The Journal of Symbolic Logic
Volume 88, Number 2, June 2023

MODEL THEORY AND PROOF THEORY OF THE GLOBAL REFLECTION PRINCIPLE

MATEUSZ ZBIGNIEW ŁEŁYK

Abstract. The current paper studies the formal properties of the Global Reflection Principle, to wit the assertion "All theorems of Th are true," where Th is a theory in the language of arithmetic and the truth predicate satisfies the usual Tarskian inductive conditions for formulae in the language of arithmetic. We fix the gap in Kotlarski's proof from [15], showing that the Global Reflection Principle for Peano Arithmetic is provable in the theory of compositional truth with bounded induction only (CT_0). Furthermore, we extend the above result showing that Σ_1-uniform reflection over a theory of uniform Tarski biconditionals (UTB^-) is provable in CT_0, thus answering the question of Beklemishev and Pakhomov [2]. Finally, we introduce the notion of a prolongable satisfaction class and use it to study the structure of models of CT_0. In particular, we provide a new model-theoretical characterization of theories of finite iterations of uniform reflection and present a new proof characterizing the arithmetical consequences of CT_0.

§1. Introduction. The Global Reflection Principle (GRP) for a theory Th is the assertion that all theorems of Th are true. As the statement involves the notion of truth for the language of Th, to uncover its meaning adequately one shall pass to a proper extension of Th in a richer language. Minimal such extensions are called *axiomatic theories of truth* for Th. Each such theory arises by enriching the language of Th with a single fresh predicate $T(x)$ and adding a bunch of axioms characterizing $T(x)$ as a truth predicate for the language of Th. In the paper we focus on one of the most natural such extensions, which comprises straightforward formalizations of the usual inductive Tarski's conditions in the language of Th together with the predicate $T(x)$. Let us denote this theory with CT^-.[1]

The GRP lies at the intersection of at least three, to some extent independent, areas of research. The first, which was the starting point for the current paper, is the *Tarski Boundary project*, that seeks to characterize the extensions of $\mathrm{CT}^- + \mathrm{Th}$ which are conservative over Th. This is non-trivial for two reasons: on the one hand, if Th can develop enough coding apparatus,[2] $\mathrm{CT}^- + \mathrm{Th}$ does not prove any new sentences from the language of Th. On the other hand, it is an immediate

Received March 9, 2021.

2020 *Mathematics Subject Classification.* 03C62, 03F30, 03F35, 03H15.

Key words and phrases. satisfaction classes, reflection principles, global reflection principle, axiomatic theories of truth.

[1] The minus sign signalizes the lack of induction for the extended language. Defined as $\mathrm{CT}\restriction$ in [10] and CT in [2].

[2] As shown by [17], this holds whenever Th interprets the elementary arithmetic EA, which, for the purposes of this paper, can be taken to be $I\Delta_0 + \exp$.

0022-4812/23/8802-0012
DOI:10.1017/jsl.2022.39

consequence of second Gödels Incompleteness theorem, that $\mathrm{CT}^{-} + \mathrm{Th}{+}\mathrm{GRP}$ is a nonconservative extension of Th. Moreover, the "natural" extensions of $\mathrm{CT}^{-} + \mathrm{Th}$ that are nonconservative over Th all prove GRP for Th. Hence, in a sense, GRP is the source of nonconservativity in the realm of truth theories and it seems highly desirable to know what are the minimal resources needed to prove it.

The second area is proof theory, especially ordinal analysis as initiated in [21] and further developed, e.g., in [2]. In this approach natural truth-free counterparts of GRP, Uniform Reflection Principles (REF), play central roles in determining the quantitative information about the consequences of theories of predicative strength. To get things right, one adds stratified truth predicates to the picture and studies (partial) uniform reflection principles over axiomatic theories of truth (this is the method used in [2]). This is how the GRP enters the scene.

Last but not least, axiomatic theories of truth play an important role in the field of model theory of Peano Arithmetic (cf. [12, 14]). Any subset $S \subseteq \mathcal{M}$ such that $(\mathcal{M}, S) \models \mathrm{CT}^{-}$ is essentially a full satisfaction class. If $(\mathcal{M}, S)$ additionally satisfies GRP, then S contains all the theorems of Th, in the sense of $\mathcal{M}$. Satisfaction classes provide a very powerful tool in constructing interesting models of PA and investigating their structure.[3]

The current paper contributes to all the three areas. More specifically:

1. We prove that Δ_0-induction for the truth predicate is enough to prove GRP for PA. Coupled with the earlier results by Kotlarski [15], this shows that, over $\mathrm{CT}^{-} + \mathrm{EA}$, Δ_0 induction for the truth predicate is equivalent to GRP for PA (we denote this theory with CT_0). This improves on earlier results from [25] and provides a direct fix to Kotlarski's argument in [15].[4] Additionally, coupled with various developments from the literature, our result shows that the Global Reflection Principle for PA is a very robust notion, being equivalent to various others, apparently very different, truth-theoretic principles, as witnessed by the Many Faces Theorem (Corollary 3.18).
2. We extend the above result, answering the open problem posed by Beklemishev and Pakhomov in [2]. We show that not only GRP is provable in CT_0 but also Σ_1-Uniform Reflection over a weak truth extension of EA, called $\mathrm{UTB}^{-} + \mathrm{EA}$ (which adds to the arithmetical part uniquely *Uniform Tarski Biconditionals*). The result has some bearings on the analysis performed in [2].
3. We provide a new conservativity proof for CT_0. Unlike in the first one from [15] we are able to show directly that CT_0 is arithmetically conservative over ω-iterations of uniform reflection over PA (denote this theory with $\mathrm{REF}^{\omega}(\mathrm{PA})$[5]). The proof is based on an essentially model-theoretic idea of *prolonging* a (partial) satisfaction class in an end-extension. This proves to be sufficiently robust to characterize finite iterations of Uniform Reflection. We show that a model $(\mathcal{M}, S)$, where S is a partial inductive satisfaction class, satisfies n iterations of uniform reflection if and only if a nonstandard restriction of S can be prolonged n times.

[3] For example the construction of a recursively saturated rather classless model of PA by Schmerl [14] employs them in a crucial way.

[4] Around 2012 a serious gap in the proof of Theorem 2.2 was discovered by Richard Heck and Albert Visser.

[5] The direct conservativity argument for these theories is presented in [2] as well.

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

The paper is organised as follows: in Section 2 we introduce all the relevant preliminaries and context. In particular we develop handy and uniform conventions regarding the definable models and satisfaction classes. Section 3 is devoted to the proof of GRP in CT_0. In particular we describe the history of the problem and comment on flaws in Kotlarski's aforementioned proof [15]. The proof is a streamlined version of the one presented by the author in [18]. In Section 4 we extend the result from the previous section answering the question of Beklemishev and Pakhomov in [2] in the positive: we give a proof of Σ_1-uniform reflection over $\mathrm{UTB}^- + \mathrm{EA}$ in CT_0. The proof makes crucial use of the Arithmetized Completeness Theorem. Additionally, the section offers some strengthenings of this main result. In Section 5 we give a proof of the conservativity of CT_0 over ω-iterations of uniform reflection over PA. Extending the work of Kaye and Kotlarski [13] we characterize the theory of n-iterated uniform reflection over PA in terms of models of the form $(\mathcal{M}, S)$ where S is a partial inductive satisfaction class. Finally we examine the structure of models of CT_0 and prove a variation of the main result of Section 4.

To enhance the reading, $\square$ (as usual) denotes the end of a proof, while $\boxplus$ means that the proof is omitted. $\triangle$ signalizes the end of a definition, remark, convention, etc.

§2. Setting the stage.

In this section we gather all the technical preliminaries needed to follow our reasoning and at the same time develop a useful framework for proving our main results. In particular most of the results contained herein can be found (sometimes under slightly different wording) in [9, 12].

For starters, PA denotes Peano Arithmetic and $\mathcal{L}$ denotes its language, which we stipulate to contain $+, \times, 0, 1, \leq$ as primitive symbols. While studying extensions of PA in a richer language, we use a handy convention known from [14]: PA^* denotes any theory in the extended language that admits all instantiations of the induction scheme for the extended language. Similarly, $I\Sigma_n^*$ denotes the extension of PA with induction for Σ_n formulae of the extended language. If $\mathcal{L}'$ is any language then, $\mathcal{L}'_S$ and $\mathcal{L}'_T$ denote the result of extending $\mathcal{L}'$ with a single binary predicate S or a single unary predicate T, respectively. Most of the extensions of PA that we study are formulated either in the language $\mathcal{L}_S$ or $\mathcal{L}_T$. Last but not least, EA denotes elementary arithmetic, i.e., the extension of $I\Delta_0$ with a single Π_2 assertion "exp is total." All the theories we study extend EA, possibly in a richer language. $\Delta_0(\exp)$ denote the class of bounded formulae in the language with a symbol for the exponential function exp: it will be used throughout the paper because various syntactical functions needed to state the axioms for the satisfaction predicate are in fact $\Delta_0(\exp)$. However, since most of the theories we consider are extensions of PA, the presence of exp as a primitive symbol do not increase their strength. We explain this in more detail in Section 3.

To smoothly deal with class sized-objects (such as definable models of arithmetical theories) various definitions will be stated in the canonical predicative two-sorted extension of PA, i.e., ACA_0. Uppercase letters $X, Y, Z, X_1, Y_1, Z_1, \ldots$ denote second-order variables. In all applications we shall reason about definable classes (perhaps in a richer language) which will be substituted for the free second-order variables. The two sorted language is denoted $\mathcal{L}_2$.

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

2.1. Coding conventions. We assume a standard coding of syntax in PA (defined as in [9]): primitive symbols of the language are assigned numbers in a recursive way, and then terms, formulae, sentences, etc. are treated as well-formed sequences of such numbers. The notion of sequence is based on the definable Ackermanian membership predicate $\in$. $\mathrm{Term}(x)$, $\mathrm{ClTerm}(x)$, $\mathrm{Var}(x)$, $\mathrm{Form}(x)$, and $\mathrm{Sent}(x)$ denote the arithmetical formulae expressing that x is an arithmetical term, a closed term, a variable, an *arithmetical* formula, and an *arithmetical* sentence, respectively. $x \in \mathrm{Subf}(y)$ expresses that x is a subformula of y. We define $x \in \mathrm{Term}(y)$ and $x \in \mathrm{ClTerm}(y)$ analogously (y is required to be either a formula or a term). $x \in \mathrm{FV}(y)$ expresses that x is a variable which has a free occurrence in (a formula) y.

The choice of coding apparatus is irrelevant as long as the coding is PA provably monotone, i.e., the following is provable in PA:

$$\forall \phi, \psi \ \ (\phi \in \mathrm{Subf}(\psi) \rightarrow \phi \leq \psi).$$

We require a similar condition for (the given formalisation) of $x \in \mathrm{Term}(y)$. Various codings which violate this condition are studied in [8, 11].

Throughout the paper we distinguish between variables of the metalanguage, for which we reserve the symbols $x, y, z, x_0, x_1, \dots, y_0, y_1, \dots,$ and variables of the arithmetized language, which are denoted $v, v_0, v_1, \dots$. We assume a fixed correspondence between the first and the second ones. $\bar{x}$, $\bar{v}$, ... denote sequences of variables. For a formula ϕ, $\ulcorner \phi \urcorner$ denotes its Gödel code.

2.2. Some model theory of PA. All the definitions and conventions regarding models of PA are as in [12]. By default $\mathcal{M}, \mathcal{N}, \mathcal{K}$ (possibly with indices) range over *nonstandard* models of PA and M, N, K denote their respective universes. If $\mathcal{M}$ is any model (possibly for $\mathcal{L}_S$) and $\phi(\bar{x})$ a formula (possibly with parameters from $\mathcal{M}$; $\bar{x}$ denotes a sequence of variables), then $\phi^{\mathcal{M}}$ denotes the set definable by ϕ in $\mathcal{M}$, i.e., $\{\bar{a} \in M \mid \mathcal{M} \models \phi(\bar{a})\}$. If $\mathcal{M} \models \mathrm{PA}$ and $X \subseteq M^n$, then $X\restriction_{<b}$ denotes the restriction of X to all elements smaller than b. In the case when $\mathcal{N} \subseteq \mathcal{M}$, $X\restriction_{\mathcal{N}}$ denotes $\bigcup_{b \in N} X\restriction_{<b}$ (the restriction of a relation to the submodel).

Let $I \subseteq M$. We write $d > I$ if d is greater than all the elements of I. I is called an *initial segment* of $\mathcal{M}$ if I is closed downwards with respect to $\leq$. We say that I is a *cut* if I is an initial segment which is closed under successor, i.e.,

$$\forall x \ \ x \in I \rightarrow x + 1 \in I.$$

If I is a cut of $\mathcal{M}$, then we call $\mathcal{M}$ an *end-extension* of I and write $I \subseteq_e \mathcal{M}$ (note that I need not be a submodel of $\mathcal{M}$). Any element $c \in M$ such that $c > \omega$ is called nonstandard.

The following is one of the most basic consequences of induction in models of PA^*:

LEMMA 2.1 (Overspill). *If $\mathcal{M} \models \mathrm{PA}^*$, then no nontrivial cut of $\mathcal{M}$ is definable.* ⊞

In particular, if $\emptyset \subsetneq I \subsetneq_e M$ is a cut and $\phi(x)$ is any formula such that

$$\forall a \in I \ \ \mathcal{M} \models \phi(a),$$

then there exists a $d > I$ such that $\mathcal{M} \models \phi(d)$.

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

One last notion which is very tightly linked to the topic of satisfaction classes is *recursive saturation*:

DEFINITION 2.2. Fix $\mathcal{M}$ and $\bar{a} \in M$. Let $p(x)$ be a set of formulae with at most one variable x and parameters $\bar{a}$. We say that $p(x)$ is *recursive* (or *computable*) if so is the set

$$\{\ulcorner \phi(x, \bar{y}) \urcorner \mid \phi(x, \bar{a}) \in p(x)\}.$$

We say that $p(x)$ is a type over $\mathcal{M}$ if every finite subset of $p(x)$ is satisfied in $\mathcal{M}$. We say that $\mathcal{M}$ is realized if there is a $b \in M$ such that for every $\phi(x) \in p(x)$, $\mathcal{M} \models \phi(b)$. We say that $\mathcal{M}$ is *recursively saturated* (or *computably saturated*) if every recursive type over $\mathcal{M}$ is realized in $\mathcal{M}$. $\triangle$

2.3. Some model theory in PA. Models of theories extending Robinson's arithmetic are infinite objects; thus inside arithmetic they become essentially second-order objects. In what follows *a set* means a second-order object and we distinguish it from *a coded set* (a first-order object). The notion of a Δ_n set is explained below. $x \in X$ should be understood as a membership relation between a first- and a second-order object, whereas $x \in y$ denotes the Ackermanian membership (mentioned earlier) between first-order objects.

We recall the notion of a Δ_n-set (see [9]): $\mathrm{Sat}_{\Sigma_n}(x, y)(\mathrm{Sat}_{\Pi_n}(x, y))$ denotes the arithmetically definable partial satisfaction predicate for Σ_n (Π_n respectively) formulae (as in [12] or [9]) and $\mathrm{Tr}_{\Sigma_n}(x)(\mathrm{Tr}_{\Pi_n}(x))$ abbreviates $\mathrm{Sat}_{\Sigma_n}(x, \varepsilon)$ ($\mathrm{Sat}_{\Pi_n}(x, \varepsilon)$). We stress that the construction of $\mathrm{Sat}_{\Sigma_n}(\mathrm{Sat}_{\Pi_n})$ is elementary in n, so it gives rise to an EA-provably total Δ_1 map sending n to (the formula) Sat_{Σ_n}.

In PA, a Σ_n *set* (Π_n *set*) is any $\Sigma_n(\Pi_n)$ formula $\phi(v)$ with precisely one free variable. We define a Δ_n set to be a pair of formulae (ϕ, ψ) such that ϕ is Σ_n, ψ is Π_n, and

$$\forall x \ \big(\mathrm{Sat}_{\Sigma_n}(\phi, x) \equiv \mathrm{Sat}_{\Pi_n}(\psi, x)\big).$$

The notions of a $\Sigma_n(\Pi_n, \Delta_n)$ relation is defined analogously. If X is a Δ_k set given by the Σ_k formula $\phi(v)$ and a Π_k formula $\psi(v)$, then $x \in_k X$ abbreviates $\mathrm{Sat}_{\Sigma_k}(\phi, x)$. Note that $x \in_k X$ is Δ_k.

Observe that a set $A \subseteq M$ is *definable from parameters* in a model $\mathcal{M}$ if and only if, for some $k \in \omega$, there exists a Δ_k set X such that

$$A := \{x \in M \mid x \in_k X\}.$$

In the paper, except for side remarks, in which case the definitions below can clearly be adapted, we will only need to talk about models for very specific signatures consisting of two binary functions $+$, $\times$, one binary relational symbol $S(x, y)$, reserved for a satisfaction class and two constants, 0 and 1.

We note that since in PA models for theories extending some basic arithmetic (which we are uniquely interested in) are class-size objects, we do not always have a satisfaction relation for them. Models without the satisfaction relation will called *partial* to contrast them with the *full* ones for which the truth of an arbitrary sentence can be decided.

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

DEFINITION 2.3 (ACA_0; partial model). We say that $\mathbf{M} = (U_{\mathbf{M}}, +_{\mathbf{M}}, \times_{\mathbf{M}}, S_{\mathbf{M}}, 0_{\mathbf{M}}, 1_{\mathbf{M}})$ is a *partial model* if:

1. $U_{\mathbf{M}}, S_{\mathbf{M}}$ are sets and $S_{\mathbf{M}} \subseteq U^2_{\mathbf{M}}$,
2. $+_{\mathbf{M}}, \times_{\mathbf{M}}$ are functions of type $U^2_{\mathbf{M}} \to U_{\mathbf{M}}$,
3. $0_{\mathbf{M}} \in M$, $1_{\mathbf{M}} \in M$.

We say that $\mathbf{M}$ is a Δ_n-model if it is a Δ_n set satisfying the above conditions. △

DEFINITION 2.4. If $\mathcal{N}$ is any model of PA then we say that $\mathbf{M}$ is a partial $\mathcal{N}$*-definable model* if for some $k \in \omega$,

$$\mathcal{N} \models \text{“}\mathbf{M} \text{ is a partial } \Delta_k \text{ model.”}$$ △

Note that, according to our convention, “$\mathcal{N}$-definable” means “$\mathcal{N}$-definable with parameters.”

EXAMPLE 2.5 (ACA_0). For every set S, $\langle v = v, v_1 + v_2 = v_3, v_1 \cdot v_2 = v_3, S, 0, 1\rangle$ is a partial model. We denote it with $\mathbf{V}[S]$. $v = v, v_1 + v_2 = v_3$, and $v_1 \cdot v_2 = v_3$ denote sets definable with respective formulae. $\mathbf{V}$ denotes $\mathbf{V}[\emptyset]$. △

REMARK 2.6. If $\mathcal{N} \models \mathrm{PA}$ and $\mathbf{M} = (U_{\mathbf{M}}, +_{\mathbf{M}}, \times_{\mathbf{M}}, S_{\mathbf{M}}, 0_{\mathbf{M}}, 1_{\mathbf{M}})$ is a partial $\mathcal{N}$-definable model, then, outside of $\mathcal{N}$, it gives rise to a model for the signature $\{+, \times, S, 0, 1\}$. Indeed, we may define model $\mathcal{M}$ by putting

$$\mathcal{M} := ((U_{\mathbf{M}})^{\mathcal{N}}, (+_{\mathbf{M}})^{\mathcal{N}}, (\times_{\mathbf{M}})^{\mathcal{N}}, (S_{\mathbf{M}})^{\mathcal{N}}, 0_{\mathbf{M}}, 1_{\mathbf{M}}).$$

Such a model will be denoted by $(\mathbf{M})^{\mathcal{N}}$. △

DEFINITION 2.7.

