and similarly for ψ_is. Finally, define $\chi = (\vee, (\chi_1', \chi_2'))$, $\sigma_\varphi = \sigma'_{\varphi_1} * \sigma'_{\varphi_2}$, $\sigma_\varphi = \sigma'_{\psi_1} * \sigma'_{\psi_2}$ and $\gamma = \gamma_1 \restriction \mathsf{FV}(\chi_1) \cup \gamma_2 \restriction \mathsf{FV}(\chi_2)$.

- $\varphi = (\exists, (x, \varphi_1))$. Suppose there is a $\psi \in \mathsf{Form}^{\mathcal{M}}$ and $\beta \in \mathsf{Asn}^{\mathcal{M}}$ where $\varphi(\mathsf{c} \circ \alpha) = \psi(\mathsf{c} \circ \beta)$, i.e, $(\exists, (x, \varphi_1(x)))(\mathsf{c} \circ \alpha) = (\exists, (x, \psi_1(x)))(\mathsf{c} \circ \beta)$. To obtain α' and β' where $\mathsf{Asn}(\alpha', \varphi_1)$ and $\mathsf{Asn}(\beta', \psi_1)$ and $\varphi_1(\mathsf{c} \circ \alpha') = \psi_1(\mathsf{c} \circ \beta')$, define $\alpha' \supseteq \alpha, \beta' \supseteq \beta$, and $\alpha'(x) = \beta'(x)$. By the inductive hypothesis, there exist $\chi_1, \gamma', \sigma_\varphi, \sigma_\psi : \mathsf{FV}(\chi_1) \to \mathsf{Var}$ such that

(3.12) $$\varphi_1 = \chi_1 * \sigma_\varphi,\ \psi_1 = \chi_1 * \sigma_\psi;$$

(3.13) $$\gamma' = \alpha' \circ \sigma_\varphi,\ \gamma' = \beta' \circ \sigma_\psi.$$

To prevent undesirable binding, define σ_φ and σ_ψ such that both map x to itself and nothing else maps to x, so $x \in \mathsf{FV}(\chi_1)$. Define $\chi = (\exists, (x, \chi_1))$, $\gamma = \gamma' \restriction \mathsf{FV}(\chi)$, $\sigma'_\varphi = \sigma_\varphi \restriction \mathsf{FV}(\chi)$, $\sigma'_\psi = \sigma_\psi \restriction \mathsf{FV}(\chi)$ (which removes x from the domain). Observe that $(\exists, (x, \chi_1(x))) * \sigma'_\varphi = (\exists, (x, \chi_1(x) * \sigma'_\varphi))$ and $(\exists, (x, \varphi_1(x) * \sigma'_\varphi)) = (\exists, (x, \varphi_1(x))) * \sigma'_\varphi$. $(\exists, (x, \chi_1(x) * \sigma'_\varphi)) = (\exists, (x, \varphi_1(x) * \sigma'_\varphi))$ since $\varphi_1 = \chi_1 * \sigma_\varphi$; $\psi_1 = \chi_1 * \sigma_\psi$, and σ_φ agrees with σ'_φ except for x. Similarly for ψ. Since $\alpha' = \gamma' \circ \sigma_\varphi$, $\alpha' \restriction \mathsf{FV}(\chi) = (\gamma' \circ \sigma_\varphi) \restriction \mathsf{FV}(\chi)$, since $\sigma_\varphi(v) = x$ iff $v = x$,
$(\gamma' \circ \sigma_\varphi) \restriction \mathsf{FV}(\chi) = (\gamma' \restriction \mathsf{FV}(\chi)) \circ \sigma'_\varphi = \gamma \circ \sigma'_\varphi = \alpha$. Similarly for β. Therefore $(\varphi, \alpha) \sim (\psi, \beta)$.

This exhausts all cases. □

Remark 131. *From the proof, we also observe that $\sim$ has the following properties when interpreted in* PA*:*

- *if* $((\neg, \varphi_0), \alpha_0) \sim ((\neg, \varphi_1), \alpha_1)$*, then* $(\varphi_0, \alpha_0) \sim (\varphi_1, \alpha_1)$.
- *if* $(\vee, (\varphi_0, \varphi_1)), \alpha) \sim ((\vee, (\varphi_0', \varphi_1')), \alpha')$*, then* $(\varphi_0, \alpha \restriction \mathsf{FV}(\varphi_0)) \sim (\varphi_0', \alpha' \restriction \mathsf{FV}(\varphi_0'))$ *and* $(\varphi_1, \alpha \restriction \mathsf{FV}(\varphi_1)) \sim (\varphi_1', \alpha' \restriction \mathsf{FV}(\varphi_1'))$.
- *if* $\varphi = (\exists, (t, \psi))$ *and* $\varphi' = (\exists, (t', \psi'))$ *and* $(\varphi, \alpha) \sim (\varphi', \alpha')$*, then* $t = t'$ *and for some* e*,* $(\varphi, \alpha[t : e]) \sim (\varphi', \alpha'[t : e])$.
- *if* $(\varphi, \alpha) \sim (\psi, \beta)$*, then* $c(\varphi) = c(\psi)$*. (where* c *is defined as in definition 67)*

The proof of the first two items in Proposition 94 is similar. We only present the more complex case of showing $T(S)$ is a truth class.

Lemma 132 (PA)**.** *$T(S)$ is a truth class.*

Proof. We check that $(\mathcal{M}, T(S)) \vDash \mathsf{PA}^{\mathsf{FS}}$. There are several cases:

- To show $(\mathcal{M}, T(S)) \vDash \forall x(\mathsf{T}(x) \to \mathsf{Sent}(x))$, consider arbitrary $x \in T(S)$. There exists ψ, α where $\varphi = \psi(\mathsf{c} \circ \alpha)$ and $(\psi, \alpha) \in S$. Since $(\psi, \alpha) \in S$ means $\mathsf{Asn}(\alpha, \psi)$, for every $v \in \mathsf{FV}(\psi)$, exists $a \in M$ where $\alpha(v) = a$. So $\mathsf{FV}(\varphi) = \mathsf{FV}(\psi(\mathsf{c} \circ \alpha)) = \emptyset$, $\varphi \in \mathsf{Sent}$.
- To show $(\mathcal{M}, T(S)) \vDash (x = (R, (\mathsf{c}_{y_0}, \ldots, c_{y_{n-1}}))) \to (R(y_0, \ldots, y_{n-1}) \leftrightarrow \mathsf{T}(x))$,
let $\varphi = (R, (\mathsf{c}_{y_0}, \ldots, \mathsf{c}_{y_{n-1}}))$.
$\Rightarrow$: Suppose $\varphi \in T(S)$, there exist ψ, α where $\varphi = \psi(\mathsf{c} \circ \alpha)$ and $(\psi, \alpha) \in S$. Therefore ψ is of the shape $Rv_0 \ldots v_{n-1}$, and $\alpha(v_i)$=y_i. By definition of a satisfaction class, $R(y_0, \ldots, y_{n-1})$.
$\Leftarrow$: Suppose $R(y_0, \ldots, y_{n-1})$, let $\psi = (R, (v_0 \ldots v_{n-1}))$ and $\alpha \in \mathsf{Asn}(\psi)$ where $\alpha(v_i) = y_i$. By definition of satisfaction class, $(\psi, \alpha) \in S$. Observe that $\varphi = \psi(\mathsf{c} \circ \alpha)$, so $\varphi \in T(S)$.

- To show $(\mathcal{M}, T(S)) \vDash \forall x(x = (\neg, y) \to (\mathsf{T}(x) \leftrightarrow \neg\mathsf{T}(y)))$, let $\varphi = (\neg, \varphi_1)$.
  $\Rightarrow$ Suppose $\varphi \in T(S)$. Then there exist ψ, α where $\varphi = \psi(\mathsf{c} \circ \alpha)$, $(\psi, \alpha) \in S$. ψ must be of the shape $(\neg, \psi_1)$ where $\varphi_1 = \psi_1(\mathsf{c} \circ \alpha)$. Since S is a satisfaction class, $(\psi_1, \alpha) \notin S$. Suppose, for a contradiction, that $\varphi_1 \in T(S)$. There exist ψ', α' where $\varphi_1 = \psi'(\mathsf{c} \circ \alpha')$ and $(\psi', \alpha') \in S$. We observe that $\psi_1(\mathsf{c} \circ \alpha) = \varphi_1 = \psi'(\mathsf{c} \circ \alpha')$. By lemma 130 and S extensional, $(\psi', \alpha') \in S$ iff $(\psi_1, \alpha) \in S$, but $(\psi_1, \alpha) \notin S$, contradiction.
  This is the crucial point where extensionality is used: if S were not extensional, the commutation with negation would fail.
  $\Leftarrow$: Suppose $\varphi_1 \notin T(S)$. Therefore $\forall \psi, \alpha(\varphi_1 = \psi(\mathsf{c} \circ \alpha) \to (\psi, \alpha) \notin S)$. It is easy to see that there exists ψ, α where $\varphi = \psi(\mathsf{c} \circ \alpha)$. Consider such pair, so $(\psi, \alpha) \notin S$, by definition of satisfaction class $((\neg, \psi), \alpha) \in S$, observe that $(\neg, \varphi_1) = (\neg, \psi)(\mathsf{c} \circ \alpha)$, so $(\neg, \varphi_1) \equiv \varphi \in T(S)$.