1. For a natural number n, $\underline{n}$ denotes the canonical numeral naming n, i.e.,
$$\underbrace{1 + (1 + \cdots + (1 + 0)\ldots)}_{n \text{ times } 1}.$$
$y = \underline{x}$ denotes the formalisation of this relation in PA. We shall often treat $\underline{x}$ as a term symbol depending on variable x.
2. (ACA_0) An assignment is any function with domain $\mathrm{dom}(f) \subseteq \mathrm{Var}$. For a partial model $\mathbf{M}$, α is an $\mathbf{M}$*-assignment* if its range is contained in $U_{\mathbf{M}}$. We denote it with $\alpha \in \mathrm{Asn}(\mathbf{M})$. α is an $\mathbf{M}$*-assignment for* ϕ, symbolically $\alpha \in \mathrm{Asn}(\phi, \mathbf{M})$, if α is an $\mathbf{M}$-assignment and $\mathrm{FV}(\phi) = \mathrm{dom}(\alpha)$. We naturally extend this definition to coded sequences of terms and formulae: if s is such a sequence, then
$$\alpha \in \mathrm{Asn}(s, \mathbf{M}) := \forall i \in \mathrm{dom}(s) \;\; \alpha \in \mathrm{Asn}(s_i, \mathbf{M}).$$
$\alpha \in \mathrm{Asn}(s)$ has an analogous meaning.
3. (ACA_0) If α is any assignment and ϕ a formula, then by $\phi[\alpha]$ we denote the result of the simultaneous substitution of $\underline{\alpha(v)}$ for every free occurrence of v, for every $v \in \mathrm{FV}(\phi) \cap \mathrm{dom}(\alpha)$. $t[\alpha]$ for a term t is defined analogously. If we are interested in a single substitution in a formula ϕ, then we write $\phi[x/v]$, or $\phi[x]$ if v is clear from context, to mean $\phi[\alpha]$ where α is an assignment such that $\mathrm{dom}(\alpha) = \{v\}$ and $\alpha(v) = x$. Abusing the notation a little bit, for a term t, $\phi[t/v]$ denotes the result of substituting t for every free occurrence of v.

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

EXAMPLE 2.8 (PA). If $\phi = \big((v_0 + v_1 = v_2) \wedge (\exists v_2 \ v_2 = v_2)\big)$ and $\alpha(v_0) = 2$, $\alpha(v_2) = 3$, then

$$\phi[\alpha] = \big(((1 + 1 + 0) + x_1 = (1 + 1 + 1 + 0)) \wedge (\exists v_2 \ v_2 = v_2)\big).$$

Note that this is the same as $\big((\underline{2} + x_1 = \underline{3}) \wedge (\exists v_2 \ v_2 = v_2)\big)$. We shall use both formats. △

4. (ACA$_0$) If α is any X-assignment, then $\alpha{\restriction}_\phi$ abbreviates $\alpha{\restriction}_{\mathrm{FV}(\phi)}$, where $f{\restriction}_A$ denotes the restriction of a function f to a set A. If ϕ is clear from context, we will write $\alpha{\restriction}.$ instead of $\alpha{\restriction}_\phi$.
5. (ACA$_0$) If t is a term and $\alpha \in \mathrm{Asn}(t)$, then t^α denotes *the value of t under α*. It is the same as the value of $t[\alpha]$ ($t[\alpha]$ is a closed term).
6. (PA) If α and β are any two assignments and v is a variable, then $\alpha \leq_v \beta$ expresses that β extends α by assigning something to the variable v, i.e., $\mathrm{dom}(\beta) = \mathrm{dom}(\alpha) \cup \{v\}$ and for all $w \in \mathrm{dom}(\alpha)$, $\alpha(w) = \beta(w)$. Note that if $\alpha \leq_v \beta$ and $v \in \mathrm{dom}(\alpha)$, then $\alpha = \beta$.
7. (PA) If c is any (coded) set of variables and a is a number, then $[a]_c$ denotes the constant assignment sending everything in c to a. If a variable v is clear from context then we will omit it writing $[a]$ instead of $[a]_v$. △

DEFINITION 2.9 (ACA$_0$). Let $\mathbf{M}$ be a partial model. An $\mathbf{M}$-*evaluation of terms* is a partial function f of type

$$\mathrm{Term} \times \mathrm{Asn}(\mathbf{M}) \to \mathbf{M}$$

such that for every term t, $\{t\} \times \mathrm{Asn}(t, \mathbf{M}) \subseteq \mathrm{dom}(f)$ and for every terms s, t and every $\mathbf{M}$-assignment α:

1. $\langle t, \alpha\rangle \in \mathrm{dom}(f) \leftrightarrow \mathrm{FV}(t) \subseteq \mathrm{dom}(\alpha)$,
2. $\alpha \subseteq \beta \wedge \langle t, \alpha\rangle \in \mathrm{dom}(f) \to f(t, \alpha) = f(t, \beta)$,
3. $f(0, \alpha) = 0^{\mathbf{M}}$, $f(1, \alpha) = 1^{\mathbf{M}}$,
4. $f(v, \alpha) = \begin{cases} \alpha(v), \text{if } v \in \mathrm{dom}(\alpha), \\ \text{undefined, otherwise,} \end{cases}$
5. $f((s + t), \alpha) = f(s, \alpha) +^{\mathbf{M}} f(t, \alpha)$, $f((s \cdot t), \alpha) = f(s, \alpha) \cdot^{\mathbf{M}} f(t, \alpha)$. △

OBSERVATION 2.10 (ACA$_0$). *For every partial model* $\mathbf{M}$ *there exists the unique* $\mathbf{M}$-*evaluation of terms. We shall denote it with* $\mathrm{val}_{\mathbf{M}}$. *Moreover, if the model is* Δ_k, *then* $\mathrm{val}_{\mathbf{M}}$ *can be taken to be* Δ_k *as well.* ⊞

DEFINITION 2.11 (ACA$_0$). Let $\mathbf{M}$ be a partial model. If X is a set of formulae closed under subformulae, then let $s(X)$ denote the set of proper subformulae of formulae from X. S' is called an X-satisfaction relation for $\mathbf{M}$ if the conditions below holds.

1. $X \subseteq \mathrm{Form}_{\mathcal{L}_S} \wedge \forall\phi\forall\psi\big(\psi \in \mathrm{Subf}(\phi) \wedge \phi \in X \to \psi \in X\big)$.
2. $\forall y, z\big(S'(y, z) \to y \in X \wedge z \in \mathrm{Asn}(y, \mathbf{M})\big)$.
3. $\forall s, t\forall\alpha \in \mathrm{Asn}(s, t, \mathbf{M}) \quad \big(S'(s = t, \alpha) \equiv \mathrm{val}_{\mathbf{M}}(s, \alpha) = \mathrm{val}_{\mathbf{M}}(t, \alpha)\big)$.
4. $\forall s, t\forall\alpha \in \mathrm{Asn}(s, t, \mathbf{M}) \quad \big(S'(S(s, t), \alpha) \equiv \langle \mathrm{val}_{\mathcal{M}}(s, \alpha), \mathrm{val}_{\mathcal{M}}(t, \alpha)\rangle \in S_{\mathbf{M}}\big)$.
5. $\forall\phi \in s(X)\forall\alpha \in \mathrm{Asn}(\phi, \mathbf{M}) \quad \big(S'(\neg\phi, \alpha) \equiv \neg S'(\phi, \alpha)\big)$.
6. $\forall\phi, \psi \in s(X)\forall\alpha \in \mathrm{Asn}(\phi, \psi, \mathbf{M}) \quad \big(S'(\phi \vee \psi, \alpha) \equiv S'(\phi, \alpha{\restriction}.) \vee S'(\psi, \alpha{\restriction}.)\big)$.
7. $\forall\phi \in s(X)\forall v\forall\alpha \in \mathrm{Asn}(\exists v\phi, \mathbf{M})$ $\big(S'(\exists v\phi, \alpha) \equiv \exists\beta \geq_v \alpha\big(\beta \in \mathrm{Asn}(\mathbf{M}) \wedge$ $S'(\phi, \beta{\restriction}.)\big)$.

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

Let $\mathrm{CS}^{-}(X, \mathbf{M}, S')$ denote the conjunction of the above sentences of $\mathcal{L}_2$ (we treat $\mathbf{M}$, S', X as second-order variables). In the context of S, $S(\phi, \alpha\restriction.)$ always mean $S(\phi, \alpha\restriction_{\phi})$. △

Definition 2.12. (PA; measures of complexity of formulae)

1. The depth of a formula ϕ is the length of the longest path in the syntactic tree of ϕ. Equivalently, the depth of ϕ is defined recursively: the depth of an atomic formula is 0, $\exists$ and $\neg$ raise the complexity by one, and the depth of the disjunction is the maximum of the depths of the disjuncts plus one. $\phi \in \mathrm{dp}(x)$ expresses that the depth of ϕ is at most x.
2. Let us fix a canonical syntactical (elementary) transformation, which for a formula $\phi(\bar{x})$ returns a formula in the Σ_c form, that is logically equivalent to $\phi(\bar{x})$. Denote with $\phi(\bar{x})^{\Sigma}$ the result of applying this transformation. We assume that $\mathrm{FV}(\phi(\bar{x})) = \mathrm{FV}(\phi(\bar{x})^{\Sigma})$. For a number c, let Σ^*_c denote the class of formulae $\phi(x)$ such that $\phi(x)^{\Sigma} \in \Sigma_c$. △

Definition 2.13 (ACA_0). For a number c, a *c-full model* is a tuple $(\mathbf{M}, \mathrm{Sat}_{\mathbf{M}})$ where $\mathbf{M}$ is a partial model, and $\mathrm{Sat}_{\mathbf{M}} \subseteq \mathrm{Form}_{\mathcal{L}_S} \times \mathrm{Asn}(\mathbf{M})$ is a Σ^*_c-satisfaction relation for $\mathbf{M}$. A model $(\mathbf{M}, \mathrm{Sat}_{\mathbf{M}})$ is a *full model* if it is a c-full model for every c. Moreover, a tuple $(\mathbf{M}, \mathrm{Sat}_{\mathbf{M}})$ is a *depth-c*-full model if $\mathrm{Sat}_{\mathbf{M}}$ is a $\mathrm{dp}(c)$-satisfaction relation for $\mathbf{M}$, i.e., $\mathrm{CS}^{-}(\mathrm{dp}(c), \mathbf{M}, \mathrm{Sat}_{\mathbf{M}})$ holds. △

We stress that $(\mathbf{M}, \mathrm{Sat}_{\mathbf{M}})$ being c-full presupposes that $\mathbf{M}$ and $\mathrm{Sat}_{\mathbf{M}}$ satisfy full induction (treated as additional predicates).

Observe that if for every $n \in \omega$, $(\mathbf{M}, \mathrm{Sat}_{\mathbf{M}})$ is an $\mathcal{N}$-definable n-full model, then we have two satisfaction classes for $\mathbf{M}$ at our disposal: the metatheoretical one and $\mathrm{Sat}_{\mathbf{M}}$. The two relations agree in the following sense: for every $\phi(x_1, \dots, x_n) \in \mathcal{L}_S$ and for all $a_1, \dots, a_n \in (U_{\mathbf{M}})^{\mathcal{N}}$,

$$\mathcal{M} \models \phi[a_1/x_1, \dots, a_n/x_n] \iff \mathcal{N} \models \mathrm{Sat}_{\mathbf{M}}(\ulcorner \phi(x_0, \dots, x_n)\urcorner, [a_1, \dots, a_n]),$$

where $[a_1, \dots, a_n]$ denotes the assignment $x_i \mapsto a_i$, $i \leq n$.

Convention 2.14. *We reserve calligraphic letters $\mathcal{M}, \mathcal{N}, \mathcal{K}$ to talk, both internally and externally, about models with satisfaction relations, while $\mathbf{M}$, $\mathbf{N}$, $\mathbf{K}$ will denote arbitrary partial models.*

By the Tarski's undefinability of truth theorem one obtains that if $\mathcal{M}$ is any model of PA, then there is no formula $\mathrm{Sat}_{\mathbf{V}}$ with parameters from $\mathcal{M}$ such that for every n, $(\mathbf{V}, \mathrm{Sat}_{\mathbf{V}})$ is an n-full model. However, relativizing the standard partial truth predicates (see [9]) one obtains the following observation.

Observation 2.15 (ACA_0). *If $\mathbf{M}$ is any partial model, then for every k there are uniquely determined predicates $\mathrm{Sat}^k_{\mathbf{M}}$ such that $(\mathbf{M}, \mathrm{Sat}^k_{\mathbf{M}})$ is a k-full model.* ⊞

Proposition 2.16 (ACA_0). *Let X be closed under subformulae. Suppose that $\mathrm{Sat}_{\mathbf{M}}$ is an X-satisfaction class for $\mathbf{M}$. Let $Y \subseteq X$ be a set of sentences such that $\mathbf{M} \models_{\mathrm{Sat}_{\mathbf{M}}} Y$. Then Y is consistent.*

Proof. We reason in ACA_0 and assume the contrary. Then there is a sequent-calculus proof of the sequent

$$\Gamma \Rightarrow 0 = 1$$

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

in the pure first-order logic, where $\Gamma \subseteq Y$ is a finite set. By cut-elimination we may assume that this proof has a subformula property, so every formula occurring in it is a subformula of a formula from $\Gamma \cup \{0 = 1\}$. By induction on the length of the proof we can show that for every sequent $\Theta \Rightarrow \Delta$ it holds that

$$\forall \phi \in \Theta\big(\mathbf{M} \models_{\mathrm{Sat}_{\mathbf{M}}} \phi\big) \rightarrow \exists \psi \in \Delta\big(\mathbf{M} \models_{\mathrm{Sat}_{\mathbf{M}}} \psi\big).$$

This contradicts that the proof ends with $0 = 1$. □

The above notions of partial and full model lead to the definition of two interpretability relations between structures:

Definition 2.17 (Interpretable models; see [12]). Let $\mathcal{M}$ and $\mathcal{N}$ be two models of an extension of PA (not necessarily satisfying PA*). We say that $\mathcal{M}$ *interprets* $\mathcal{N}$ if there exists a partial $\mathcal{M}$-definable model $\mathbf{N}$ such that

$$\mathcal{N} = (\mathbf{N})^{\mathcal{M}}.$$

We say that $\mathcal{M}$ *strongly interprets* $\mathcal{N}$ iff there exists $\mathbf{K}$ witnessing that $\mathcal{M}$ interprets $\mathcal{N}$ and there exists an $\mathcal{M}$-definable satisfaction predicate $\mathrm{Sat}_{\mathbf{K}}$ making $\mathbf{K}$ a full model. Interpretability and strong interpretability will be denoted by $\lhd$ and $\lhd_S$, respectively. △

Observe that, as defined neither interpretability nor strong interpretability is preserved under isomorphism, in the sense that from $\mathcal{M} \lhd \mathcal{N}$ and $\mathcal{N} \simeq \mathcal{K}$ we cannot conclude that $\mathcal{M} \lhd \mathcal{K}$. The next two propositions uncover the important properties of $\lhd$. The following routine notion will come in handy:

Definition 2.18 (ACA$_0$; relativization). Suppose that $\mathbf{M}$ is a partial model. For every formula ϕ we define its *relativization* $\phi^{\mathbf{M}}$ by induction on the complexity of ϕ:

$$\begin{aligned}
(s(\bar{v_s}) = t(\bar{v_t}))^{\mathbf{M}} &:= \mathrm{val}_{\mathbf{M}}(s, [\bar{v_s}]) = \mathrm{val}_{\mathbf{M}}(t, [\bar{v_t}]),\\
(S(t(\bar{v})))^{\mathbf{M}} &:= \mathrm{val}_{\mathbf{M}}(t, [\bar{v}]) \in S_{\mathbf{M}},\\
(\phi \vee \psi)^{\mathbf{M}} &:= (\phi)^{\mathbf{M}} \vee (\psi)^{\mathbf{M}},\\
(\neg\phi)^{\mathbf{M}} &:= \neg(\phi)^{\mathbf{M}},\\
(\exists x \phi)^{\mathbf{M}} &:= \exists x \in U_{\mathbf{M}}\ \ (\phi)^{\mathbf{M}}.
\end{aligned}$$

Above, $\mathrm{val}_{\mathbf{M}}(t, [\bar{v_s}]) = y$ abbreviates the formula

$$\exists \alpha\big(\alpha \in \mathrm{Asn}(t, \mathbf{M}) \wedge \bigwedge_i \alpha(v_i) = x_i \wedge y = \mathrm{val}_{\mathbf{M}}(t, \alpha)\big). \qquad △$$

Proposition 2.19. *If $\mathcal{M} \lhd \mathcal{N}$ and $\mathcal{N} \lhd \mathcal{K}$, then $\mathcal{M} \lhd \mathcal{K}$.*

Proof. Suppose that $\mathcal{N} = (\mathbf{N})^{\mathcal{M}}$ and $\mathcal{K} = (\mathbf{K})^{\mathcal{N}}$. Suppose further that $\mathbf{N}$ is partial Δ_k model in $\mathcal{M}$. Hence using partial satisfaction predicate for Σ_k formulae, we can see that the $\mathbf{K}^{\mathbf{N}}$ (see Definition 2.18) makes sense in $\mathcal{M}$, and, in $\mathcal{M}$, $\mathbf{K}^{\mathbf{N}}$ is a partial Δ_k model. Moreover it is easy to observe that

$$\left(\mathbf{K}^{\mathbf{N}}\right)^{\mathcal{M}} = \mathcal{K},$$

which ends the proof. □

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

The following proposition will play a crucial role in some of our arguments. Its proof consists in internalizing the argument from Remark 2.6 and makes use of the arithmetization of the relativization function introduced above.

PROPOSITION 2.20 (Enayat–Visser). *Suppose that $\mathcal{M} \lhd_S \mathcal{N}$ and $\mathcal{N} \lhd \mathcal{K}$, then $\mathcal{M} \lhd_S \mathcal{K}$.*

PROOF. By Proposition 2.19 we have $\mathcal{M} \lhd \mathcal{K}$ and $(\mathbf{K})^{\mathbf{N}}$ is a partial $\mathcal{M}$-definable model witnessing the interpretability. We define the satisfaction relation for $(\mathbf{K})^{\mathbf{N}}$ via the formula

$$\mathrm{Sat}_{\mathbf{K}^{\mathbf{N}}}(x, y) := \mathrm{Form}_{\mathcal{L}_S}(x) \wedge y \in \mathrm{Asn}(x, \mathbf{K}^{\mathbf{N}}) \wedge \mathrm{Sat}_{\mathbf{N}}(x^{\mathbf{K}}, y). \qquad \square$$

In PA we can prove that every consistent theory admits a full model. Since in most cases both the theory and the model are infinite objects, this is in fact a parametrized family of theorems:

THEOREM 2.21 (Arithmetized Completeness Theorem). *For every $n \in \omega$,* PA* *proves the sentence*

$$\textit{Every } \Delta_n \textit{ consistent theory has a } \Delta_{n+1} \textit{ full model.} \qquad \boxplus$$

Since the proof of this fact (apart from axioms for arithmetical operations) depends only on the presence of induction, it can be proved also in every extension of PA which includes full induction scheme (for the extended language). This will be crucial in the second part of the paper. Let us complete this introductory part with two classical observations which give us some information about the structure of interpretable models. The first one shows that in fact interpretability can be seen as refined end-extendibility.