- To show $(\mathcal{M}, T(S)) \vDash \forall x(x = (\vee, (y_1, y_2)) \to (\mathsf{T}(x) \leftrightarrow \mathsf{T}(y_1) \vee \mathsf{T}(y_2)))$, let $\varphi = (\vee, (\varphi_1, \varphi_2))$.
  $\Rightarrow$: Suppose $\varphi \in T(S)$. There exist ψ, α with $\varphi = \psi(\mathsf{c} \circ \alpha)$ and $\psi, \alpha \in S$. ψ must be of the shape $(\vee, (\psi_1, \psi_2))$, where $\varphi_1 = \psi_1(\mathsf{c} \circ \alpha) = \psi_1(\mathsf{c} \circ (\alpha \restriction \mathsf{FV}(\psi_1)))$ and $\varphi_2 = \psi_2(\mathsf{c} \circ \alpha) = \psi_2(\mathsf{c} \circ \alpha \restriction \mathsf{FV}(\psi_2)))$. Since S is a satisfaction class, $(\psi, \alpha) \in S$ entails $(\psi_1, \alpha \restriction \mathsf{FV}(\psi_1)) \in S$ or $(\psi_2, \alpha \restriction \mathsf{FV}(\psi_2)) \in S$. So $\varphi_1 \in T(S)$ or $\varphi_2 \in T(S)$.
  $\Leftarrow$: Suppose $\varphi_1 \in T(S)$ or $\varphi_2 \in T(S)$. W.l.o.g, suppose $\varphi_1 \in T(S)$. There exists ψ_1, α_1 with $\varphi_1 = \psi_1(\mathsf{c} \circ \alpha_1)$ and $(\psi_1, \alpha_1) \in S$. Consider any ψ', α where $\varphi = \psi'(\mathsf{c} \circ \alpha)$. Since $\varphi = (\vee, (\varphi_1, \varphi_2))$, ψ' must be of the shape $(\vee, (\psi_1', \psi_2'))$. So $\psi'(\mathsf{c} \circ \alpha) = (\vee, (\psi_1', \psi_2'))(\mathsf{c} \circ \alpha) = (\vee, (\psi_1'(\mathsf{c} \circ \alpha), \psi_2'(\mathsf{c} \circ \alpha)) = (\vee, \psi_1'(\mathsf{c} \circ (\alpha \restriction \mathsf{FV}(\psi_1'))), \psi_2'(\mathsf{c} \circ (\alpha \restriction \mathsf{FV}(\psi_2'))))$. By unique reading, $\varphi_1 = \psi_1'(\mathsf{c} \circ (\alpha \restriction \mathsf{FV}(\psi_1')))$ and $\varphi_2 = \psi_2'(\mathsf{c} \circ (\alpha \restriction \mathsf{FV}(\psi_2')))$. Therefore $\psi_1'(\mathsf{c} \circ (\alpha \restriction \mathsf{FV}(\psi_1'))) = \psi_1(\mathsf{c} \circ \alpha_1)$, by lemma 130, $(\psi_1', \alpha \restriction \mathsf{FV}(\psi_1')) \in S$ iff $(\psi_1, \alpha_1) \in S$. So $(\psi_1', \alpha \restriction \mathsf{FV}(\psi_1')) \in S$. By definition of a satisfaction class, $((\vee, (\psi_1', \psi_2')), \alpha) \in S$. So $\varphi \in T(S)$.

- To show $(\mathcal{M}, T(S)) \vDash \forall x(x = (\exists, (v, y)) \to (\mathsf{T}(x) \leftrightarrow \exists z\mathsf{T}(y(\mathsf{c}_z))))$, consider $\varphi = (\exists, (v, \varphi_1(v)))$.
  $\Rightarrow$: Suppose $\varphi \in T(S)$. There exist ψ, α where $\varphi = \psi(\mathsf{c} \circ \alpha)$ and $(\psi, \alpha) \in S$. ψ is of the shape $(\exists, (v, \psi_1(v)))$, so $\varphi_1(v) = \psi_1(v)(\mathsf{c} \circ \alpha)$. Since $(\psi, \alpha) \in S$, by definition of satisfaction classes, there exists $\alpha' \supseteq \alpha$ where $(\psi_1(v), \alpha') \in S$. We consider $\psi_1(\mathsf{c} \circ \alpha')$. let c_z be that $\alpha'(v) = z$. Observe that $\varphi_1(\mathsf{c}_z) = \psi_1(\mathsf{c} \circ \alpha')$. Therefore $\varphi_1(\mathsf{c}_z) \in T(S)$.
  $\Leftarrow$: Suppose $(\exists, (z, \varphi_1(\mathsf{c}_z))) \in T(S)$. There exist ψ_1, α where $\varphi(\mathsf{c}_z) = \psi(\mathsf{c} \circ \alpha)$ and $(\psi_1, \alpha) \in S$. Consider ψ', α' where $\varphi = \psi'(\mathsf{c} \circ \alpha')$, we show that $(\psi', \alpha') \in S$. ψ' must be of the shape $(\exists, (v, \psi_1'(v)))$. Therefore $\psi'(\mathsf{c} \circ \alpha') = (\exists, (v, \psi_1'(v)))(\mathsf{c} \circ \alpha') = (\exists, (v, \psi_1'(v)(\mathsf{c} \circ \alpha'))) = (\exists, (v, \psi_1'(\mathsf{c} \circ \alpha')(v)))$. Define $\alpha'' = \alpha' \cup \{(v, z)\}$, so $\varphi_1(\mathsf{c}_z) = \psi_1'(\mathsf{c} \circ \alpha'')$. So $\psi_1'(\mathsf{c} \circ \alpha'') = \varphi_1(\mathsf{c}_z) = \psi_1(\mathsf{c} \circ \alpha)$. By lemma 130, $(\psi_1', \alpha'') \in S$ iff $(\psi_1, \alpha) \in S$, so $(\psi_1', \alpha'') \in S$. By definition of satisfaction class, $((\exists, (v, \psi_1'(v))), \alpha') \in S$. So $\varphi \in T(S)$.