DEFINITION 2.22. If $\mathcal{M}$ is a model for a language $\mathcal{L}'$ extending $\mathcal{L}$, then $\mathcal{M}\restriction_{\mathcal{L}}$ denote its $\mathcal{L}$-reduct. △

PROPOSITION 2.23 (Folklore). *Let $\mathcal{M}, \mathcal{N}$ be models of* PA*. *If $\mathcal{M}$ interprets $\mathcal{N}$, then there exists a unique $\mathcal{M}$-definable isomorphism between $\mathcal{M}\restriction_{\mathcal{L}}$ and initial segment of $\mathcal{N}\restriction_{\mathcal{L}}$.*

PROOF. Suppose $\mathbf{N} := \langle U_{\mathbf{N}}, +_{\mathbf{N}}, \times_{\mathbf{N}}, S_{\mathbf{N}}, 0_{\mathbf{N}}, 1_{\mathbf{N}} \rangle$ is an $\mathcal{M}$ definable partial model witnessing the interpretability of $\mathcal{N}$ in $\mathcal{M}$. Let $\mathrm{val}_{\mathbf{N}}$ be a valuation function for $\mathbf{N}$. We define the embedding $\imath : \mathcal{M} \to \mathcal{N}$ via the formula

$$\imath(x) := \mathrm{val}_{\mathbf{N}}(\underline{x}, \varepsilon),$$

where $\underline{x}$ is a canonical numeral (in the sense of $\mathcal{M}$) naming x. By the earlier remarks there exists a satisfaction predicate $\mathrm{Sat}_{\mathbf{N}}$ making $\mathbf{N}$ a 1-full model. The fact that $\imath$ is an initial embedding follows since for every x we can build a quantifier-free sentence (in the sense of $\mathcal{M}$)

$$\forall v \ \left(v < \underline{x} \to \bigvee_{z<x} v = \underline{z} \right),$$

and by induction on x show that every such sentence is true in $\mathbf{N}$ according $\mathrm{Sat}_{\mathbf{N}}$. But this is equivalent to $\imath$ being an initial embedding. Now, if $\imath'$ is any other $\mathcal{M}$

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

definable isomorphism between $\mathcal{M}$ and an initial segment of $\mathcal{N}$, then it follows that

$$\mathcal{M} \models \text{“}\iota'(0) = 0_{\mathbf{N}} \wedge \forall x\big(\iota'(x+1) = \iota'(x) +_{\mathbf{N}} 1_{\mathbf{N}}\big).\text{”}$$

Then, by induction it follows that $\mathcal{M} \models \forall x\big(\iota(x) = \iota'(x)\big)$. □

If we strengthen the assumption to strong interpretability, then we can conclude that the interpreted model is always “longer.”

Proposition 2.24 (Folklore). *Let $\mathcal{M}, \mathcal{N}$ be models of* PA*. *Suppose that $\mathcal{M}$ strongly interprets $\mathcal{N}$ and let $\iota : \mathcal{M}\restriction_{\mathcal{L}} \to \mathcal{N}\restriction_{\mathcal{L}}$ be the embedding from Proposition 2.23. Then ι is not an elementary embedding. In particular, $\mathcal{M}$ is isomorphic to a proper initial segment of $\mathcal{N}$.*

Proof. Let $\mathcal{M}$, $\mathcal{N}$, ι be as above and let $\mathbf{N}$ be a partial $\mathcal{M}$-definable model such that $\mathcal{N} = (\mathbf{N})^{\mathcal{M}}$. That ι is not elementary follows from Tarski undefinability of truth theorem. Indeed, since ι and $\mathrm{Sat}_{\mathbf{N}}$ are $\mathcal{M}$ definable, we can define in $\mathcal{M}$ a predicate $S(x, y)$ by putting

$$S(\phi, \alpha) \equiv \mathrm{Sat}_{\mathbf{N}}(\phi, \iota \circ \alpha).$$

With such a definition for every formula $\phi(x_1, \dots, x_n)$ and all $a_1, \dots, a_n \in M$ we have

$$\mathcal{M} \models S(\ulcorner\phi(x_1, \dots, x_n)\urcorner, [a_1, \dots, a_n]) \iff \mathcal{N} \models \phi(\iota(a_1), \dots, \iota(a_n)).$$

Then, if ι were elementary, then the condition on the right-hand side would be equivalent to $\mathcal{M} \models \phi(a_1, \dots, a_n)$ which contradicts Tarski's theorem. The last part follows easily from the above and Proposition 2.23. □

Let us note one immediate corollary. Recall that $\mathcal{M}$ is κ-like if $|M| = \kappa$ but every proper initial segment of $\mathcal{M}$ has cardinality strictly smaller than κ.

Corollary 2.25. *If $\mathcal{M}$ is κ-like, then $\mathcal{M}$ is not strongly interpreted in any model of* PA.

Proof. Obviously, if $\mathcal{N}$ interprets $\mathcal{M}$, then $|N| \geq |M|$. Moreover, if $\mathcal{M}$ is strongly interpretable in $\mathcal{N}$, then it has a proper initial segment of cardinality $|M|$. □

Moreover, models strongly interpretable in a nonstandard model of PA have to be recursively saturated. This is a corollary to the proposition below (see also [14]):

Proposition 2.26. *Suppose d is a nonstandard element of $\mathcal{N}$. If $\mathcal{M}$ is isomorphic to a depth-d-full model, then $\mathcal{M}$ is recursively saturated.*

Proof. Suppose that $\mathcal{M} = (\mathbf{M}, \mathrm{Sat}_{\mathbf{M}})$ is an $\mathcal{N}$ definable depth-d-full model, $\mathcal{N}$ and d being as above. Fix an arbitrary recursive type $p(x) = \{\phi_i(x, a) \mid i \in \omega\}$ with (without loss of generality) a single parameter a and let $\sigma(x, y)$ be the Δ_1 formula representing its recursive enumeration, i.e., for every $i \in \omega$,

$$\mathrm{PA} \vdash \forall w \ \ \big(\sigma(i, w) \leftrightarrow w = \underline{\ulcorner\phi_i(x, z)\urcorner}\big).$$

Now, since the depth of every ϕ_i is less than d and $p(x)$ is a type we have for every $n \in \omega$

$$\mathcal{N} \models \exists z \forall i \leq n \forall w \ \ \big(\sigma(i, w) \rightarrow \mathrm{Sat}_{\mathbf{M}}(w, [x \mapsto z, y \mapsto a])\big)$$

($[x \mapsto z, y \mapsto a]$ denotes the unique assignment sending the variable x to z and the variable y to a). Hence, by overspill for some nonstandard c we have

$$\mathcal{N} \models \exists z \forall i \leq c \forall w \ \ \big(\sigma(i,w) \rightarrow \mathrm{Sat}_{\mathbf{M}}(w, [x \mapsto z, y \mapsto a])\big),$$

which shows that $p(x)$ is realised in $\mathcal{M}$. □

2.4. Satisfaction classes. Satisfaction classes provide truth conditions for $\mathbf{V}$.[6] Usually they are studied in the context of nonstandard models of PA. Let $\mathcal{M}$ be such a model.

Definition 2.27. We say that $S \subseteq M^2$ is a *partial satisfaction class* on $\mathcal{M}$ if there is a nonstandard c such that $(\mathcal{M}, S) \models \mathrm{CS}^-(\mathrm{dp}(c+1), \mathbf{V}, S)$. Equivalently the following holds in $(\mathcal{M}, S)$:

1. $\forall x, y \big(S(x,y) \rightarrow \mathrm{Form}(x) \wedge x \in \mathrm{dp}(c) \wedge y \in \mathrm{Asn}(x)\big)$.
2. $\forall s, t \forall \alpha \in \mathrm{Asn}(s,t) \ \ \big(S(s = t, \alpha) \equiv s^\alpha = t^\alpha\big)$.
3. $\forall \phi \in \mathrm{dp}(c) \forall \alpha \in \mathrm{Asn}(\phi) \ \ \big(S(\neg\phi, \alpha) \equiv \neg S(\phi, \alpha)\big)$.
4. $\forall \phi, \psi \in \mathrm{dp}(c) \forall \alpha \in \mathrm{Asn}(\phi, \psi) \ \ \big(S(\phi \vee \psi, \alpha) \equiv S(\phi, \alpha\restriction_\phi) \vee S(\psi, \alpha\restriction_\psi)\big)$.
5. $\forall \phi \in \mathrm{dp}(c) \forall v \forall \alpha \in \mathrm{Asn}(\exists v \phi) \ \ \big(S(\exists v \phi, \alpha) \equiv \exists \beta \geq_v \alpha \big(S(\phi, \beta\restriction_\phi)\big)\big)$.

Henceforth, the conjunction of 1–5 will be denoted by $\mathrm{CS}^-(c)$. If additionally $(\mathcal{M}, S) \models \mathrm{PA}^*$, then S is called a *partial inductive satisfaction class*. If $(\mathcal{M}, S) \models \forall x \ \ \mathrm{CS}^-(x)$, then S is called a *full satisfaction class*. Further define:

$$\begin{aligned} \mathrm{UTB}^- &:= \{\mathrm{CS}^-(\underline{n}) \mid n \in \omega\}, \\ \mathrm{UTB}_n &:= \mathrm{UTB}^- + I\Sigma_n(S), \\ \mathrm{UTB} &:= \bigcup_{n \in \omega} \mathrm{UTB}_n. \end{aligned}$$

Now we define an analogue of arithmetical hierarchy over the theory of a full satisfaction class.

$$\begin{aligned} \mathrm{CS}^- &:= \forall x \ \ \mathrm{CS}^-(x), \\ \mathrm{CS}_n &:= \mathrm{CS}^- + I\Sigma_n(S), \\ \mathrm{CS} &:= \bigcup_{n \in \omega} \mathrm{CS}_n. \end{aligned}$$

If S is a partial satisfaction class on $\mathcal{M}$ and $b \in M$, then we put

$$S_b := \{\langle \phi, \alpha \rangle \in S \mid \phi \in \Sigma_b^*\}.$$

Note that S_b is Δ_0 definable from S, and hence for every n, if S is Σ_n inductive, then so is S_b. △

Note that if S is a full satisfaction class, then $(\mathcal{M}, S)$ strongly interprets $\mathcal{M}$ and $\mathbf{V}$ is a Δ_0 partial model witnessing the interpretation. However, $(\mathbf{V}, S)$ need not be full, as we need not have any induction for S. In particular, it does not follow that

$$(\mathcal{M}, S) \models \forall \phi \in \mathrm{Ax}_{\mathrm{PA}} \ \ S(\phi, \varepsilon),$$

[6] Obviously one can study satisfaction classes for languages with additional predicates, but we will not be interested in such objects.

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

which, arguably, would mean that $\mathcal{M}$ knows that $\mathbf{V}$ is a model of PA. Let us call such a satisfaction class PA-*correct*. It can be shown that for a countable $\mathcal{M}$ the following conditions are equivalent:

1. There exists a full satisfaction class on $\mathcal{M}$.
2. There exists a PA-correct full satisfaction class on $\mathcal{M}$.
3. $\mathcal{M}$ is recursively saturated.

The implication 3. ⇒ 2. has been shown for the first time in [16]. References [7, 17] contain different proofs. The implication 1. ⇒ 3. is a consequence of Lachlan's theorem (see Theorem 2.31).

The name UTB^{-} stands for *Uniform Tarski Biconditionals*[7] and is normally used for the theory having as axioms all sentences of the form

$$\forall\alpha \in \mathrm{Asn}(\phi)\ \ S(\underline{\ulcorner\phi\urcorner}, \alpha) \equiv \phi((\alpha(x_1), \dots, (\alpha(x_n)),$$

for every $\phi(x_1, \dots, x_n) \in \mathrm{Form}_{\mathcal{L}}$.[8] One can show that, over EA, this set of sentences is equivalent to the one we've officially taken as a definition of UTB^{-}. By Observation 2.15 each finite portion of UTB^{-} is definable in PA and consequently we obtain the following proposition (which formalizes in EA).

Proposition 2.28 (EA). *If* Th *is any extension of* PA, *then* $\mathrm{UTB} + \mathrm{Th}$ *is conservative over* Th. ⊞

Furthermore, observe that $(\mathcal{M}, S) \models \mathrm{CS}^{-}$ iff S is a full satisfaction class on $\mathcal{M}$ and $(\mathcal{M}, S) \models \mathrm{CS}$ iff S is a full inductive satisfaction class in the sense of [14]. For further usage let us observe that the relation of CS to CS_n is similar to that between PA and $I\Sigma_n$. In particular there are definable partial Sat_{Σ_n} satisfaction predicates for $\Sigma_n\mathcal{L}_S$ formulae. Each Sat_{Σ_n} is a $\Sigma_n\mathcal{L}_S$ formula. As a consequence we obtain:

Proposition 2.29. *For every n,* $\mathrm{EA} + \mathrm{CS}_{n+1} \vdash \mathrm{Con}_{\mathrm{EA}+\mathrm{CS}_n}$. ⊞

We note that if S is a partial satisfaction class on a model $\mathcal{M}$, then for an arbitrary standard formula $\phi(x_0, \dots, x_n) \in \mathcal{L}$,

$$(\mathcal{M}, S) \models \forall\alpha\ \ \left(S(\ulcorner\phi\urcorner, \alpha) \equiv \phi(\alpha(x_1), \dots, \alpha(x_n))\right).$$

In particular it follows from Tarski's theorem that S is never definable in $\mathcal{M}$ (even if we allow parameters).

Nonstandard satisfaction classes provide a very useful tool for investigating nonstandard models of PA. The first point of interest is that their existence implies recursive saturation. For starters we cite a proposition which directly follows from Proposition 2.26:

Proposition 2.30 (Folklore; see [12]). *If S is a partial inductive satisfaction class in $\mathcal{M}$, then $\mathcal{M}$ is recursively saturated.*

Proof. Suppose that $(\mathcal{M}, S) \models \mathrm{CS}(c)$ for some nonstandard $c \in M$. Then $\mathcal{M}$ is isomorphic to a depth-c-full $\mathcal{M}$-definable model. Hence by Proposition 2.26 it is recursively saturated. □

[7] This theory is defined as UTB↾ in [10] and as UTB in [2].

[8] In the above α is a bound variable, so $\phi((\alpha(x_1), \dots, (\alpha(x_n))$ denotes the formula $\exists y_1, \dots, y_n\ \left(\bigwedge_{i\leq n} y_i = (\alpha)_i \wedge \phi(y_1, \dots, y_n)\right)$.

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

Interestingly, with a much more complicated proof one can strengthen the above proposition lifting the assumption that the chosen satisfaction class is inductive.

THEOREM 2.31 (Lachlan; see [12]). *If for some nonstandard c, $(\mathcal{M}, S) \models \mathrm{CS}^-(c)$, then $\mathcal{M}$ is recursively saturated.* ⊞

The converse to Lachlan's theorem fails, as was shown by Smith.

THEOREM 2.32 (Smith [22]). *If $(\mathcal{M}, S) \models \mathrm{CS}^-$, then there is S' such that for some nonstandard $c \in M$ $(\mathcal{M}, S') \models \mathrm{CS}_0(c)$.* ⊞

The condition that $(\mathcal{M}, S') \models I\Delta_0(S)$ implies that S' is piecewise coded (in the sense of [9]) or, using set-theoretical notions, a class on $\mathcal{M}$. Since $(\mathcal{M}, S') \models \mathrm{CS}^-(c)$ it follows that S' is not definable (even allowing parameters) in $\mathcal{M}$ (or, is a proper class). Since there are recursively saturated models of PA in which every class is definable (with parameters; see [14]), Smith's result shows that there are recursively saturated models which do not carry a full satisfaction class.

A common strengthening of theorems of Smith's and Lachlan's was obtained by Wcisło in [25]:

THEOREM 2.33 (Wcisło). *If $(\mathcal{M}, S) \models \mathrm{CS}^-(c)$ then there is an S' and a nonstandard c such that $(\mathcal{M}, S') \models \mathrm{CS}(c)$.* ⊞

An interesting open problem in the model theory of PA is whether the converse to the above theorem is true, i.e., whether every $\mathcal{M} \models \mathrm{PA}$ which admits a partial inductive satisfaction class admits a full satisfaction class. If one allows to prolong the given model, then there is a positive answer to this question.

THEOREM 2.34 (Visser). *If $(\mathcal{M}, S) \models \mathrm{CS}(c)$ for some $c \in M$, then there are $\mathcal{M} \preceq_e \mathcal{N}$ and S' such that $(\mathcal{N}, S') \models \mathrm{CS}^-$ and $S \subseteq S'$.* ⊞

The proof of this theorem is given in [20, Theorem 43]. In Section 5 we shall give an analogous result for $\mathcal{M} \models \mathrm{REF}^{\omega}(\mathrm{PA})$ and CS_0 instead of CS^-.

REMARK 2.35. The theory CS^- is a cousin of a compositional truth theory CT^- which admits a unary predicate T. All compositional axioms of CS^- can be easily adapted to this new setting; however in the case of the universal quantifier we have two natural ways to go. The first candidate is the "numeral" version, i.e.,

$$\forall v \forall \phi(v) \;\; \big(T(\forall v \phi) \equiv \forall x T(\phi[x/v])\big).$$

We stress that $\phi[x/v]$ denotes the result of substituting the numeral naming x for every free occurrence of the variable v. If such an axiom is adopted, then the resulting theory, denote it $n\mathrm{CT}^-$, can define the satisfaction predicate satisfying CS^- via the formula

$$S(\phi, \alpha) := \mathrm{Form}_{\mathcal{L}}(\phi) \wedge \alpha \in \mathrm{Asn}(\phi) \wedge T(\phi[\alpha]).$$

Let us stress that the above formula is $\Delta_0(\exp)$. The second option is the "term" version of CT^-, denote it $t\mathrm{CT}^-$, where the axiom for $\forall$ is the following:

$$\forall v \forall \phi(v) \;\; \big(T(\forall v \phi) \equiv \forall t \in \mathrm{Term} \;\; T(\phi[t/v])\big).$$

Using Enayat–Visser methods from [7] it can be shown that $n\mathrm{CT}^-$ and $t\mathrm{CT}^-$ are independent of each other, i.e., neither of them implies the other one (over the

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

remaining axioms of CS^-). Moreover, it is an open problem whether $t\mathrm{CT}^-$ can define the predicate of CS^-. In this paper CT^- will be introduced in Section 3 and will denote the numeral version, i.e., $n\mathrm{CT}^-$. △

2.5. Reflection principles. Reflection principles are various (families of) statements expressing the soundness of a given theory Th in a way which is transparent for Th. In other words, their aim is to capture the meaning of the metatheoretical assertion:

Every theorem of Th *is true.*

In order to avoid the problem of choosing the presentation for (an abstractly given theory Th), we will assume that Th is an elementary formula, which, provably in EA, defines a set of sentences. Such a formula will be called a Gödelized theory. We use PA to abbreviate the canonical elementary formula saying "x is an axiom of PA^- or an axiom of induction." Having a satisfaction predicate $S(x,y)$ at our disposal we can express the above in the form of the Global Reflection Principle

$$\forall \phi \ \ \mathrm{Pr}_{\mathrm{Th}}(\phi) \rightarrow S(\phi, \varepsilon). \tag{GR(Th)}$$

If S satisfies UTB^-, then from this one can derive instantiation of the uniform reflection

$$\forall x_1 \dots \forall x_n \big(\mathrm{Pr}_{\mathrm{Th}}(\ulcorner \phi \urcorner[\underline{x_1}, \dots, \underline{x_n}]) \rightarrow \phi(x_1, \dots, x_n)\big). \tag{REF(Th)}$$

Hence REF(Th) contains all the formulae of the above form for the language of Th. If Γ is a set of formulae of the language of Th, then Γ-REF(Th) denote the restriction of REF(Th) to formulae from class Γ. Below we will need also its iterated versions:

$$\begin{aligned}
\Gamma\text{-REF}^0(\mathrm{Th}) &:= \mathrm{Th},\\
\Gamma\text{-REF}^{n+1}(\mathrm{Th}) &:= \mathrm{Th} + \Gamma\text{-REF}(\mathrm{REF}^n(\mathrm{Th})),\\
\Gamma\text{-REF}^{\omega}(\mathrm{Th}) &:= \bigcup_{n \in \omega} \Gamma\text{-REF}^n(\mathrm{Th}).
\end{aligned}$$

In the successor step we tacitly fix the canonical representation of Γ-$\mathrm{REF}^n(\mathrm{Th})$.

The last definition which is relevant to formalizing soundness claims introduces the oracle provability predicates.

Definition 2.36. Let Th be any elementary theory. $\mathrm{Proof}^X_{\mathrm{Th}}(x,y)$ denotes a $\Delta^0_0(\exp)$ formula with a second-order variable X which canonically formalizes the relation: "y is a proof of sentence x from axioms of Th and sentences belonging to X." $\mathrm{Pr}^X_{\mathrm{Th}}$ is the Σ^0_1 provability predicate based on it. △

The oracle provability predicate defined above enables us to (uniformly) define a closure conditions on various satisfaction classes. For example we shall often encounter the assertion

$$\forall \phi \big(\mathrm{Pr}^S_{\emptyset}(\phi) \rightarrow S(\phi, \varepsilon)\big),$$

which should be read as "Every first-order consequence of true sentences is true," where "true" abbreviates that $S(\phi, \varepsilon)$ holds. In the above assertion we simply

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

substitute the definable class $\{x \mid S(x,\varepsilon)\}$ for the free second-order variable X. Let us also observe that formally $\mathrm{Pr}^X_{\mathrm{Th}}$ is the same as $\mathrm{Pr}_{\mathrm{Th}\cup X}$; however, for heuristic reasons we prefer to keep the lower index for absolute definitions and the upper one for arbitrary sets of formulae.

2.5.1. Reflection and internal models. The theory REF(Th) admits a model-theoretical characterisation in terms of strongly definable models. Below $\mathsf{K}_A(\mathcal{M})$ denote the set of elements of a model M which are definable in $\mathcal{M}$ with parameters from the set A.

Theorem 2.37 (Kotlarski–Kaye [13]). *For an arbitrary recursively saturated $\mathcal{M}$ the following are equivalent*:

1. $\mathcal{M} \models \mathrm{REF}(\mathrm{Th})$.
2. *There exists a full $\mathcal{M}$-definable model $\mathcal{N} = (\mathbf{N}, \mathrm{Sat}_{\mathbf{N}})$ such that*:
   (a) $\mathcal{M} \equiv \mathcal{N}$.
   (b) $\mathsf{K}_\emptyset(\mathcal{M}) = \mathsf{K}_\emptyset(\mathcal{N})$.
   (c) $\mathcal{M} \models \forall \phi \in \mathrm{Th}\ \ \mathrm{Sat}_{\mathbf{N}}(\phi,\varepsilon)$.

Thanks to condition (a) imposed on $\mathcal{N}$ in the above, Theorem 2.37 can be iterated an arbitrary finite number of times. Let us call the pair $(\mathcal{M},\mathcal{N})$ a *KK-pair* if it satisfies conditions 1. – 3. in the thesis of the above theorem. Thus we obtain:

Corollary 2.38. $\mathcal{M} \models \mathrm{REF}(\mathrm{Th})$ *if and only if there exists $\{\mathcal{M}_i\}_{i\in\omega}$ such that $\mathcal{M}_0 = \mathcal{M}$ and for each i, $(\mathcal{M}_i, \mathcal{M}_{i+1})$ is a KK-pair.*

In Section 5 we shall offer a similar in spirit model-theoretical characterization of $\mathrm{REF}^n(\mathrm{Th})$ and $\mathrm{REF}^\omega(\mathrm{Th})$. The main difference will be that we shall work with models with satisfaction classes.

2.5.2. Reflection and satisfaction classes. Full satisfaction classes in nonstandard models embody the conception of a satisfaction relation for $\mathbf{V}$. However, as we have already remarked, not every satisfaction class provides us with a reasonable truth predicate for $\mathbf{V}$. One property that one would require from such a truth predicate is the closure under the internal provability relation. In particular the satisfaction relation for $\mathbf{V}$ should make true all the (internal) theorems of first-order logic. This corresponds to the sentence

$$\forall \phi\ \left(\mathrm{Pr}_\emptyset(\phi) \to S(\phi,\varepsilon)\right) \qquad (\mathrm{GR}(\varnothing))$$

being true in a model $(\mathcal{M}, S)$. However, as shown for the first time in [4], over CS⁻ the above sentence implies GR(PA). In particular, in a countable recursively saturated model $\mathcal{M}$ in which there is a proof of inconsistency of PA there is no such a reasonable class for $\mathbf{V}$ (although there are many unreasonable ones). A characterization of models admitting a well-behaved satisfaction class was essentially first given by Kotlarski (in [15]):[9]

[9] The attribution here is qualified by "essentially," since Kotlarski proves this theorem for a different axiomatization of $\mathrm{REF}^\omega(\mathrm{PA})$ (see [18]). The current formulation requires going through the characterization of formal ω-consistency by Smoryński [23]. This paper contains a different direct proof of this theorem.

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

Theorem 2.39 (Kotlarski). *Suppose $\mathcal{M}$ is a countable recursively saturated model of* PA. *Then, there exists a full satisfaction class S such that* $(\mathcal{M}, S) \models \mathrm{GR(PA)}$ *if and only if* $\mathcal{M} \models \mathrm{REF}^{\omega}(\mathrm{PA})$. ⊞

A different natural question is how much induction is required to prove the global reflection for PA. Here a partial answer was given by Wcisło in [25], who showed that the satisfaction predicate satisfying $\mathrm{CS}^{-} + \mathrm{GR(PA)}$ is definable in CS_0:

Theorem 2.40 (Wcisło). *There exists an $\mathcal{L}_S$ formula $S'(x, y)$ such that*

$$\mathrm{EA} + \mathrm{CS}_0 \vdash [\mathrm{CS}^{-} + \mathrm{GR(PA)}][S'/S]. \qquad ⊞$$

In the above $\phi[S'/S]$ denotes the result of a uniform substitution of a formula $S'(s, t)$ for each occurrence of a formula $S(s, t)$ (renaming variables if necessary). In the next section we improve this result and show that GR(PA) is provable in $\mathrm{EA} + \mathrm{CS}_0$.

§3. Provability of the global reflection principle.

In this section we confirm Kotlarski's claim [15] that, over EA, the Δ_0 induction for a satisfaction predicate is enough to prove the Global Reflection Principle for PA. We start by explaining the original strategy and our fix. Unless said otherwise, all theories by default extend EA. However, it is very easy to see that $\mathrm{CS}_0 + \mathrm{EA} \vdash \mathrm{PA}$, since for every arithmetical formula $\phi(x)$, $\psi(x) := T(\phi[x])$ is a $\Delta_0(\mathcal{L}_T + \exp)$ and consequently, we have an induction axiom for it.

3.1. Kotlarski's proof.

Kotlarski's proof of GR(PA) in CS_0 starts by observing that each Δ_0 inductive satisfaction class makes all (in the sense of the ground model) the axioms of induction true. The argument runs as follows: working in CS_0 fix a formula $\phi(v)$ with a free variable v. Then, $S(\phi(v), [x])$ is a $\Delta_0(\exp)$ formula with a free variable x, where $[x]$ denotes the assignment $\{\langle v, x\rangle\}$. Hence, the following is an axiom of CS_0:

$$S(\phi(v), [0]) \wedge \forall x \big(S(\phi(v), [x]) \to S(\phi(v), [x+1])\big) \longrightarrow \forall x S(\phi(v), [x]).$$

Here Kotlarski's proof of PA-correctness ends. However, it is not obvious whether the above is equivalent to $S(\mathrm{Ind}(\phi(v)), \varepsilon)$, where $\mathrm{Ind}(\psi)$ denotes the axiom of induction for a formula ψ. Repeated applications of the compositional clauses yield the equivalence of $S(\mathrm{Ind}(\phi(v)), \varepsilon)$ with

$$S(\phi[0/v], \varepsilon) \wedge \forall x \big(S(\phi(v), [x]) \to S(\phi[v+1/v], [x])\big) \longrightarrow \forall x S(\phi(v), [x]).$$

Firstly, on the grounds of CS^{-} alone $S(\phi(v), [0])$ neither implies $S(\phi[0/v], \varepsilon)$ nor is implied by it. Similarly with $S(\phi(v), [x+1])$ and $S(\phi[v+1/v], [x])$. To see this one should think of a nonstandard $\phi(v)$ in which v occurs nonstandardly deep in $\phi(v)$ (i.e., at a nonstandard level of $\phi(v)'$s syntactical tree). For example one can take $\phi(v)$ to be

$$\underbrace{0 = 0 \vee (0 = 0 \vee (0 = 0 \vee \cdots}_{a \text{ times } 0=0} \vee v = v) \cdots),$$

for a nonstandard element a. This is the first problem.

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

The second problem lies in showing that a Δ_0–inductive satisfaction class is closed under provability, i.e., proving the sentence

$$\forall\phi\ \left(\mathrm{Pr}^S_\emptyset(\phi)\rightarrow S(\phi,\varepsilon)\right).$$

In the above $\mathrm{Pr}^S_\emptyset$ derives from the oracle provability predicate defined in Definition 2.36. Kotlarski's idea was to work (internally in PA) with a Hilbert-style proof calculus with Modus Ponens as the only rule of reasoning and then using $\Delta_0(\exp)$ induction for an $\mathcal{L}_S$ formula

$$\theta(x,p) := \text{“If } \phi \text{ is the } x\text{– th sentence in } p,\ \text{then } S(\phi,\varepsilon)\text{.”}$$

In $\theta(x,p)$ all quantifiers can be bounded by p,[10] that can be taken as a parameter. So it is indeed a $\Delta_0(\exp)$-formula. In the base step ϕ is either a logical axiom or a true sentence, so it's truth is either trivial (the latter case) or seems to follow from the compositional axioms. In the inductive step, we have to check that if ϕ and $\phi\rightarrow\psi$ are true, then so is ψ. This is indeed guaranteed by the compositional axioms.

However, problems arise while verifying the base step. For example (working in a nonstandard model) we might encounter the following logical axiom:

$$\psi := \forall v\phi(v)\rightarrow\phi(t),$$

for some nonstandard formula ϕ and a term t. Then $S(\psi,\varepsilon)$ is equivalent to

$$\forall y S(\phi(v),[y])\rightarrow S(\phi(t),\varepsilon),$$

so we encounter problems similar to the ones discussed while dealing with the truth of induction axioms. Moreover, there are more generic problems: if one does not want to incorporate the rule of universal generalisation, then one has to accept universal generalisations of all instances of propositional tautologies as axioms. In particular (working in a nonstandard model $(\mathcal{M},S)$) in the base step one might encounter an axiom of the form

$$\xi := \forall v_1\ldots\forall v_a\left(\phi\vee\neg\phi\right).$$

If it is true that that for any full satisfaction class S on $\mathcal{M}$, $(\mathcal{M},S)\models\forall\alpha\in\mathrm{Asn}(\phi)S(\phi\vee\neg\phi,\alpha)$, inferring that $(\mathcal{M},S)\models S(\xi,\varepsilon)$ requires some argument which cannot be carried out in CS^- alone.

3.2. The idea. To fill in the gaps in Kotlarski's reasoning it is sufficient to establish within CS_0 a kind of induction on the buildup of formulae. Indeed, what we missed were (inter alia) the following properties:

$$\forall\phi\in\mathrm{Form}_{\mathcal{L}}\forall\alpha\in\mathrm{Asn}(\phi)\left(S(\phi,\alpha)\equiv S(\phi[\alpha],\varepsilon)\right).$$

$$\forall\phi\in\mathrm{Form}^{\leq 1}_{\mathcal{L}}\forall s,t\forall\alpha\in\mathrm{Asn}(s)\forall\beta\in\mathrm{Asn}(t)\left(s^\alpha=t^\beta\rightarrow S(\phi[s/v],\alpha)\equiv S(\phi[t/v],\beta)\right).$$

The above hold (provably in CS^-) if ϕ is an atomic formula and clearly are preserved by taking disjunctions, negations, and applying existential quantification. However, in order to secure the step for $\exists$ we need a Π_1 assumption, saying that the equivalence

$$S(\psi,\alpha)\equiv S(\psi[\alpha],\varepsilon)$$

[10] Or, to be more accurate, by objects of size exponential in p.

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

holds under an arbitrary assignment α. It turns out, however, that such assumptions can be expressed with a Δ_0 formula. Firstly, the above is clearly equivalent to

$$\forall \alpha \in \mathrm{Asn}(\psi) S(\psi \equiv \psi[\alpha], \alpha). \tag{1}$$

Secondly, if S commuted with the blocks of universal quantifiers, the above could have been further reduced to

$$S(\mathrm{ucl}(\psi \equiv \psi[\alpha]), \varepsilon), \tag{2}$$

where $\mathrm{ucl}(\cdot)$ is a (definable) function which given a formula returns its universal closure. Our problem thus reduces to showing the equivalence between conditions of types (1) and (2). The standard strategy is to use induction on the length of the quantifier prefix. However, the proof of this once again uses Π_1 induction. To bypass this problem, we shall first establish commutation with blocks of uniformly bounded universal quantifiers, i.e., the principle

$$\forall \phi \forall x \big(S(\mathrm{bucl}(\phi, x), \varepsilon) \equiv \forall \alpha \prec [x]_\phi \big(\alpha \in \mathrm{Asn}(\phi) \to S(\phi, \alpha)\big)\big), \tag{B$\forall$C}$$

where $\mathrm{bucl}(\phi, x)$ denotes the universal closure of ϕ, in which every quantifier in the prefix is bounded by (the term) x and $\alpha \prec [x]_\phi$ says that $\mathrm{dom}(\alpha) = \mathrm{FV}(\phi)$ and each value of α is less than x. Having this, we will express $\forall \alpha S(\phi, \alpha)$ in a Δ_0 way via

$$S(\forall v \mathrm{bucl}(\phi, v), \varepsilon),$$

where v is a variable which do not occur in ϕ. We now proceed to the details.

3.3. The proof. One more preparation step will be helpful. We shall expand the language $\mathcal{L}_S$ with symbols for all primitive recursive functions and extend CS_0 with the defining equations of them. Let $\mathcal{L}_S^+$ denote the expanded language and CS_0^+ the extended theory. Since, trivially, CS_0^+ is $\mathcal{L}_S$-conservative over CS_0, it is sufficient to prove (GR(Th)) in CS_0^+. Let us observe that the latter theory proves Δ_0 induction for the language with new function symbols.

Proposition 3.1. $\mathrm{CS}_0^+ \vdash I\Delta_0(\mathcal{L}_S^+)$.

Proof. Fix a model $\mathcal{M} \models \mathrm{CS}_0^+$ and a $\Delta_0(\mathcal{L}_S^+)$ formula with parameters $\phi(x)$. Without loss of generality all the terms occurring in $\phi(x)$ are of the form $f(\underline{n}, y)$ where $f(x, y)$ is a two place p.r. function. Let $\psi_f(x, y, z)$ be the Δ_0 formula of $\mathcal{L}$ defining the graph of f. Fix an arbitrary $a \in M$, assume that $\phi(a)$ holds, and let b be greater than a and all parameters from $\phi(x)$. Without loss of generality assume that b is nonstandard. Let c be such that

$$\mathcal{M} \models \forall x < b \forall y < b \psi_f(x, y) < c.$$

Now let $\phi'(x)$ result from $\phi(x)$ by recursively changing all subformulae of $\phi(x)$ of the form

$$R(f(\underline{k}, x), f(\underline{n}, y))$$

into

$$\exists z \exists w \big(z < c \wedge w < c \wedge \psi_f(\underline{n}, y, w) \wedge \psi_f(\underline{k}, x, z) \wedge R(w, z)\big),$$

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

where z, w are fresh variables. $\phi'(x)$ is a $\Delta_0(\mathcal{L}_S)$ formula and clearly we have

$$\mathcal{M} \models \forall x < b \ \big(\phi(x) \equiv \phi'(x)\big).$$

Hence, since $a < b$ and $\phi(a)$ holds, $\phi'(a)$ holds as well. Since for $\phi'(x)$ we have an induction axiom there is the least $d < a$ such that $\phi'(d)$ holds. Hence d is the least element satisfying $\phi(x)$. □

In the proof below we shall use the following primitive recursive functions (we identify p.r. relations with their characteristic functions):

- $\prec$ is a primitive recursive product ordering on functions. That is, if α and β are two functions, then $\alpha \prec \beta$ holds if α is smaller than β in the product ordering, i.e.,

  $$\mathrm{dom}(\alpha) = \mathrm{dom}(\beta) \wedge \forall x \in \mathrm{dom}(\alpha) \ \ \alpha(x) < \beta(x).$$

  $\preccurlyeq$ is the partial ordering based on $\prec$.
- $\mathrm{bucl}(\phi, x)$ denotes the universal closure of ϕ, in which every quantifier in the prefix is bounded by (the term) x.
- $\mathrm{bucl}(\phi, x, c)$ returns the formula

  $$\forall v_{i_1} < x \dots \forall v_{i_y} < x \ \ \phi,$$

  where $x_{i_1}, \dots, x_{i_y}$ are all the elements of the set c listed in the order of decreasing indices (i.e., $i_y < i_{y-1} < \dots < i_1$). In particular x has to be a term and c a set of variables. Officially $\forall x < t\psi$ abbreviates $\forall x\big(x < t \to \psi\big)$; hence the above formula is slightly more complicated than it seems to be. Moreover c need not contain uniquely variables which are free in ϕ; hence some of the quantifiers in the prefix of $\mathrm{bucl}(\phi, x, c)$ might be dummy.
- $[x]_c$ returns the constant function assigning value x to every variable from the set c.
- For a coded set of variables c and a number a, $c{\uparrow_a}$ denotes the set consisting of first a elements of c and $c{\downarrow^a}$ the set consisting of last a elements of c. For simplicity we assume that the ordering of variables is given by their indices.
- Syntactical relations mentioned at the beginning of Section 2.