□

Lemma 133 (PA). *Given extensional satisfaction class S and truth class T, $S(T(S)) = S$ and $T(S(T)) = T$.*

Proof. $S \subseteq S(T(S))$: consider arbitrary $(\varphi, \alpha) \in S$, by definition $(\varphi, \alpha) \in S(T(S))$ iff there is $\psi \in \mathcal{L}_{\mathsf{PA}}^+$, $\beta \in \mathsf{Asn}$ where $\varphi(\mathsf{c} \circ \alpha) = \psi(\mathsf{c} \circ \beta)$ and $(\psi, \beta) \in S$. Take $\psi = \varphi$ and $\beta = \alpha$. $S(T(S)) \subseteq S$: suppose $(\varphi, \alpha) \in S(T(S))$, then there exists $\psi \in \mathcal{L}_{\mathsf{PA}}^+$, $\beta \in \mathsf{Asn}$ where $\varphi(\mathsf{c} \circ \alpha) = \psi(\mathsf{c} \circ \beta)$ and $(\psi, \beta) \in S$. We know that $(\varphi, \alpha) \sim (\psi, \beta)$ and that S is extensional, so $(\varphi, \alpha) \in S$.

$T(S(T)) \subseteq T$: suppose $\varphi \in T(S(T))$. Therefore there exist $\psi \in \mathcal{L}_{\mathsf{PA}}^+$, $\alpha \in \mathsf{Asn}(\psi)$, $\varphi = \psi(\mathsf{c} \circ \alpha)$, $(\psi, \alpha) \in S(T)$. Unpacking $S(T)$'s definition gives $\psi(\mathsf{c} \circ \alpha) \in T$. But $\varphi = \psi(\mathsf{c} \circ \alpha)$, so $\varphi \in T$. $T \subseteq T(S(T))$: suppose $\varphi \in T$. We

want $\psi \in \mathcal{L}^+_{\mathsf{PA}}, \alpha \in \mathsf{Asn}(\psi), \varphi = \psi(\mathsf{c} \circ \alpha), \psi(\mathsf{c} \circ \alpha) \in T$. We construct ψ and α by recursion:

- For $\varphi = (R, (c_{x_1}, \ldots, c_{x_n}))$, let $\psi = (R, (v_{x_1}, \ldots, v_{x_n}))$ and $\alpha(v_{x_i}) = x_i$.
- For $\varphi = (\neg, \varphi')$ with ψ' and α' such that $\varphi' = \psi'(\mathsf{c} \circ \alpha')$, let $\psi = (\neg, \psi')$ and $\alpha = \alpha'$.
- For $\varphi = (\vee, (\varphi_1, \varphi_2))$ with ψ_1, ψ_2 and α_1, α_2 such that $\varphi_1 = \psi_1(\mathsf{c} \circ \alpha_1)$ and $\varphi_2 = \psi_2(\mathsf{c} \circ \alpha_2)$. Let $\psi = (\vee, (\psi_1, \psi_2))$ and $\alpha = \alpha_1 \restriction \mathsf{FV}(\psi_1) \cup \alpha_2 \restriction \mathsf{FV}(\psi_2)$.
- For $\varphi = (\exists, (v, \varphi'))$ with φ' and α' such that $\varphi' = \psi'(\mathsf{c} \circ \alpha')$. Let $\psi = (\exists, (v, \psi'))$ and $\alpha = \alpha' \setminus \{v\}$.