Lemma 3.2 (CS_0^+). *For every formula ϕ, every a, every set of variables c, and every assignment α for* $\mathrm{bucl}(\phi, \underline{a}, c)$,

$$S(\mathrm{bucl}(\phi, \underline{a}, c), \alpha) \equiv \forall \beta \prec [a]_c S(\phi, \alpha \cup \beta{\restriction}.).$$

Proof. Fix ϕ, a, c and let b be the cardinality of c. We formalize the standard argument on the length of the quantifier prefix in $\mathrm{bucl}(\phi, \underline{a}, c)$: starting from the assumption that $\forall \beta \prec [a]_c S(\phi, \alpha \cup \beta{\restriction}.)$ we check that we can prefix ϕ with b quantifiers and arrive at $S(\mathrm{bucl}(\phi, \underline{a}, c), \alpha)$. Formally, we use induction on y up to b in the $\Delta_0(\mathcal{L}_S^+)$-formula

$$\forall \beta \prec [a]_{c{\downarrow^{b-y}}} S(\mathrm{bucl}(\phi, \underline{a}, c{\uparrow_y}), \alpha \cup \beta{\restriction}.) \equiv \forall \beta \prec [a]_c S(\phi, \alpha \cup \beta{\restriction}.).$$

Observe that $c{\uparrow_b} = c{\downarrow^b} = c$, $c{\downarrow^0} = \emptyset$. Hence $\mathrm{bucl}(\phi, \underline{a}, c{\uparrow_0}) = \phi$, $[a]_{c{\downarrow^b}} = [a]_c$. Moreover, ε is the unique assignment γ satisfying $\gamma \prec [a]_\emptyset$. Consequently, for $y = b$

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

the left-hand side is equivalent to $S(\mathrm{bucl}(\phi,\underline{a},c),\alpha)$ and for $y=0$ both sides of the equivalence are the same.

Assume that the inductive hypothesis holds for d and let i_d be the index of the $(d+1)$-st variable in c. Observe that

$$\mathrm{bucl}(\phi,\underline{a},c{\uparrow}_{d+1}) = \forall v_{i_d} < \underline{a}^\frown \mathrm{bucl}(\phi,\underline{a},c{\uparrow}_d).$$

Hence by compositional axioms, $\forall\beta \prec [a]_{c\downarrow^{b-(d+1)}} S(\mathrm{bucl}(\phi,\underline{a},c{\uparrow}_{d+1}),\alpha\cup\beta\restriction.)$ is equivalent to

$$\forall\beta \prec [a]_{c\downarrow^{b-(d+1)}} \forall\gamma \geq_{v_{i_d}} \alpha\cup\beta\restriction_{\mathrm{bucl}(\phi,\underline{a},c\uparrow_{d+1})} \big(\gamma(v_{i_d}) < a \to S(\mathrm{bucl}(\phi,\underline{a},c{\uparrow}_d),\gamma\restriction.)\big). \tag{$*$}$$

Observe that the above is equivalent to

$$\forall\beta\big(\beta \prec [a]_{c\downarrow^{b-d}} \to S(\mathrm{bucl}(\phi,\underline{a},c{\uparrow}_d),\alpha\cup\beta\restriction.)\big),$$

which, by induction hypothesis, is equivalent to $\forall\beta\prec[a]_c S(\phi,\alpha\cup\beta\restriction.)$. □

The above lemma motivates the following abbreviation: we define

$$\mathrm{cl}(\phi,c) := \forall v^\frown \mathrm{bucl}(\phi,v,c).$$

In the above v is a variable with the least index among those which do not occur in ϕ. In particular $\mathrm{cl}(x,y)$ is a (partial) primitive recursive function, so we have a symbol for it in CS_0^+. $\mathrm{cl}(\phi)$ abbreviates $\mathrm{cl}(\phi,\mathrm{FV}(\phi))$.

Now, the following corollary clearly follows from Lemma 3.2.

Corollary 3.3 (CS_0^+). *For all ϕ, $c\subseteq\mathrm{Var}$ and $\alpha\in\mathrm{Asn}(\mathrm{cl}(\phi,c))$ it holds that* $S(\mathrm{cl}(\phi,c),\alpha)\equiv\forall\beta\in\mathrm{Asn}\big(\mathrm{dom}(\beta)=c\to S(\phi,\alpha\cup\beta\restriction.)\big)$.

Proof. Fix ϕ and c. By the compositional axioms and Lemma 3.2 $S(\forall v\mathrm{bucl}(\phi,v,c),\alpha)$ is equivalent to $\forall x\forall\beta\prec[x]_c S(\phi,\alpha\cup\beta\restriction.)$, which is clearly equivalent to $\forall\beta\big(\mathrm{dom}(\beta)=c\to S(\phi,\alpha\cup\beta\restriction.)\big)$, since each assignment for ϕ is dominated by an assignment of the form $[a]_\phi$ for some a. □

Now, we demonstrate how to use the above corollary for establishing, within CS_0^+, the induction on the buildup of formulae. It will be convenient to isolate a few more definitions.

Definition 3.4 (PA). An *occurrence* of a variable v in a formula ϕ (term s) is a path in a syntactic tree of ϕ (term s) ending with v. An occurrence of a subformula ψ of a formula ϕ is defined analogously. The fact that ψ is a subformula occurrence in ϕ is denoted $\psi\leq_o\phi$. The (coded) set of occurrences of variables in a formula ϕ (term s) will be denoted $\mathrm{Occ}(\phi)$ ($\mathrm{Occ}(s)$).

A *substitution of terms for a formula* ϕ is a (coded) function η such that $\mathrm{dom}(\eta)\subseteq\mathrm{Occ}(\phi)$ and $\mathrm{rg}(\eta)\subseteq\mathrm{ClTerm}$. For a formula ϕ, $\phi[\eta]$ denotes the result of applying η to ϕ.

If η is a substitution of terms for ϕ and ψ is a subformula of ϕ, then η naturally gives rise to a substitution of terms for ψ (we look only at those paths that pass through ψ and take their suffixes starting from ψ). Such a substitution will be denoted by $\eta\restriction_\psi$ or simply $\eta\restriction.$ if it is clear from context which formula should occur in the subscript.

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

The substitution of terms η for a formula ϕ *agrees* with an assignment $\alpha \in \mathrm{Asn}(\phi)$ if whenever $p \in \mathrm{dom}(\eta)$ is an occurrence of a variable v in ϕ, then $(\eta(p))^\circ = \alpha(v)$. Let v_p denote the variable whose occurrence p is.

Let ϕ be a formula and ψ an occurrence of its subformula. $\mathrm{Var}(\phi/\psi)$ denotes the (coded) set of variables whose occurrence in ψ is free in ψ but bounded in ϕ.

For a subformula ψ of ϕ define $\mathrm{cl}_\phi(\psi) := \forall v^\frown \mathrm{bucl}(\psi, v, \mathrm{Var}(\phi/\psi))$ where v is the least variable not occurring in ϕ. △

Example 3.5. Let $\psi := \big(x_0 = x_1 \wedge \exists x_2(x_2 = ((0+1)+1)\cdot x_0)\big)$ and $\phi := (\forall x_0 \psi) \wedge x_0 = x_0$. Then

$$\mathrm{Var}(\phi/\psi) = \{x_0\}.$$

Consequently $\mathrm{cl}_\phi(\psi) = \forall x_3 \forall x_0 < x_3\big(x_0 = x_1 \wedge \exists x_2\big(x_2 = ((0+1)+1)\cdot x_0\big)\big)$. △

Lemma 3.6 (CS_0^+). *Assume that ϕ is a formula, $\zeta \in \mathrm{Asn}(\phi)$, and η is a substitution of terms for ϕ which agrees with ζ. Then it holds that $S(\phi \equiv (\phi[\eta]), \zeta)$.*

Proof. Fix a formula ϕ, any assignment $\zeta \in \mathrm{Asn}(\phi)$, and a substitution of terms η which agrees with ζ. We reason by induction on y in the $\Delta_0(\mathcal{L}_S^+)$-formula

$$\theta(y) := \forall \psi \leq_o \phi\big(\psi \in \mathrm{dp}(y) \to S(\mathrm{cl}_\phi(\psi \equiv \psi[\eta{\restriction}_\psi]), \zeta{\restriction}.)\big).$$

Let us observe that $S(\mathrm{cl}_\phi(\psi \equiv \psi[\eta{\restriction}_\psi]), \zeta{\restriction}.)$ makes sense: each occurrence of a free variable of $\psi \equiv \psi[\eta{\restriction}_\psi]$ is either an occurrence of a free variable of ϕ and hence the variable gets assigned a value by $\zeta{\restriction}_\psi$, or is bounded in ϕ and so, belonging to $\mathrm{Var}(\phi/\psi)$, gets bounded by a quantifier occurring in a prefix of $\mathrm{cl}_\phi(\psi \equiv \psi[\eta{\restriction}_\psi])$.

Let z be the least variable not occurring in ϕ. We show that $\theta(0)$ holds. The unique sentences of depth 0 are atomic sentences, so let us fix two terms s, t and argue that

$$S\big(\mathrm{cl}_\phi(s = t \equiv (s = t[\eta{\restriction}_{s=t}])), \zeta{\restriction}.\big)$$

holds. Let $c = \mathrm{Var}(\phi/\psi)$. Consequently

$$\mathrm{cl}_\phi(s = t \equiv (s = t[\eta{\restriction}_{s=t}])) = \forall z^\frown \mathrm{bucl}(s = t \equiv (s = t[\eta{\restriction}_{s=t}]), z, c).$$

Call the above sentence on the right-hand side ξ. By Corollary 3.3 we know that $S(\xi, \zeta{\restriction}.)$ is equivalent to

$$\forall \alpha \in \mathrm{Asn}\big(\mathrm{dom}(\alpha) = c \to S(s = t \equiv (s = t[\eta{\restriction}_{s=t}]), \zeta{\restriction}_\xi \cup \alpha)\big).$$

The last sentence holds, since, by the compositional conditions for atomic sentences and connectives, it is equivalent to the assertion that for every $\alpha \in \mathrm{Asn}$ such that $\mathrm{dom}(\alpha) = c$

$$(s^{\zeta{\restriction}_\xi \cup \alpha} = t^{\zeta{\restriction}_\xi \cup \alpha}) \equiv (s[\eta{\restriction}_{s=t}]^{\zeta{\restriction}_{s=t[\eta{\restriction}_{s=t}]} \cup \alpha} = t[\eta{\restriction}_{s=t}]^{\zeta{\restriction}_{s=t[\eta{\restriction}_{s=t}]} \cup \alpha}). \tag{\$}$$

To avoid double restrictions let us abbreviate $\zeta{\restriction}_\xi$ with ζ_ξ. Observe that $\zeta_\xi{\restriction}_{s=t} = \zeta_\xi$. To prove (\$) it is sufficient to show that each occurrence of a variable in either s or t gets assigned the same value on both sides. This holds since if $o \in Occ(s)$, then either $o \in \mathrm{dom}(\eta{\restriction}_{s=t})$ or not. In the former case by our assumption $\zeta_\xi(v_o) = \zeta(v_o) = (\eta(o))^\circ$. If $o \notin \mathrm{dom}(\eta{\restriction}_{s=t})$, then $o \notin \mathrm{dom}(\eta)$ and $v_o \in \mathrm{dom}(\zeta_\xi{\restriction}_{s=t[\eta{\restriction}_{s=t}]} \cup \alpha) \subseteq \mathrm{dom}(\zeta_\xi \cup \alpha)$, and hence o get assigned the same value on both sides of the equivalence. Consequently

$$s[\eta\restriction_{s=t}]^{\zeta_\xi\restriction_{s=t[\eta\restriction_{s=t}]}\cup\alpha} = s^{\zeta_\xi\cup\alpha}.$$

The same holds for t in place of s. Now assume that the thesis holds for y and consider (an occurrence of) ψ of depth $y+1$. We shall do the case of $\psi = \psi_1 \vee \psi_2$ and $\psi = \exists v\psi_3$. We treat them simultaneously. As previously put $\xi := \mathrm{cl}_\phi\big(\psi \equiv (\psi[\eta\restriction_\psi])\big)$, $\xi_i := \mathrm{cl}_\phi\big(\psi_i \equiv (\psi_i[\eta\restriction_{\psi_i}])\big)$, $c = \mathrm{Var}(\phi/\psi)$, and $c_i = \mathrm{Var}(\phi/\psi_i)$. Applying the inductive assumption and Corollary 3.3 we have for $i \in \{1,2,3\}$

$$\forall\alpha \in \mathrm{Asn}\big(\mathrm{dom}(\alpha) = c_i \to S(\psi_i \equiv (\psi_i[\eta\restriction_{\psi_i}]), \zeta\restriction_{\xi_i} \cup \alpha)\big).$$

Now observe that $\zeta\restriction_{\xi_i}\restriction_{\psi_i} = \zeta\restriction_{\xi_i}$ and if $\mathrm{dom}(\alpha) = c_i$, then $\alpha\restriction_{\psi_i} = \alpha$. By this and compositional conditions, for arbitrary α such that $\mathrm{dom}(\alpha) = c_i$, the succedent of the above implication is equivalent to

$$S(\psi_i, \zeta\restriction_{\xi_i} \cup \alpha) \equiv S(\psi_i[\eta\restriction_{\psi_i}], \zeta\restriction_{\xi_i}\restriction_{\psi_i[\eta\restriction_{\psi_i}]} \cup \alpha\restriction_{\psi_i[\eta\restriction_{\psi_i}]}). \tag{3}$$

In the case $\psi = \psi_1 \vee \psi_2$ we have $c = c_1 \cup c_2$. Now, by compositional conditions, the following are equivalent for an arbitrary assignment α such that $\mathrm{dom}(\alpha) = c$:

$$S(\psi_1 \vee \psi_2 \equiv (\psi_1 \vee \psi_2[\eta\restriction_{\psi_1\vee\psi_2}]), \zeta\restriction_\xi \cup \alpha). \tag{4}$$

$$\left(\bigvee_{i\in\{0,1\}} S(\psi_i, \zeta\restriction_\xi\restriction_{\psi_i} \cup \alpha\restriction_{\psi_i})\right) \equiv \left(\bigvee_{i\in\{0,1\}} S(\psi_i[\eta\restriction_{\psi_i}], \zeta\restriction_\xi\restriction_{\psi_i[\eta\restriction_{\psi_i}]} \cup \alpha\restriction_{\psi_i[\eta\restriction_{\psi_i}]})\right). \tag{5}$$

Now observe that $\zeta\restriction_\xi\restriction_{\psi_i} = \zeta\restriction_{\xi_i}\restriction_{\psi_i} = \zeta\restriction_{\xi_i}$. Indeed, v is a free variable in $\mathrm{cl}_\phi(\psi \equiv \psi[\eta\restriction_\psi])$ $(= \xi)$ and a free variable in ψ_i, if and only if v is a free variable in $\mathrm{cl}_\phi(\psi_i \equiv \psi_i[\eta\restriction_{\psi_i}])$ $(= \xi_i)$ and in ψ_i. Hence $\mathrm{dom}(\zeta\restriction_\xi\restriction_{\psi_i}) = \mathrm{dom}(\zeta\restriction_{\xi_i}\restriction_{\psi_i})$ and this completes our claim. The same reasoning shows also that $\zeta\restriction_\xi\restriction_{\psi_i[\eta\restriction_{\psi_i}]} = \zeta\restriction_{\xi_i}\restriction_{\psi_i[\eta\restriction_{\psi_i}]}$. Finally, if $\mathrm{dom}(\alpha) = c$, then $\mathrm{dom}(\alpha\restriction_{\psi_i}) = c_i$. It follows that (3) implies the above condition (5) and the case of $\vee$ is done.

In the case of $\exists$, we observe that

$$c_3 = \begin{cases} c \cup \{v\}, \text{ if } v \in \ \mathrm{FV}(\psi_3), \\ c, \text{ otherwise.} \end{cases}$$

The following are equivalent for every $\alpha \in \mathrm{Asn}$ such that $\mathrm{dom}(\alpha) = c$:

$$S(\exists v\psi_3 \equiv (\exists v\psi_3[\eta\restriction_{\exists v\psi_3}]), \zeta\restriction_\xi \cup \alpha). \tag{6}$$

$$\begin{aligned}&\Big(\exists\beta \geq_v (\alpha \cup \zeta\restriction_\xi)\restriction_{\exists v\psi_3} S(\psi_3, \beta\restriction_{\psi_3})\Big)\\ &\qquad\equiv \Big(\exists\beta \geq_v (\alpha \cup \zeta\restriction_\xi)\restriction_{\exists v\psi_3[\eta\restriction_{\exists v\psi_3}]} S(\psi_3[\eta\restriction_{\psi_3}], \beta\restriction_{\psi_3[\eta\restriction_{\psi_3}]})\Big).\end{aligned} \tag{7}$$

$$\begin{aligned}&\Big(\exists\beta \geq_v (\alpha \cup \zeta\restriction_\xi) S(\psi_3, \beta\restriction_{\psi_3})\Big)\\ &\qquad\equiv \Big(\exists\beta \geq_v (\alpha \cup \zeta\restriction_\xi)\restriction_{\exists v\psi_3[\eta\restriction_{\exists v\psi_3}]} S(\psi_3[\eta\restriction_{\psi_3}], \beta\restriction_{\psi_3[\eta\restriction_{\psi_3}]})\Big).\end{aligned} \tag{8}$$

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

Observe that $v \notin \mathrm{dom}(\zeta\upharpoonright_{\xi}) \cup \mathrm{dom}(\zeta\upharpoonright_{\xi_3})$, because every occurrence of v in $\exists v\psi_3$ is bounded in ϕ. Hence $\zeta\upharpoonright_{\xi} = \zeta\upharpoonright_{\xi_3}$, and consequently $\zeta\upharpoonright_{\xi}\upharpoonright_{\psi_3} = \zeta\upharpoonright_{\xi_3}\upharpoonright_{\psi_3} = \zeta\upharpoonright_{\xi_3}$. With this observation the proof in the case $v \notin \mathrm{FV}(\psi_3)$ is straightforward, for (8) immediately reduces to (for all α such that $\mathrm{dom}(\alpha) = c_3$)

$$\Big(S(\psi_3, \alpha \cup \zeta\upharpoonright_{\xi_3})\Big) \equiv \Big(S(\psi_3[\eta\upharpoonright_{\psi_3}], (\alpha \cup \zeta\upharpoonright_{\xi_3})\upharpoonright_{\psi_3[\eta\upharpoonright_{\psi_3}]})\Big).$$

The above is the same as our induction assumption. So we may assume that $v \in \mathrm{FV}(\psi_3)$. Now fix α such that $\mathrm{dom}(\alpha) = c$. Suppose first that $\beta \geq_v (\alpha \cup \zeta\upharpoonright_{\xi_3})$ is such that $S(\psi_3, \beta\upharpoonright_{\psi_3})$ holds. Let us observe that in this case, $\beta\upharpoonright_{\psi_3} = \beta$. Moreover $\mathrm{dom}(\alpha) \cap \mathrm{dom}(\zeta\upharpoonright_{\xi_3}) = \emptyset$; hence for some α' such that $\mathrm{dom}(\alpha') = c_3, \beta = \alpha' \cup \zeta\upharpoonright_{\xi_3}$. Hence $S(\psi_3[\eta\upharpoonright_{\psi_3}], \beta\upharpoonright_{\psi_3[\eta\upharpoonright_{\psi_3}]})$ follows by induction assumption (3). It is left to show that $\beta\upharpoonright_{\psi_3[\eta\upharpoonright_{\exists v\psi_3}]} \geq_v (\alpha \cup \zeta\upharpoonright_{\xi})\upharpoonright_{\exists v\psi_3[\eta\upharpoonright_{\exists v\psi_3}]}$. This holds since no occurrence of v is in $\mathrm{dom}(\eta\upharpoonright_{\psi_3}) = \mathrm{dom}(\eta\upharpoonright_{\exists v\psi_3})$ and $\beta \geq_v (\alpha \cup \zeta\upharpoonright_{\xi_3})$. So now assume that for some $\beta \geq_v (\alpha \cup \zeta\upharpoonright_{\xi_3})\upharpoonright_{\exists v\psi_3[\eta\upharpoonright_{\exists v\psi_3}]} S(\psi_3[\eta\upharpoonright_{\psi_3}], \beta\upharpoonright_{\psi_3[\eta\upharpoonright_{\psi_3}]})$ holds. As no occurrence of v is in $\mathrm{dom}(\eta\upharpoonright_{\psi_3})$, we can infer that $S(\psi_3[\eta\upharpoonright_{\psi_3}], \beta)$ holds. Extend α to α' such that $\mathrm{dom}(\alpha') = c_3$ and $\alpha'(v) = \beta(v)$. Then $(\alpha' \cup \zeta\upharpoonright_{\xi_3})\upharpoonright_{\psi_3[\eta\upharpoonright_{\psi_3}]} = \beta$ and by (3) we obtain $S(\psi_3, \alpha' \cup \zeta\upharpoonright_{\xi_3})$. Hence there exists β' such that $\beta' \geq_v \alpha \cup \zeta\upharpoonright_{\xi_3}$ and $S(\psi_3, \beta')$. This ends the whole proof. □

COROLLARY 3.7 (CS_0^+). *For every $\phi, \psi, \alpha \in \mathrm{Asn}(\phi), \beta \in \mathrm{Asn}(\psi)$, if $\phi[\alpha] = \psi[\beta]$, then $S(\phi, \alpha) \equiv S(\psi, \beta)$.*

PROOF. By Lemma 3.6 $S(\phi, \alpha)$ is equivalent to $S(\phi[\alpha], \varepsilon)$ and $S(\psi, \beta)$ is equivalent to $S(\psi[\beta], \varepsilon)$. □

COROLLARY 3.8. $\mathrm{CS}_0^+ \vdash \forall \phi(v) S(\mathrm{Ind}(\phi(v)), \varepsilon)$.