Checking that the recursive definition is well-formed is similar to the proof of Lemma 130. □

Truth is Equivalent to Extensional Satisfaction in PA^-

To generalize the result to PA^-, we first have to show that the syntactical concepts required to express extensionality are expressible in PA^-. This involves the arithmetization of substitution, $\varphi * \sigma$, $\alpha \circ \sigma$, and $\varphi(\mathsf{c} \circ \alpha)$. By corollary 60, it suffices to show that these are definable by primitive recursion.

Definition 134 (Substitution for a Formula)**.** *We define* $\mathsf{Subst}(\sigma, \varphi)$ *as expressing* $\sigma : \mathsf{FV}(\varphi) \to \mathsf{Var}$*, and that if* x *is a free variable of* φ*, then* x *is not in the scope of any quantifier that binds* $\sigma(x)$*.*

$$\begin{aligned}\mathsf{Subst}(\sigma, \varphi) :\leftrightarrow & \mathsf{Seq}(\sigma) \wedge \mathsf{Form}(\varphi) \wedge \forall i \leq \ell(\sigma)\, [\exists v(v = \pi_1((\sigma)_i) \wedge v \in \mathsf{FV}(\varphi)) \wedge \exists v(v = \pi_2((\sigma)_i) \wedge \mathsf{Var}(v))] \\ & \wedge \forall v \in \mathsf{FV}(\varphi)\, [\exists y((v, y) \in \sigma \wedge \forall x, y\, ((v, x) \in \sigma \wedge (v, y) \in \sigma \to x = y))] \\ & \wedge \forall v \in \mathsf{FV}(\varphi)\, \mathsf{WellFormed}(v, \sigma(v), \varphi).\end{aligned}$$

where $\mathsf{WellFormed}(v, u, \varphi)$ *is defined by the following primitive recursive function:*

$$f(v, u, n) : \mathsf{Var} \times \mathsf{Var} \times \mathsf{Form} \mapsto \{1, 0\}$$

$$f(v, u, n) = \begin{cases} 1 & \text{if } n = (R, (t_0, ..., t_{n-1})), \\ f(v, u, \psi) & \text{if } n = (\neg, \psi), \\ \mathit{min}(f(v, u, \psi), f(v, u, \chi)) & \text{if } n = (\vee, (\psi, \chi)), \\ f(v, u, \psi) & \text{if } n = (\exists, (s, \psi)), s \neq u, \\ f(v, u, \psi) & \text{if } n = (\exists, (s, \psi)), s = u, v \notin \psi, \\ 0 & \text{if } n = (\exists, (s, \psi)), s = u, v \in \psi. \end{cases}$$

where $v \notin \psi$ *means that* v *is not a variable of* ψ*, which can be handled in an analogous way as* $\mathsf{FV}(\psi)$ *by a primitive recursive function. Let* $\mathsf{WellFormed}(v, u, n) := f(v, u, n)$*.*

Definition 135. *Suppose* $\varphi(v_0, \ldots, v_{n-1})$ *is a Feferman-style formula encoding with* $v_0 \ldots v_{n-1}$ *free, and* α *is an*

assignment for φ. There are primitive recursive functions s_α and $f_{F,\alpha}$ with the following property:

$$f_\alpha(t,F) = \begin{cases} \mathsf{c}_{\alpha(t)} & \textit{if } t \in F, \\ t & \textit{otherwise.} \end{cases}$$

$$s_\alpha(n,F) = \begin{cases} (R,(f_\alpha(t_0,F),...,f_\alpha(t_{n-1},F))) & \textit{if } n = (R,(t_0,...,t_{n-1})), \\ (\neg, s(\psi,F)) & \textit{if } n = (\neg,\psi), \\ (\vee,(s(\psi,F),s(\chi,F))) & \textit{if } n = (\vee,(\psi,\chi)), \\ (\exists,(v,s(\chi,F))) & \textit{if } n = (\exists,(v,\psi)). \end{cases}$$

Corollary 136. *The graphs of s_α and f_α are PA^- definable.*

We then define $\psi = \varphi(\mathsf{c} \circ \alpha) :\leftrightarrow \mathsf{Form}(\varphi) \wedge \mathsf{Form}(\psi) \wedge \mathsf{Asn}(\alpha,\varphi) \wedge (\varphi,\psi) \in s_\alpha(\varphi,\mathsf{FV}(\varphi))$. Since both its extension and anti-extension are $\Sigma_1(\mathsf{Seq})$, $\psi = \varphi(\mathsf{c} \circ \alpha)$ is expressible in PA^-.