PROOF. By the previous considerations at the beginning of this section, for a fixed $\phi(v)$, $S(\mathrm{Ind}(\phi(v)), \varepsilon)$ is equivalent to

$$S(\phi[0/v], \varepsilon) \wedge \forall x\big(S(\phi(v), [x]) \rightarrow S(\phi[v+1/v], [x])\big) \longrightarrow \forall x S(\phi(v), [x]).$$

By Lemma 3.6 and Corollary 3.7 we have

$$S(\phi[0/v], \varepsilon) \equiv S(\phi(v), [0]),$$
$$\forall x\big(S(\phi[v+1/v], [x]) \equiv S(\phi(v), [x+1])\big).$$

Hence, finally $S(\mathrm{Ind}(\phi(v)), \varepsilon)$ is equivalent to the following axiom of $\Delta_0(\mathcal{L}_S^+)$-induction:

$$S(\phi(v), [0]) \wedge \forall x\big(S(\phi(v), [x]) \rightarrow S(\phi(v), [x+1])\big) \longrightarrow \forall x S(\phi(v), [x]). \quad □$$

Recall that $\phi(t/v)$ denotes the substitution of t for all (free) occurrences of v in $\phi(v)$.

COROLLARY 3.9 (CS_0^+). *For every formula $\phi(v)$, term t (possibly having some variables), which is substitutable for v in $\phi(v)$ and every $\alpha \in \mathrm{Asn}(\phi(t/v))$, if $S(\forall v\phi(v), \alpha\upharpoonright_{.})$, then $S(\phi(t/v), \alpha)$.*

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

Proof. Fix $\phi(v)$, t, α, as above and suppose $S(\forall v\phi(v), \alpha\restriction_{.})$. By compositional conditions we know that for every β such that $\beta \geq_v \alpha\restriction_{\forall v\phi(v)}$, $S(\phi(v), \beta\restriction_{.})$ holds. Define β such that $\beta\restriction_{\forall v\phi(x)} = \alpha\restriction_{\forall v\phi(x)}$ and $\beta(v) = t^{\alpha}$. Hence, by our assumption we have $S(\phi(v), \beta)$. Let η_0 be a substitution of $t[\alpha]$ for every occurrence of v in $\phi(v)$. Then η_0 agrees with β, so by Lemma 3.6 we have $S(\phi(v)[\eta_0], \beta\restriction_{.})$. Observe that $\beta\restriction_{\phi(v)[\eta_0]} = \alpha\restriction_{\phi(v)[\eta_0]}$; hence we have $S(\phi(v)[\eta_0], \alpha\restriction_{.})$. Let π be a (coded) set of occurrences of the free variables from $\phi(t/v)$ that are within the new occurrences of t (observe that there might be some occurrences of t in $\phi(v)$). Let η be a substitution of numerals such that for every occurrence $p \in \pi$, $\eta(p) = \underline{\alpha(v_p)}$ (recall that v_p is the variable whose occurrence is p). By the definition of η, we have $\phi(v)[\eta_0] = \phi(t/v)[\eta]$; hence we can conclude that $S(\phi(t/v)[\eta], \alpha\restriction_{.})$ holds. Since η agrees with α, Lemma 3.6 yields $S(\phi(t/v), \alpha)$. □

Theorem 3.10. $\mathrm{CS}_0 \vdash \forall\phi\big(\mathrm{Pr}^{S}_{\emptyset}(\phi) \rightarrow S(\phi, \varepsilon)\big)$.

Proof. We reason in CS_0^{+}. We fix a sequent calculus for the first-order logic with equality, as in [24] (this choice is just a matter of convenience[11]). We fix a proof p of a sentence ϕ and by induction on its length argue that whenever a sequent $\Gamma \Rightarrow \Delta$ occurs in p, then

$$S(\mathrm{cl}\Big(\bigwedge\Gamma \rightarrow \bigvee\Delta\Big), \varepsilon) \tag{9}$$

holds, where $\bigwedge\Gamma$ and $\bigvee\Gamma$ denote (the canonically parenthesized) conjunction and disjunction over sentences from sets Γ and Δ, respectively. To simplify the notation $\bigwedge\Gamma \rightarrow \bigvee\Delta$ will be abbreviated using the sequent notation as $\Gamma \Rightarrow \Delta$. In the course of the induction we rely on the fact that the following sentence is provable in CS_0:

$$\forall\alpha \in \mathrm{Asn}(\Gamma)\Big(S\Big(\bigvee\Gamma, \alpha\Big) \equiv \exists\phi \in \Gamma S(\phi, \alpha\restriction_{\phi})\Big). \tag{DC}$$

The proof of (DC) in CT_0 consists in a straightforward induction on the size of Γ and a similar argument can be given in the case of $\bigwedge$ yielding a dual equivalence (see also [3, 25] for precise arguments). We go back to the main induction on the length of the fixed proof p. In the base step we have to establish that all initial sequents satisfy (9). These include initial sequents for equality and all sequents of the form $\phi \Rightarrow \phi$. In both cases the proof follows the same pattern: first using Corollary 3.3 we get rid of the quantifier prefix and then verify that the formula following it is satisfied by every assignment. In the case of the initial sequents for equality we use the conditions for atomic sentences from CS^{-}, in case of $\phi \Rightarrow \phi$ we use the axioms for $\neg$ and $\vee$.

In the induction step the cases of quantifier rules are the unique non-obvious ones. Let us consider the *dictum de omni* rule:

$$\frac{\Gamma, \phi(t) \Rightarrow \Delta}{\Gamma, \forall v\phi(v) \Rightarrow \Delta},$$

[11] Strictly speaking this calculus is formulated only for the language with both $\vee$ and $\wedge$ and both $\exists$, $\forall$, but we can always extend the language by defining the missing symbols and adding axioms defining them.

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

where $\phi(v)$ is a formula and t is free for v in $\phi(v)$. We may safely assume that v is a free variable in ϕ. For simplicity abbreviate $\phi(t/v)$ with simply $\phi(t)$. So suppose for every $\alpha \in \operatorname{Asn}\big((\Gamma + \phi(t)) \Rightarrow \Delta\big)$ it holds that

$$S\left((\Gamma + \phi(t)) \Rightarrow \Delta, \alpha\right). \tag{10}$$

Fix an arbitrary $\alpha \in \operatorname{Asn}\big((\Gamma + \forall v\phi(v)) \Rightarrow \Delta\big)$ and assume that for every $\theta \in \Gamma \cup \{\forall v\phi(v)\}$, $S(\theta, \alpha\restriction.)$ holds. Consider any $\beta \in \operatorname{Asn}\big((\Gamma + \phi(t)) \Rightarrow \Delta\big)$ such that $\beta\restriction_{\left(\Gamma+\forall v\phi(v)\right)\Rightarrow\Delta} = \alpha$. Since $S(\forall v\phi(v), \beta\restriction.)$, then by Lemma 3.9, $S(\phi(t), \beta\restriction.)$ as well. Hence for every $\theta \in \Gamma \cup \{\phi(t)\}$ it holds that $S(\theta, \beta\restriction.)$. By the induction assumption, for some $\psi \in \Delta$, $S(\psi, \beta\restriction.)$ and since $\beta\restriction_\psi = \alpha\restriction_\psi$, this ends the proof.

Now consider the rule of universal generalisation

$$\frac{\Gamma \Rightarrow \Delta, \phi(v)}{\Gamma \Rightarrow \Delta, \forall v\phi(v)},$$

where v does not occur free in the lower sequent. As previously assume that for all $\alpha \in \operatorname{Asn}\big(\Gamma \Rightarrow (\Delta + \phi(v))\big)$,

$$S\big(\Gamma \Rightarrow (\Delta + \phi(v)), \alpha\big).$$

Fix an arbitrary $\beta \in \operatorname{Asn}\big(\Gamma \Rightarrow (\Delta + \forall v\phi(v))\big)$ and arbitrary $\gamma \geq_v \beta$ and assume that for all $\psi \in \Gamma$, $S(\psi, \gamma\restriction.)$. Since $\gamma\restriction_\Gamma = \beta\restriction_\Gamma$ (v is not a free variable in Γ) it follows that there is $\psi \in \Delta \cup \{\phi(v)\}$ such that $S(\psi, \gamma\restriction.)$. If $\psi \in \Delta$, then we are done, since $\gamma\restriction_\Delta = \beta\restriction_\Delta$. Otherwise $S(\phi(v), \gamma\restriction.)$ and it follows that $S(\forall v\phi(v), \beta\restriction.)$, since γ was arbitrary.

Now, the thesis of the theorem follows, since, for arbitrary ϕ, if $\Pr^S_\emptyset(\phi)$ holds, then there is a sequent calculus proof of $\Gamma \Rightarrow \phi$, where for every $\psi \in \Gamma$ we have $S(\psi, \varepsilon)$. Hence, by the proof above we obtain

$$S\left(\operatorname{cl}\left(\bigwedge \Gamma \rightarrow \phi\right), \varepsilon\right),$$

and since $\bigwedge \Gamma \rightarrow \phi$ does not admit any free variables, then we have $S(\bigwedge \Gamma \rightarrow \phi, \varepsilon)$. The thesis follows by the conjunctive correctness and compositional conditions: since for every $\psi \in \Gamma$ we have $S(\psi, \varepsilon)$, then $S(\bigwedge \Gamma, \varepsilon)$ holds and an application of Modus Ponens yields $S(\phi, \varepsilon)$. □

Corollary 3.11. *For every Gödelized theory* Th, *we have*

$$\mathrm{CS}_0 + \forall x\big(\mathrm{Th}(x) \rightarrow S(x, \varepsilon)\big) \vdash \forall\phi\big(\Pr_{\mathrm{Th}}(\phi) \rightarrow S(\phi, \varepsilon)\big).$$

Hence,

$$\mathrm{CS}_0 + \forall x\big(\mathrm{Th}(x) \rightarrow S(x, \varepsilon)\big) \vdash \mathrm{REF}^\omega(\mathrm{Th}).$$

Proof. The first part follows directly from Theorem 3.10. Having it, we prove the second one: by induction on n we prove that

$$\mathrm{CS}_0 + \forall x\big(\mathrm{Th}(x) \rightarrow S(x, \varepsilon)\big) \vdash \forall\phi\big(\Pr_{\mathrm{REF}^n(\mathrm{Th})}(\phi) \rightarrow S(\phi, \varepsilon)\big).$$

For $n = 0$ this follows from the first part. Fix n and assume the thesis holds for it. By the compositional clauses for CS^- we have

$$\mathrm{CS}_0 + \forall x\big(\mathrm{Th}(x) \rightarrow S(x, \varepsilon)\big) \vdash \forall\phi\big(S(\ulcorner \Pr_{\mathrm{REF}^n(\mathrm{Th})}(\phi) \rightarrow \phi \urcorner, \varepsilon)\big).$$

Consequently, reapplying the first part of this corollary for $\mathrm{REF}^n(\mathrm{Th})$ substituted for Th we get the induction thesis for $n+1$. □

The corollary below easily follows from the corollary above and Corollary 3.8.

Corollary 3.12. $\mathrm{CS}_0 \vdash \forall\phi\big(\mathrm{Pr}_{\mathrm{PA}}(\phi) \rightarrow S(\phi,\varepsilon)\big)$. ⊞

Corollary 3.13. $\mathrm{CS}_0 \vdash \mathrm{REF}^{\omega}(\mathrm{PA})$. ⊞

3.4. Corollaries.

3.4.1. Compositional satisfaction vs. compositional truth. The above results transfer immediately to the setting of the following theory of truth:

Definition 3.14. CT^- is the $\mathcal{L}\cup\{T\}$ theory extending EA with the following axioms:

1. $\forall x\big(T(x) \rightarrow \mathrm{Sent}(x)\big)$.
2. $\forall s,t \in \mathrm{ClTerm}\big(T(s=t) \equiv (s)^{\circ} = (t)^{\circ}\big)$.
3. $\forall\phi,\psi \in \mathrm{Sent}\big(T(\phi\vee\psi) \equiv (T(\phi)\vee T(\psi))\big)$.
4. $\forall\phi \in \mathrm{Sent}\big(T(\neg\phi) \equiv \neg T(\phi)\big)$.
5. $\forall\phi(v) \in \mathrm{Form}^{\leq 1}\big(T(\exists v\phi(v)) \equiv \exists x T(\phi[x]\big)$.

As usual CT_n denotes the result of extending the following theory with induction axioms for Σ_n formulae of the extended language. △

Let $\mathcal{L}_T^+$ denote the extension of $\mathcal{L}_T$ with function symbols for all p.r. recursive functions and CT_0^+ denote the extension of CT_0 with all defining axioms for fresh functions symbols in $\mathcal{L}_T^+$. Then we have an analogue of Proposition 3.1:

Proposition 3.15. $\mathrm{CT}_0^+ \vdash \mathrm{I}\Delta_0(\mathcal{L}_T^+)$. ⊞

Now we show that the result on the provability of (GR(Th)) in CS_0 transfers to the setting with the truth predicate.

Corollary 3.16. $\mathrm{CT}_0 \vdash \forall\phi \in \mathrm{Sent}\big(\mathrm{Pr}_{\mathrm{PA}}(\phi) \rightarrow T(\phi)\big)$.

Proof. Work in CT_0^+ and put

$$S(x,y) := \mathrm{Form}(x) \wedge y \in \mathrm{Asn}(x) \wedge T(x[y]).$$

The above is a $\Delta_0(\mathcal{L}_T^+)$ formula, so obviously we have $\mathrm{I}\Delta_0(\mathcal{L}_S^+)$. Now, we show that $S(x,y)$ behaves compositionally. We focus on the $\exists$-axiom. Pick a formula $\phi(v)$ and $\alpha \in \mathrm{Asn}(\exists v\phi(v))$. Observe that the following equivalences hold:

$$\begin{aligned} S(\exists v\phi(v),\alpha) &\equiv T(\exists v\phi(v)[\alpha]) \\ &\equiv \exists x T(\phi[\alpha][x])) \\ &\equiv \exists\beta \geq_v \alpha T(\phi[\beta]) \\ &\equiv \exists\beta \geq_v \alpha S(\phi,\beta). \end{aligned}$$

In the second and the third equivalence we use the fact that $v \notin \mathrm{dom}(\alpha)$. Hence (in CT_0) by Theorem 3.10 we have

$$\forall\phi \in \mathrm{Sent}\big(\mathrm{Pr}_{\mathrm{PA}}(\phi) \rightarrow S(\phi,\varepsilon)\big).$$

Translating it back to the language with the truth predicate, we get our thesis. □

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

The proof of the above corollary proceeds by defining the satisfaction predicate satisfying CS_0^+ in CT_0^+. In fact, the same translation works also in the context of the non-inductive versions of both theories, CS^- and CT^-. However, it is not known, whether the reverse is true in the context of these theories, i.e., whether CS^- can define the truth predicate of CT^-. Using the Enayat–Visser method [7] of constructing pathological models for CS^- one can show that standard methods of defining truth from satisfaction do not work. However, the results from the previous section witness that Δ_0 induction is sufficient to overcome these deficiencies of CS^-.

Proposition 3.17. *The truth predicate satisfying* CT_0 *is definable in* CS_0.

Proof. Working in CS_0, put

$$T(x) := S(x, \varepsilon).$$

As previously, $T(x)$ is a $\Delta_0(\mathcal{L}_S)$ formula, so $\mathrm{CS}_0 \vdash \mathrm{I}\Delta_0(\mathcal{L}_T)$. Since sentences are the unique formulae for which ε is an assignment, so axiom 1. of CS^- implies the corresponding axiom of CT^-. Once again we focus on the compositional axioms for $\exists$. Working in CS_0 fix ϕ and without loss of generality assume that $v \in \mathrm{FV}(\phi)$. Observe that the following equivalences hold:

$$\begin{aligned} T(\exists v\phi) &\equiv S(\exists v\phi, \varepsilon) \\ &\equiv \exists \beta \geq_v \varepsilon S(\phi, \beta\restriction.) \\ &\equiv \exists x S(\phi, [x]) \\ &\equiv \exists x S(\phi[x], \varepsilon) \\ &\equiv \exists x T(\phi[x]). \end{aligned}$$

The proof of the fourth equivalence involves the crucial use of Lemma 3.6. □

3.4.2. Many faces theorem. Corollary 3.16 coupled with some known results from the literature, shows that the Global Reflection Principle is a very robust notion. Not only it is equivalent to bounded induction but is immune to, apparently significant, variations. This is summarized in the corollary below (we state it for the theory of compositional truth; however all the equivalences should transfer to the setting of a satisfaction predicate without significant changes). $\mathrm{Pr}^T_{\mathrm{Sent}}(\phi)$ asserts that ϕ is provable from the set of true sentences in pure sentential logic, while DC is a truth variant of the principles from the proof of Theorem 3.10.

Corollary 3.18. *Over* $\mathrm{CT}^- + \mathrm{EA}$ *the following theories are equivalent*:

1. $I\Delta_0(\mathcal{L}_T)$;
2. $\forall \phi\big(\mathrm{Pr}_{\mathrm{PA}}(\phi) \rightarrow T(\phi)\big)$;
3. $\forall \phi\big(\mathrm{Pr}_{\emptyset}(\phi) \rightarrow T(\phi)\big)$;
4. $\forall \phi\big(\mathrm{Pr}^T_{\emptyset}(\phi) \rightarrow T(\phi)\big)$;
5. $\forall \phi\big(\mathrm{Pr}^T_{\mathrm{Sent}}(\phi) \rightarrow T(\phi)\big)$;
6. DC.

The implication 3.⇒4. is established in [5]. The equivalence between 5. and 1. is demonstrated in [3]. Much later it was significantly fine-tuned in [6], yielding the implication 6.⇒1. ⊞

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

3.4.3. Fullness. The following is one of the most useful properties of CS_0. It implies that every model of CS_0 is full, a theorem first demonstrated by Wcisło and presented in [25]. The proof below is an observation also due to Wcisło which crucially uses the provability of $(\mathrm{GR}(\mathrm{Th}))$. It's proof is included also in [20] (Fact 33) but we give it here for completeness. In the definition below we fix a canonical elementary translation transforming a given formula ϕ into one in the Σ_n form. We assume that it formalizes in PA.

Recall (Definition 2.12) that $\phi(x)^\Sigma$ denotes the canonical Σ_c form of $\phi(x)$ and $\Sigma_c^* := \{\phi \mid \phi^\Sigma \in \Sigma_c\}$. Moreover recall that S_c denotes the restriction of S to all formulae which are equivalent to sentences of Σ_c complexity (in the sense of $\mathcal{M}$). More precisely

$$S_c := \{\langle \phi, \alpha\rangle \mid (\mathcal{M}, S) \models \phi \in \Sigma_c^* \wedge S(\phi^\Sigma, \alpha)\}.$$

Theorem 3.19. *Suppose that* $(\mathcal{M}, S) \models \mathrm{CS}^- + \mathrm{GR}(\mathrm{PA})$. *Then for every* c, $(\mathcal{M}, S_c) \models \mathrm{CS}(\Sigma_c^*)$.

Proof. Fix a model $(\mathcal{M}, S) \models \mathrm{CS}^- + \mathrm{GR}(\mathrm{PA})$ and $c \in M$. It will be easier to switch to the truth predicate, so put $T := S(x, \varepsilon)$. Since for every formula $\phi \in \Sigma_c^*$ and every $\alpha \in \mathrm{Asn}(\phi)$ we have

$$(\mathcal{M}, S) \models S(\phi, \alpha) \equiv T(\phi[\alpha]),$$

it is sufficient to show that $(\mathcal{M}, T_c) \models \mathrm{Ind}(\mathcal{L}_T)$, where T_c is the restriction of T to the formulae of Σ_c^* complexity, i.e.,

$$T_c := (\Sigma_c^*)^{\mathcal{M}} \cap (T)^{(\mathcal{M}, S)}.$$

From now on we work in $(\mathcal{M}, T)$. By the classical metamathematics of PA, for every c there is a formula Sat_{Σ_c} such that for every $\phi \in \Sigma_c^*$ and every $\alpha \in \mathrm{Asn}(\phi)$ we have

$$\mathrm{Pr}_{\mathrm{PA}}\left(\phi[\alpha] \equiv \mathrm{Sat}_{\Sigma_c}(\phi^\Sigma, \alpha)\right).$$

Hence, by $\mathrm{GR}(\mathrm{PA})$ we conclude that for every sentence $\phi \in \Sigma_c^*$

$$T(\phi) \equiv T(\mathrm{Sat}_{\Sigma_c}(\phi^\Sigma, \varepsilon)).$$

Put $\xi(v) := \mathrm{Sat}_{\Sigma_c}(v, \varepsilon)$ and $T'(x) := T(\xi[x]) \wedge x \in \Sigma_c^*$. Then we see that

$$T_c = (T')^{(\mathcal{M}, T)} \cap (\Sigma_c^*)^{\mathcal{M}}.$$

Consequently, T' satisfies the compositional axioms of CT^- for formulae from the Σ_c^* class. We shall now show $(\mathcal{M}, T') \models \mathrm{Ind}(\mathcal{L}_T)$. Thus let $\eta[T']$ be an arbitrary axiom of induction for a formula with T' (we mark all occurrences of T' in η). We may assume that $\eta[T']$ is in the semirelational form (as defined in [19]). Since, using the notation of [19], T' is of the form $T * \xi$, by Lemma 25 in [19] we have

$$(\mathcal{M}, T) \models \eta[T'] \equiv T(\eta[\xi]).$$

However, $\eta[\xi]$ is an axiom of induction (in the sense of $\mathcal{M}$) for an arithmetical formula $\eta[\xi]$, hence $T(\eta[\xi])$ by $\mathrm{GR}(\mathrm{PA})$. □

Remark 3.20. A very similar reasoning was used in Kotlarski in [15]. However various parts of this paper are negatively influenced by the significant gaps already discussed at the beginning of this section. We decided to reprove it in a rigorous way. Essentially the same proof is given in [20]. △

Corollary 3.21. *For any* $(\mathcal{M}, S) \models \mathrm{CS}^{-}$ *the following conditions are equivalent*:

1. $(\mathcal{M}, S) \models \mathrm{GR}(\mathrm{PA})$.
2. *For every* $c \in M$, $(\mathcal{M}, S) \models \mathrm{CS}(\Sigma^*_c)$.
3. $(\mathcal{M}, S) \models \mathrm{CS}_0$.

Proof. We show the remaining implication 2. ⇒ 3. By the classical fact in the metamathematics of PA, a subset of the model satisfies Δ_0-induction if and only if it is piecewise coded, i.e., it is sufficient to show

$$(\mathcal{M}, S) \models \forall c \exists d \forall x < c\big(S((x)_0, (x)_1) \equiv x \in d\big),$$

where $(x)_i$ denotes the i-th projection of x. Working in $(\mathcal{M}, S)$ fix an arbitrary c. Obviously, if a formula $\phi < c$ then $\phi \in \Sigma^*_c$. Hence, it is sufficient to find a d such that

$$(\mathcal{M}, S_c) \models \forall x < c\big(S((x)_0, (x)_1) \equiv x \in d\big).$$

This clearly can be done as S_c satisfies full induction. □

§4. Consequences of the global reflection principle. In this section we focus on the Δ_0-inductive *truth* predicate. We remind the Reader that by default all theories extend EA. We extend the result from the previous section and prove the following theorem.

Theorem 4.1. *For every* $\phi(x) \in \Sigma_1(\mathcal{L}_T)$ *and every* $n \in \omega$ *the following sentence is provable in* CT_0:

$$\forall x\big(\mathrm{Pr}^T_{\mathrm{UTB}_n}(\phi[x]) \rightarrow \phi(x)\big).$$

Corollary 4.2. $\mathrm{CT}_0 \vdash \Sigma_1(\mathcal{L}_T)\text{-REF}(\mathrm{UTB}^{-})$.

The above answers affirmatively the question of Beklemishev and Pakhomov from [2].

Convention 4.3. *Working in an extension of* UTB^{-}, *it makes sense to treat the predicate* T *as a theory composed of all true sentences. Thus, we shall often write* (*for a Gödelized theory* Th)

$$\Gamma\text{-REF}(\mathrm{Th} + T)$$

to denote the theory consisting of all sentences

$$\forall x\big(\mathrm{Pr}^T_{\mathrm{Th}}(\phi[x]) \rightarrow \phi(x)\big),$$

for $\phi(x) \in \Gamma$. △

Let us start by explaining that Theorem 4.1 really improves on the results from the previous section. Let Th be a Gödelized theory extending EA in a language $\mathcal{L}'$ and let $\mathrm{UTB}^{-}(\mathrm{Th})$ denote the extension of Th with UTB^{-} axioms. It is enough to observe that over EA

$$\Delta_0(\mathcal{L}_{\mathrm{UTB}^{-}(\mathrm{Th})})\text{-REF}(\mathrm{UTB}^{-}(\mathrm{Th})) \vdash \mathrm{GR}(\mathrm{Th}).$$

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

Indeed, working in $\mathrm{EA} + \Delta_0(\mathcal{L}_{\mathrm{UTB}^-(\mathrm{Th})})$-REF$(\mathrm{UTB}^-(\mathrm{Th}))$ fix an arbitrary $\mathcal{L}_{\mathrm{Th}}$ sentence ϕ and assume $\mathrm{Pr}_{\mathrm{Th}}(\phi)$. By the axioms of $\mathrm{UTB}^-(\mathrm{Th})$ we immediately obtain $\mathrm{Pr}_{\mathrm{UTB}^-(\mathrm{Th})}(\ulcorner T(\phi)\urcorner)$. Therefore, by the Δ_0 reflection for $\mathrm{UTB}^-(\mathrm{Th})$ we obtain, $T(\phi)$.

Proof of Theorem 4.1 starts with a lemma:[12]

Lemma 4.4. *For every* $\phi \in \Delta_0(\mathcal{L}_T)$,

$$\mathrm{CT}_0 \vdash \forall x \big(\mathrm{Pr}^T_{\mathrm{UTB}}(\phi[x]) \to \phi(x)\big).$$

Proof. Fix $(\mathcal{M}, T) \models \mathrm{CT}_0$, $\phi(x) \in \Delta_0(\mathcal{L}_T)$ and $a \in M$ and any proof p such that $(\mathcal{M}, T) \models \mathrm{Proof}^T_{\mathrm{UTB}}(p, \phi[a])$. Since in ϕ all the quantifiers are bounded, then there is a $b \in M$ such that for every $(\mathcal{M}', T')$ satisfying (1) $\mathcal{M} \subseteq_e \mathcal{M}'$ and (2) $T'\restriction_{<b} = T\restriction_{<b}$ we have

$$(\mathcal{M}, T) \models \phi(a) \iff (M', T') \models \phi(a). \tag{A}$$

Recall that for $X \subseteq M$, $X\restriction_{<b}$ denotes the set of elements of X below b. Fix any such b. In short, b is big enough so that any end-extension of $(\mathcal{M}, T)$ which agrees with T up to b is absolute with respect to $\phi(a)$. Without loss of generality assume that $p < b$ and let c be big enough so that any formula (with code) smaller than b belongs to $(\Sigma^*_{c-1})^{\mathcal{M}}$. By Theorem 3.19, $(\mathcal{M}, T_c) \models \mathrm{CT}(\Sigma^*_c)$. Now we work in $(\mathcal{M}, T_c)$. Consider the theory

$$\mathrm{Th} := \mathrm{PA} + \{\phi \in \Sigma^*_c \mid T(\phi)\}.$$

By GR(PA) (Corollary 3.12) Th is consistent and by the trivial conservativity proof for UTB it follows that UTB + Th is consistent as well. Hence, by the Arithmetized Completeness Theorem (for PA*) there is a full model $((\mathcal{N}, T'), \mathrm{Sat}_{(\mathcal{N},T')})$ such that

$$(\mathcal{M}, T_c) \models \big[(\mathcal{N}, T') \models_{\mathrm{Sat}_{(\mathcal{N},T')}} \mathrm{UTB} + \mathrm{Th}\big]. \tag{B}$$

Since $(\mathcal{N}, T')$ is strongly interpretable in $(\mathcal{M}, T_c)$, then, by Proposition 2.24, $\mathcal{M} \subseteq_e \mathcal{N}$. Since internally in $(\mathcal{M}, T_c)$, $(\mathcal{N}, T')$ is a model of UTB + Th, then every sentence occurring in p is true in $(\mathcal{N}, T')$. So we see that $(\mathcal{N}, T') \models \phi(a)$. We show that $T'\restriction_{<b} = T\restriction_{<b}$. Fix any sentence ψ such that $\psi < b$. Then, by the choice of c, $\psi \in \Sigma^*_{c-1}$ and hence the following conditions are equivalent:

1. $\psi \in T\restriction_{<b}$.
2. $\psi \in \mathrm{Th}$.
3. $(\mathcal{M}, T_c) \models \big[\mathcal{N} \models_{\mathrm{Sat}_{(\mathcal{N},T')}} \psi\big]$.
4. $(\mathcal{N}, T') \models T(\psi)$.
5. $\psi \in T'\restriction_{<b}$.

The equivalence between 3. and 4. follows by (B). Hence it follows by (A) that $(\mathcal{M}, T) \models \phi(a)$, which suffices to prove the claim. □

Remark 4.5. The above proof generalises to the case in which PA is replaced with a (formalized) theory Th in an expanded (at most countable) language $\mathcal{L}'$ (we assume a fixed Gödel coding of $\mathcal{L}'$) such that $\mathrm{Th} \vdash \mathrm{Ind}_{\mathcal{L}'}$. This allows us to obtain:

[12] This lemma can be obtained by the methods of [2] as well. Indeed, the result (for UTB^- instead of UTB and without the oracle T) is mentioned in an open question at the end of page 15.

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

Lemma 4.6. *For every* $\phi \in \Delta_0(\mathcal{L}'_T)$,

$$\mathrm{CT}_0 + \forall\phi\big(\mathrm{Th}(\phi) \to T(\phi)\big) \vdash \forall x\big(\mathrm{Pr}^{T}_{\mathrm{UTB}(\mathcal{L}')+\mathrm{Th}}(\phi[x]) \to \phi(x)\big).$$

In order to bypass the problems with infinitely many additional predicates in $\mathcal{L}'$ it is sufficient to work with an $\mathcal{M}$-bounded fragment of Th and consider only the fragment of $\mathcal{L}'$ consisting of predicates which occur in a formula in the fixed proof p. △

The following lemma suffices to complete the proof of Theorem 4.1.

Lemma 4.7 (Bounding lemma). *For every formula* $\phi(x) \in \Delta_0(\mathcal{L}_T)$ *and every* $n \in \omega$, *the following implication is provable in* CT_0:

$$\mathrm{Pr}^{T}_{\mathrm{UTB}_n}(\exists v\phi(v)) \to \exists y \mathrm{Pr}^{T}_{\mathrm{UTB}}(\exists v < \underline{y}\phi(v)).$$

Before we prove it, let us show a proposition which was the motivation for the proof of the above lemma:

Proposition 4.8 (Σ_1-completeness for UTB^{-}). *For every* $\Delta_0(\mathcal{L}_T)$ *formula* $\phi(x)$, *if* $(\mathbb{N}, \mathrm{Th}(\mathbb{N})) \models \exists x\phi(x)$, *then for some* $n \in \omega$ $\mathrm{Th}(\mathbb{N}) + \mathrm{UTB}^{-} \vdash \exists x < \underline{n}\phi(x)$.

Proof. Suppose that for every n, $\mathrm{UTB}^{-} + \mathrm{Th}(\mathbb{N}) + \forall x < \underline{n}\neg\phi(x)$ is consistent. A trivial compactness argument then shows that $\mathrm{Th}(\mathbb{N}) + \mathrm{UTB}^{-} + \{\forall x < \underline{n}\phi(x) \mid n \in \omega\}$ is consistent as well. So let us take $(\mathcal{M}, T) \models \mathrm{Th}(\mathbb{N}) + \mathrm{UTB}^{-} + \{\forall x < \underline{n}\phi(x) \mid n \in \omega\}$ and look at $(\mathbb{N}, T\restriction_{\mathbb{N}})$. Since $\mathbb{N} \preceq_e \mathcal{M}$, $(\mathbb{N}, T\restriction_{\mathbb{N}}) \models \mathrm{UTB}^{-}$, and, consequently $T\restriction_{\mathbb{N}} = \mathrm{Th}(\mathbb{N})$, because in $\mathbb{N}$ there is only one interpretation for the UTB^{-}-truth predicate. Since $\phi(x) \in \Delta_0(\mathcal{L}_T)$, then $(\mathbb{N}, \mathrm{Th}(\mathbb{N})) \models \forall x\neg\phi(x)$, which suffices to end the proof. □

Remark 4.9. Essentially the same proof shows that already TB^{-} (a non-uniform version of UTB^{-} based solely on Tarski biconditionals for arithmetical sentences) is Σ_1-complete in the above sense. △

Proof of Lemma 4.7. Fix $n \in \omega$, $\phi(x) \in \Delta_0(\mathcal{L}_T)$, and a model $(\mathcal{M}, T) \models \mathrm{CT}_0$. Assume that for all $a \in M$, $(\mathcal{M}, T) \models \neg\mathrm{Pr}^{T}_{\mathrm{UTB}}(\forall v{<}\underline{a}\neg\phi(v))$. By formalizing in PA the trivial compactness argument, we see that for an $(\mathcal{M}, T)$-definable theory

$$\mathrm{Th} := \mathrm{UTB} + \{\phi \mid T(\phi)\} + \{\forall v < \underline{a}\neg\phi(v) \mid a \in M\},$$

$(\mathcal{M}, T) \models \mathrm{Con}_{\mathrm{Th}}$. However, contrary to what happened in the proof of Σ_1-completeness for UTB^{-} (Proposition 4.8), there need not be an $(\mathcal{M}, T)$-definable full model of Th, since we may not have Σ_1-induction for $\mathcal{L}_T$. As a remedy, we shall work with restrictions of T. For an arbitrary c we shall show that $\neg\mathrm{Pr}^{T\restriction_c}_{\mathrm{UTB}_n\restriction_c}(\forall v\neg\phi(v))$, which suffices to prove the claim by a trivial compactness argument ($\mathrm{UTB}_n\restriction_c$ denotes the first $c-1$ axioms of UTB_n). Fix c and let $c < b$ be big enough so that every formula in $T\restriction_c$ and every axiom of $\mathrm{UTB}^{-}\restriction_c$ belongs to Σ^{*}_{b-1}. Working in $(\mathcal{M}, T_b)$ consider

$$\mathrm{Th}_b := \mathrm{UTB} + \{\phi \mid T(\phi) \wedge \phi \in \Sigma^{*}_{b}\} + \{\forall v < \underline{a}\neg\phi(v) \mid a \in M\}.$$

Since $\mathrm{Th}_b \subseteq \mathrm{Th}$, then Th_b is consistent. Th_b is definable in $(\mathcal{M}, T_b) \models \mathrm{PA}^{*}$; hence we can fix a full model $((\mathcal{N}, T'), \mathrm{Sat}_{(\mathcal{N},T')})$ such that

$$(\mathcal{M}, T_b) \models \big[(\mathcal{N}, T') \models_{\mathrm{Sat}_{(\mathcal{N},T')}} \mathrm{Th}_b\big].$$

Consider a model $(\mathcal{M}, T'\restriction_M)$ (i.e., we restrict T' to model M). Since $(\mathcal{M}, T'\restriction_M) \subseteq_e (\mathcal{N}, T')$ and $\phi(x) \in \Delta_0(\mathcal{L}_T)$, then $(\mathcal{M}, T'\restriction_M) \models \forall x \neg\phi(x)$. We check that $(\mathcal{M}, T'\restriction_M) \models \mathrm{CS}^-(\Sigma_b^*)$. To this end it is sufficient to show that for every $\phi \in \Sigma_b^*$ we have

$$\phi \in T'\restriction_M \iff \phi \in T. \qquad (*)$$

So fix an arbitrary such ϕ. Now the following conditions are equivalent:

1. $\phi \in T'\restriction_M$;
2. $(\mathcal{N}, T') \models T(\phi)$;
3. $(M, T_b) \models \big[(\mathcal{N}, T') \models_{\mathrm{Sat}_{(\mathcal{N},T')}} \phi\big]$;
4. $(\mathcal{M}, T) \models T(\phi)$.