$\psi = \varphi * \sigma$ can be treated similarly by functions f_σ and $s_\sigma(n,F)$:

$$f_\sigma(t,F) = \begin{cases} \sigma(t) & \text{if } t \in F, \\ t & \text{otherwise.} \end{cases}$$

$$s_\sigma(n,F) = \begin{cases} (R,(f_\sigma(t_0,F),...,f_\sigma(t_{n-1},F))) & \text{if } n = (R,(t_0,...,t_{n-1})), \\ (\neg, s(\psi,F)) & \text{if } n = (\neg,\psi), \\ (\vee,(s(\psi,F),s(\chi,F))) & \text{if } n = (\vee,(\psi,\chi)), \\ (\exists,(v,s(\chi,F))) & \text{if } n = (\exists,(v,\psi)). \end{cases}$$

Then define $\psi = \varphi * \sigma :\leftrightarrow \mathsf{Form}(\varphi) \wedge \mathsf{Form}(\psi) \wedge \mathsf{Subst}(\sigma,\varphi) \wedge (\varphi,\psi) \in s_\sigma(\varphi,\mathsf{FV}(\varphi))$

Remark 137. *Substitution preserves the structure of formulas:*

- *if $\varphi(x) \equiv (R,(t_0,\ldots,t_n))$, then $\varphi(\mathsf{c}\circ\alpha) \equiv (R,(t'_0,\ldots,t'_n))$ where for all $i \leq n$, either $t'_i = t_i$ or $t'_i = \mathsf{c}_{\alpha(t_i)}$. $\varphi * \alpha \equiv (R,(t'_0,\ldots,t'_n))$ where for all $i \leq n$, either $t'_i = t_i$ or $t'_i = \sigma(t_i)$,*
- *if $\varphi(x) \equiv (\neg,\psi)$, then $\varphi(\mathsf{c} \circ \alpha) \equiv (\neg, \psi(\mathsf{c} \circ \alpha))$ and $\varphi * \sigma \equiv (\neg, \psi * \sigma)$,*
- *if $\varphi(x) \equiv (\vee,(\psi,\chi))$, then $\varphi(\mathsf{c} \circ \alpha) \equiv (\vee,(\psi(\mathsf{c} \circ \alpha), \chi(\mathsf{c} \circ \alpha)))$ and $\varphi * \sigma \equiv (\vee,(\psi * \sigma, \chi * \sigma))$,*
- *if $\varphi(x) \equiv (\exists,(t,\psi))$, then $\varphi(\mathsf{c} \circ \alpha) \equiv (\exists,(t,\psi(\mathsf{c} \circ \alpha)))$, and $\varphi * \sigma \equiv (\exists,(t,\psi * \sigma))$.*

Finally, we define $(\varphi,\alpha) \sim (\psi,\beta)$ as

$$\begin{aligned}(\varphi,\alpha) \sim (\psi,\beta) := \exists\chi\exists\gamma\exists\sigma_1\exists\sigma_2(&\mathsf{Asn}(\gamma,\chi) \wedge \mathsf{Subst}(\sigma_1,\chi) \wedge \mathsf{Subst}(\sigma_2,\chi) \\ &\wedge \varphi = \chi * \sigma_1 \wedge \psi = \chi * \sigma_2 \wedge \gamma = \alpha \circ \sigma_1 \wedge \gamma = \beta \circ \sigma_2)\end{aligned}$$

which is $\Sigma_1(\mathsf{Seq})$, so PA^- decides its extension. For the anti-extension, we wait until a further lemma is established.

Remark 138. *We observe that, just as in the case of Remark 75-77, PA^- has enough machinery to operate on substitutions.*

For any substitution σ on formula φ (or on assignment α), we can define $\sigma[v : u]$ that resets the value of $\sigma(v)$ to u (as long as substitutivity is preserved), and define $\sigma \restriction \mathsf{FV}(\psi)$ (or $\sigma \restriction \mathsf{Dom}(\beta)$ where $\mathsf{FV}(\psi) \subseteq \mathsf{FV}(\varphi)$ (or $\mathsf{Dom}(\beta) \subseteq \mathsf{Dom}(\alpha)$).

Lemma 139 (PA^-). *There exists a cut $\Phi(x)$ such that for any model $\mathcal{M} \vDash \mathsf{PA}^-$, for all $\varphi, \psi \in \mathsf{Form}^{\mathcal{M}}$ and $\alpha, \beta \in \mathsf{Asn}^{\mathcal{M}}$ relativized to $\Phi(x)$, $\mathcal{M} \vDash \varphi(\mathsf{c} \circ \alpha) = \psi(\mathsf{c} \circ \beta)$ implies $\mathcal{M} \vDash (\varphi, \alpha) \sim (\psi, \beta)$.*

Proof. Note that the induction on φ in Lemma 130 is internal to the model $\mathcal{M}$. We therefore apply the shortening of Form-cuts (as defined in definition 63) to the formula $\Phi(x) := x(\mathsf{c} \circ \alpha) = y(\mathsf{c} \circ \beta) \to (x, \alpha) \sim (y, \beta)$. The (arithmetized) proof in Lemma 130 should serve as a proof for Φ being Form-inductive. Suppose $\Phi(\varphi)$ for all φ of complexity k. We show $\Phi(\varphi)$ for all φ with complexity $k+1$. For simplicity, we only consider the most difficult case $\varphi = (\exists, (x, \varphi_1))$.

Let $\mathcal{M}$ be a PA^--model. Suppose there is a $\psi \in \mathsf{Form}^{\mathcal{M}}$ and $\beta \in \mathsf{Asn}^{\mathcal{M}}$ where $\mathcal{M} \vDash \varphi(\mathsf{c} \circ \alpha) = \psi(\mathsf{c} \circ \beta)$, i.e, $\mathcal{M} \vDash (\exists, (x, \varphi_1(x)))(\mathsf{c} \circ \alpha) = (\exists, (x, \psi_1(x)))(\mathsf{c} \circ \beta)$. By Remark 137, it implies that $\varphi_1(x)(\mathsf{c} \circ \alpha) = \psi_1(x)(\mathsf{c} \circ \beta)$, where $c(\varphi_1(x)) = c(\psi_1(x)) = k$. Therefore $\mathcal{M} \vDash \varphi_1(\mathsf{c} \circ \alpha) = \psi_2(\mathsf{c} \circ \beta) \to (\varphi_1, \alpha) \sim (\psi_1, \beta)$. Unpacking the definition of $\sim$ gives that $\mathcal{M}$ sees witnesses $\chi_1 \in \mathsf{Form}^{\mathcal{M}}, \gamma' \in \mathsf{Asn}^{\mathcal{M}}(\chi_1)$ and $\sigma_\varphi, \sigma_\psi \in \mathsf{Subst}^{\mathcal{M}}(\chi_1)$ where

$$\mathcal{M} \vDash \varphi_1 = \chi_1 * \sigma_\varphi,\ \mathcal{M} \vDash \psi_1 = \chi_1 * \sigma_\psi; \tag{3.14}$$

$$\mathcal{M} \vDash \gamma' = \alpha' \circ \sigma_\varphi,\ \mathcal{M} \vDash \gamma' = \beta' \circ \sigma_\psi. \tag{3.15}$$

We wish to construct witnesses for $\mathcal{M} \vDash (\varphi, \alpha) \sim (\psi, \beta)$. First, to prevent undesirable binding, we modify $\sigma_\varphi, \sigma_\psi$ such that both map x to itself and nothing else maps to x, so $x \in \mathsf{FV}(\chi_1)$. This is possible by Remark 138. Then define $\chi = (\exists, (x, \chi_1)), \gamma = \gamma' \restriction \mathsf{FV}(\chi), \sigma'_\varphi = \sigma_\varphi \restriction \mathsf{FV}(\chi), \sigma'_\psi = \sigma_\psi \restriction \mathsf{FV}(\chi)$ (which removes x from the domain).

By Remark 137, $(\exists, (x, \chi_1(x))) * \sigma'_\varphi = (\exists, (x, \chi_1(x) * \sigma'_\varphi))$ and $(\exists, (x, \varphi_1(x) * \sigma'_\varphi)) = (\exists, (x, \varphi_1(x))) * \sigma'_\varphi$. Since $\mathcal{M} \vDash \varphi_1 = \chi_1 * \sigma_\varphi$, $\mathcal{M} \vDash \psi_1 = \chi_1 * \sigma_\psi$, and $\mathcal{M}$ believes that σ_φ agrees with σ'_φ except for x, $(\exists, (x, \chi_1 * \sigma'_\varphi)) = (\exists, (x, \varphi_1 * \sigma'_\varphi))$. Similarly for ψ.

Since $\mathcal{M} \vDash \alpha' = \gamma' \circ \sigma_\varphi$, $\mathcal{M} \vDash \alpha' \restriction \mathsf{FV}(\chi) = (\gamma' \circ \sigma_\varphi) \restriction \mathsf{FV}(\chi)$. Since $\sigma_\varphi(v) = x$ iff $v = x$, $(\gamma' \circ \sigma_\varphi) \restriction \mathsf{FV}(\chi) = (\gamma' \restriction \mathsf{FV}(\chi)) \circ \sigma'_\varphi = \gamma \circ \sigma'_\varphi = \alpha$. Similarly for β. Therefore $(\varphi, \alpha) \sim (\psi, \beta)$. □

Lemma 140 (PA^-). *There exists a cut $\Phi(x)$ such that for any model $\mathcal{M} \vDash \mathsf{PA}^-$, for all $\varphi, \psi \in \mathsf{Form}^{\mathcal{M}}$ and $\alpha, \beta \in \mathsf{Asn}^{\mathcal{M}}$ relativised to $\Phi(x)$, $\mathcal{M} \vDash (\varphi, \alpha) \sim (\psi, \beta)$ implies $\mathcal{M} \vDash \varphi(\mathsf{c} \circ \alpha) = \psi(\mathsf{c} \circ \beta)$.*

Proof. Consider the formula $\Phi(\varphi) = \varphi(\mathsf{c} \circ \alpha) = \psi(\mathsf{c} \circ \beta) \to (\varphi, \alpha) \sim (\psi, \beta)$. We show that $\Phi(x)$ is Form-inductive. Base case: $\varphi = (R, (t_1, \ldots, t_n))$. Suppose there exist ψ, α, β where $\mathcal{M} \vDash (\varphi, \alpha) \sim (\psi, \beta)$. Therefore there exists χ, γ and $\sigma_\varphi, \sigma_\psi : \mathsf{FV}(\chi) \to \mathsf{Var}$ where

$$\varphi = \chi * \sigma_\varphi,\ \gamma = \alpha \circ \sigma_\varphi;$$
$$\psi = \chi * \sigma_\psi,\ \beta = \alpha_2 \circ \sigma_\psi.$$

Since substitution preserves structure, we know that $\psi = (R, (t'_1, \ldots, t'_n))$ and $\chi = (R, (s_1, \ldots, s_n))$. Suppose, for a contradiction, that $\mathcal{M} \nvDash \varphi(\mathsf{c} \circ \alpha) = \psi(\mathsf{c} \circ \beta)$. Therefore there exists i such that $\alpha(t_i) \neq \beta(t'_i)$. Since $\varphi = \chi * \sigma_\varphi$, s_i is such that $\sigma_\varphi(s_i) = t_i$, similarly $\sigma_\psi(s_i) = t'_i$. Since $\gamma = \alpha \circ \sigma_\varphi$, $\gamma(s_i) = \alpha(\sigma_\varphi(s_i)) = \alpha(t_i)$, similarly $\gamma(s_i) = \beta(\sigma_\psi(s_i)) = \beta(t'_i)$. But $\alpha(t_i) \neq \beta(t'_i)$, a contradiction.

Suppose $\Phi(x)$ for all x with complexity k. We show $\Phi(x)$ for all x with complexity $k+1$. Again we only consider $\varphi = (\exists, (v, \varphi'))$. Suppose there exist ψ, α, β where $\mathcal{M} \vDash (\varphi, \alpha) \sim (\psi, \beta)$. Therefore ψ must be of the shape $(\exists, (v', \psi'))$ and χ must be of the shape $(\exists, (u, \chi'))$. By Remark 93, there exists $e \in M$ such that $\mathcal{M} \vDash (\varphi', \alpha[v : e]) \sim (\psi', \beta[v : e])$, therefore $\mathcal{M} \vDash \varphi'(\mathsf{c} \circ \alpha[v : e]) = \psi'(\mathsf{c} \circ \beta[v : e])$. Since substitution preserves structure, $\varphi(\mathsf{c} \circ \alpha) = (\exists, (v, \varphi'))(\mathsf{c} \circ \alpha) = (\exists, (v, \varphi'(\mathsf{c} \circ \alpha)))$, and similarly, $\psi(\mathsf{c} \circ \alpha) = (\exists, (v, \psi'(\mathsf{c} \circ \beta)))$. But it is easy to see that $\psi'(\mathsf{c} \circ \beta) = \varphi'(\mathsf{c} \circ \alpha)$ since otherwise $\mathcal{M} \vDash \varphi'(\mathsf{c} \circ \alpha[v : e]) = \psi'(\mathsf{c} \circ \beta[v : e])$ wouldn't be true.

Applying the shortening of Form-cuts to Φ leads to the cut in question. □

Therefore $(\varphi, \alpha) \sim (\psi, \beta)$ iff $\varphi(\mathsf{c} \circ \alpha) = \psi(\mathsf{c} \circ \beta)$. We know that $\varphi(\mathsf{c} \circ \alpha)$ is $\Delta_1(\mathsf{Seq})$ definable. So $\mathcal{M} \vDash (\varphi, \alpha) \not\sim (\psi, \beta)$ iff $\mathcal{M} \vDash \varphi(\mathsf{c} \circ \alpha) \neq \psi(\mathsf{c} \circ \beta)$ iff $\mathcal{M} \vDash \exists x \exists y (x = \varphi(\mathsf{c} \circ \alpha) \wedge y = \psi(\mathsf{c} \circ \beta) \wedge x \neq y)$. So PA^- decides the negative extension by $\Sigma_1(\mathsf{Seq})$ completeness.

Lemma 141 (PA^-)**.** *$T(S)$ is a truth class.*

Proof. We observe that the original proof only relies on the truth and satisfaction axioms, and lemma 130, which is known to hold in PA^- by lemma 139. □

Proposition 142 (PA^-)**.** *$S(T(S)) = S$, and $T(S(T)) = T$.*

Proof. The proof of Lemma 133 generalises. □