Now, the equivalence between (2) and (3) is by the fact that $(\mathcal{N}, T')$ is an $(\mathcal{M}, T_b)$-definable full model of UTB. The equivalence between (3) and (4) holds because $\mathcal{N}$ satisfies all sentences from Σ_b^*, which T deems true in $\mathcal{M}$. Since $T'\restriction_M$ is $(\mathcal{M}, T_b)$-definable it satisfies full $\mathcal{L}_T$ induction. To sum up, we have just concluded that

$$(\mathcal{M}, T'\restriction_M) \models \mathrm{CS}(\Sigma_b^*) + \forall x \neg\phi(x).$$

We work in $(\mathcal{M}, T'\restriction_M)$. First observe that T_b' makes $\mathbf{V}$ (i.e., the class of all numbers; see Example 2.5) a b-full model for the arithmetical vocabulary (because $T_b' = T_b$, by $(*)$). For a fixed $k \in \omega$ let X_k consist of all Boolean combinations of sentences from $\Sigma_b^*(\mathcal{L}) \cup \Sigma_k^*(\mathcal{L}_T)$. Observe that for all $k \in \omega$, all the uniform Tarski biconditionals are in X_k and this class is closed under subformulae. Now for every $k \in \omega$, there exists an $(\mathcal{M}, T'\restriction_M)$ definable satisfaction class for $\mathbf{V}[T]$ and sentences from X_k. Denote such a satisfaction predicate with $\mathrm{Sat}_{\Sigma_k}^{T_b'}$. Then $\mathrm{Sat}_{\Sigma_{n+2}}^{T_b'}$ makes $\mathbf{V}[T]$ an X_n-model for the definable restricted theory

$$\{\phi \in \mathcal{L}_{\mathrm{PA}} \mid T_c(\phi)\} + \mathrm{UTB}_n\restriction_c + \forall x \neg\phi(x).$$

By Proposition 2.16 this is enough to show that this theory is consistent. □

Remark 4.10. The proof of Lemma 4.4 can be modified to yield a new proof of Theorem 1 from [2] (for finite languages). Let us restate it and prove it here:

Theorem 4.11 (Beklemishev–Pakhomov). *Let $\mathcal{L}'$ be an arbitrary finite language with a fixed Gödel numbering and* Th *be a Gödelized $\mathcal{L}'$ theory. Then,*

$$\mathrm{UTB}^-(\mathcal{L}') + \Delta_0(\mathcal{L}_T')\text{-}\mathrm{REF}(\mathrm{UTB}^-(\mathcal{L}') + \mathrm{Th})$$

is $\mathcal{L}'$-conservative over $\mathrm{REF}(\mathrm{Th})$.

Proof. Pick any $(\mathcal{M}, S) \models \mathrm{REF}(\mathrm{Th}) + \mathrm{CS}(\Sigma_c^*(\mathcal{L}'))$ (the satisfaction class is defined for $\mathcal{L}'$). Using overspill, as in the proof of Lemma 5.3 we may pick a nonstandard $d < c$ such that

$$(\mathcal{M}, S) \models \forall \phi \in \Sigma_d^*(\mathcal{L}')\big(\mathrm{Pr}_{\mathrm{Th}}^{S_d}(\phi) \rightarrow S(\phi, \varepsilon)\big).$$

Then, in $(\mathcal{M}, S)$ the $\mathcal{L}_S'$-theory $\mathrm{Th} + \mathrm{UTB}^-(\mathcal{L}') + \{\phi \in \Sigma_d^*(\mathcal{L}') \mid S(\phi, \varepsilon)\}$ is consistent, so let $(\mathcal{N}, S')$ be its full model. We claim that

$$(\mathcal{M}, S'\restriction_M) \models \mathrm{UTB}^-(\mathcal{L}') + \Delta_0(\mathcal{L}_T')\text{-}\mathrm{REF}(\mathrm{UTB}^-(\mathcal{L}') + \mathrm{Th}).$$

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

$\mathrm{UTB}^{-}(\mathcal{L}')$ holds in $(\mathcal{M}, S'\restriction_M)$ since, in $\mathcal{M}$, S and $S'\restriction_M$ coincide on all formulae from Σ_d^*. To show Δ_0 reflection, we use the fact that $(\mathcal{N}, S')$ is strongly interpretable in $(\mathcal{M}, S)$. Fix a $\Delta_0(\mathcal{L}_T')$-formula $\phi(x)$ and working in $(\mathcal{M}, S)$ assume that p is a proof of $\phi(\underline{a})$ (for some element a) from the axioms of $\mathrm{UTB}^{-}(\mathcal{L}') + \mathrm{Th}$. Since

$$(\mathcal{M}, S) \models \Big[(\mathcal{N}, S') \models_{\mathrm{Sat}_{(\mathcal{N},S')}} \mathrm{UTB}^{-} + \mathrm{Th}\Big],$$

then $(\mathcal{N}, S') \models \phi(a)$ and since $(\mathcal{M}, S'\restriction_M) \subseteq_e (\mathcal{N}, S')$, $\phi(a)$ holds in $(\mathcal{M}, S'\restriction_M)$ by absoluteness of bounded formulae. □

△

§5. The arithmetical part of CT_0.

In this section we reprove the following result of Kotlarski [15].[13]

THEOREM 5.1 (Kotlarski and Smoryński). CT_0 *is arithmetically conservative over* $\mathrm{REF}^{\omega}(\mathrm{PA})$.

The idea of Kotlarski's argument is to mimic the Henkin proof of Completeness Theorem in a countable recursively saturated model. Also, [2] gives a different, syntactic proof of Theorem 5.1. We choose a still different path and our main ingredient is the following:

THEOREM 5.2. *Let* Th *be any Gödelized* $\mathcal{L}$ *theory. Suppose that* $\mathcal{M} \models \mathrm{REF}^{\omega}(\mathrm{Th})$ *and* S *is a satisfaction class for* $\mathcal{M}$ *such that*

$$(\mathcal{M}, S) \models \mathrm{CS}(\Sigma_c^*(\mathcal{L}))$$

for some nonstandard c*. Then there exists a nonstandard* $d \in M$ *and* $(\mathcal{N}, S') \models \mathrm{CS}_0 + \forall x\big(\mathrm{Th}(x) \rightarrow S(x, \varepsilon)\big)$ *such that* $\mathcal{M} \subsetneq_e \mathcal{N}$ *and* $S_d \subseteq S'$.

Observe that the conditions $\mathcal{M} \subsetneq_e \mathcal{N}$ and $S_d \subseteq S'$ jointly imply that $\mathcal{M} \not\preceq_e \mathcal{N}$ (assuming S_d and S' are satisfaction classes). The construction of $(\mathcal{N}, S')$ proceeds in ω-stages and is motivated by Corollary 3.21: $\mathcal{N}$ will admit a cofinal ω-sequence $a_0, \dots, a_n, \dots$ and at the n-th stage of the construction we shall build a satisfaction class for all formulae of complexity a_n. The following lemma makes this idea more precise. Henceforth Th is any Gödelized theory in $\mathcal{L}$.[14]

LEMMA 5.3. *Suppose that* $\mathcal{M} \models \mathrm{REF}^{\omega}(\mathrm{Th})$, c *is a nonstandard element of* M, *and* $(\mathcal{M}, S) \models \mathrm{CS}(\Sigma_c^*(\mathcal{L}))$*. Then there exist a nonstandard* $d \in M$ *and a sequence* $\{(\mathcal{M}_n, S_n, c_n)\}_{n\in\omega}$ *such that* $(\mathcal{M}_0, S_0, c_0) = (\mathcal{M}, S_d, d)$ *and for each* n:

1. $\mathcal{M}_n \preceq_e \mathcal{M}_{n+1}$ *and* $S_n \subseteq S_{n+1}$.
2. $(\mathcal{M}_{n+1}, S_{n+1}) \models \mathrm{CS}(\Sigma_{c_{n+1}}^*(\mathcal{L})) + \forall \phi \in \Sigma_{c_{n+1}}^*\big(\mathrm{Th}(\phi) \rightarrow S(\phi, \varepsilon)\big)$.
3. $c_{n+1} \in M_{n+1} \setminus M_n$.

Thus $\{(\mathcal{M}_n, S_n, c_n)\}_{n\in\omega}$ consists of proper end-extensions and each satisfaction class S_{n+1} in the sequence decides all formulae in the sense of M_n. Let us show that Lemma 5.3 immediately implies Theorem 5.2.

[13] In order to get the arithmetical theory right one should compose Kotlarski's result with Smoryński's observation from [23].

[14] All the results generalize to the setting where $\mathcal{L}$ is substituted with an arbitrary finite language. We have decided for the reduced version to keep the definition simpler.

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

Proof of Theorem 5.2 modulo Lemma 5.3. Fix $\mathcal{M} \models \mathrm{REF}^{\omega}(\mathrm{Th})$ and S such that $(\mathcal{M}, S) \models \mathrm{CS}(\Sigma^*_c)$. Fix a chain $\{(\mathcal{M}_n, S_n, c_n)\}_{n\in\omega}$ as in the thesis of Lemma 5.3. Put $\mathcal{N} = \bigcup_n \mathcal{M}_n$, $S' = \bigcup_n S_n$. It is straightforward to verify that $(\mathcal{N}, S') \models \mathrm{CS}^{-} + \forall\phi\big(\mathrm{Th}(\phi) \to S(\phi, \varepsilon)\big)$. Let us check one direction of the quantifier axiom. Assume that $(\mathcal{N}, S') \models \exists\beta \geq_v \alpha S(\phi(v), \beta)$. Let k be big enough so that $(\mathcal{M}_n, S_n) \models \exists\beta \geq_v \alpha S(\phi(v), \beta)$ and $\phi(v) \in \Sigma^*_{c_{n-1}}$. It follows that $(\mathcal{M}_n, S_n) \models S(\exists v\phi(v), \alpha)$. Hence $(\mathcal{N}, S') \models S(\exists v\phi(v), \alpha)$. The argument in the rest of cases is similar.

To check $\Delta_0(\mathcal{L}_T)$-induction, we verify that S is coded, i.e., for every $c \in N$ there exists a $d \in N$ such that

$$\forall x < c\big(S((x)_0, (x)_1) \equiv x \in d\big)$$

holds (recall that $(x)_i$ denotes the projection of x to the i-th coordinate). Take any c and let n be big enough so that $c \in M_n$ and each formula smaller than c is of complexity $\Sigma^*_{c_n}$. By using induction in (M_{n+1}, S_{n+1}) we can find the appropriate $d \in M_n$ and since $(\mathcal{M}_n, S_{n+1}\restriction_{M_n}) \subseteq_e (\mathcal{N}, S')$, this d will work also for $(\mathcal{N}, S')$. □

The proof of Lemma 5.3 consists in *prolonging* the given satisfaction class S until its domain catches up with the universes of models constructed along the way. As it turns out this notion of a satisfaction class being *prolongable* is worth isolating.

Definition 5.4. Suppose $(\mathcal{M}, S, X) \models \mathrm{CS}(X)$ and for every n, $(\mathcal{M}, X) \models \Sigma^*_n \subseteq X$.

1. S is 0-prolongable if there is $\mathcal{N}$ such that $\mathcal{M} \subsetneq_e \mathcal{N}$, $\mathcal{N}$ is strongly interpreted in $(\mathcal{M}, S)$ via a satisfaction class $\mathrm{Sat}_{\mathcal{N}}$, and for all $\phi \in s(X)$ and $\alpha \in \mathrm{Asn}(\phi)$,

$$(\mathcal{M}, S) \models \Big[S(\phi, \alpha) \equiv \mathcal{N} \models_{\mathrm{Sat}_{\mathcal{N}}} \phi[\alpha]\Big].$$

2. S is $n+1$ prolongable if there is $\mathcal{N}$ such that $\mathcal{M} \preceq_e \mathcal{N}$, $c \in N \setminus M$, and $S' \subseteq N^2$ such that $(\mathcal{N}, S') \models \mathrm{CS}(\Sigma^*_c)$, $S \subseteq S'$, and S' is n-prolongable. △

Let us observe that if $\mathcal{N}$ witnesses the 0-prolongability of S, then $\mathcal{N} \models \mathrm{PA}$. The following lemma is the key element of our reasoning:

Convention 5.5. *Let* $\mathrm{GR}(X, S, Z)$ *denote the following sentence of* $\mathcal{L}_2$:

$$\forall\phi \in s(X)\big(\mathrm{Pr}_Z(\phi) \to S(\phi, \varepsilon)\big).$$

Moreover, if S is a satisfaction class on $\mathcal{M}$, then $(\mathcal{M}, S) \models \mathrm{GR}(X, S, S + Y)$ *will abbreviate* $(\mathcal{M}, S) \models \mathrm{GR}(X, S, \{\phi \in \mathrm{Sent}_{\mathcal{L}} \mid S(\phi, \varepsilon)\} \cup Y)$. △

Lemma 5.6. *Suppose that* $(\mathcal{M}, S, c) \models \mathrm{CS}(\Sigma^*_c) \wedge \mathrm{GR}(\Sigma^*_c, S, S + \mathrm{REF}(\mathrm{Th}))$. *Then there exists* $(\mathcal{N}, S', c') \models \mathrm{CS}(\Sigma^*_{c'}) \wedge \mathrm{GR}(\Sigma^*_{c'}, S', S' + \mathrm{Th})$ *such that*:

1. $(\mathcal{N}, S', c')$ *is strongly interpretable in* $(\mathcal{M}, S, c)$.
2. $c' \in N \setminus M$.
3. $S \subseteq S'$.

Proof. Fix $(\mathcal{M}, S, c)$ as in the assumption and work in it. Consider the following $\mathcal{L}_{S'} \cup \{c'\}$ theory, where c' is a fresh constant and S' is a fresh predicate:

$$\mathrm{GR}(\Sigma^*_{c'}, S', S' + \mathrm{Th}) + \{\phi \mid S(\phi, \varepsilon)\} + \mathrm{CS}(\Sigma^*_{c'}, S') + \{c' > \underline{a} \mid a \in M\}.$$

We shall show that the theory is consistent. The proof proceeds in two stages. In the first one, we fix a full model of $\mathrm{REF}(\mathrm{Th}) + \{\phi \mid S(\phi, \varepsilon)\}$, i.e., a full model $(\mathcal{N}', \mathrm{Sat}_{\mathcal{N}'})$ such that $\mathcal{N}' \models_{\mathrm{Sat}_{\mathcal{N}'}} \mathrm{REF}(\mathrm{Th}) + \{\phi \mid S(\phi, \varepsilon)\}$. Such a model exists by the Arithmetized Completeness Theorem, since by our assumption every consequence of $\mathrm{REF}(\mathrm{Th})$ and the set of true sentences (in the sense of S) is true, in the sense of S (recall that $(\mathcal{M}, S) \models \mathrm{PA}^*$). Let us observe that $\mathcal{N}' \models_{\mathrm{Sat}_{\mathcal{N}'}} \mathrm{PA}$, since $\mathrm{REF}(\mathrm{EA}) \vdash \mathrm{PA}$ (and the proof formalizes in EA) and $\mathrm{Th} \supseteq \mathrm{EA}$. Now, using $\mathcal{N}'$ we formalize the standard argument that any model of PA admits an elementary extension to a model with a fully inductive, partial nonstandard satisfaction class. We reason in $(\mathcal{M}, S, c)$. We show that

$$\mathrm{ElDiag}_{\mathrm{Sat}_{\mathcal{N}'}}(\mathcal{N}') + \mathrm{CS}(\Sigma^*_{c'}, S') + \{c > \underline{a} \mid a \in M\} + \mathrm{GR}(\Sigma^*_{c'}, S', S' + \mathrm{Th})$$

is consistent. Let A be a finite (in the sense of $\mathcal{M}$) fragment of this theory and let $a - 1$ be the greatest d such that $\ulcorner c > \underline{d} \urcorner \in A$. We shall find the interpretation for S' in $\mathcal{N}'$. Consider the arithmetical partial truth predicate Sat_a for formulae from Σ^*_a and interpret S' as Sat_a. Then $\mathcal{N}' \models \mathrm{CS}(\Sigma^*_a, \mathrm{Sat}_a)$. Moreover, as $\mathcal{N}' \models \mathrm{REF}(\mathrm{Th})$, $\mathcal{N}' \models \mathrm{GR}(\Sigma^*_a, \mathrm{Sat}_a, \mathrm{Sat}_a + \mathrm{Th})$ (see, e.g., [1]). So A is consistent and by the formalized compactness theorem, so is the entire theory. Let $(\mathcal{N}, S', c')$ be its full model. Then $\mathcal{N}$ is clearly an end-extension of $\mathcal{M}$ and $c' \in N \setminus M$. Moreover $(\mathcal{N}, S', c') \models \mathrm{CS}(\Sigma^*_{c'}, S') \wedge \mathrm{GR}(\Sigma^*_{c'}, S', S' + \mathrm{Th})$ and $S \subseteq S'$, by construction. □

The following lemma isolates the relation between prolongability and global reflection.

Lemma 5.7. *Suppose that $(\mathcal{M}, S, c) \models \mathrm{CS}(\Sigma^*_c)$ where c is nonstandard. Then the following conditions are equivalent*:

1. *S is n-prolongable.*
2. $(\mathcal{M}, S, c) \models \mathrm{GR}(\Sigma^*_c, S, S + \mathrm{REF}^n(\mathrm{EA}))$.

Proof. We prove by induction on n that

$$\forall \mathcal{M} \forall S \forall c \in M \setminus \omega \Big[(\mathcal{M}, S, c) \models \mathrm{CS}(\Sigma^*_c) \implies \big(S \text{ is } n\text{– prolongable} \iff (\mathcal{M}, S, c) \models \mathrm{GR}(\Sigma^*_c, S, S + \mathrm{REF}^n(\mathrm{EA}))\big)\Big].$$

For the base step fix $\mathcal{M}$, S, X as above and assume first that S is 0-prolongable. Fix $\mathcal{N}$ as in the definition of 0-prolongability. Working in $(\mathcal{M}, S)$ take any proof p of a sentence $\phi \in s(\Sigma^*_c)$ from EA. Since EA is a finite theory $\mathcal{N} \models_{\mathrm{Sat}_{\mathcal{N}}} \mathrm{EA}$. Then, since p is a proof from true axioms in the sense of $\mathrm{Sat}_{\mathcal{N}}$, then $\mathrm{Sat}_{\mathcal{N}}(\phi, \varepsilon)$. Hence, since ϕ is a formula from $s(\Sigma^*_c)$, $S(\phi, \varepsilon)$ holds by the definition of 0-prolongability.

Suppose now that $(\mathcal{M}, S, c) \models \forall \phi \in s(\Sigma^*_c)\big(\mathrm{Pr}^S_{\mathrm{EA}}(\phi) \rightarrow S(\phi, \varepsilon)\big)$. Consider the following $(\mathcal{M}, S, c)$-definable theory $\mathrm{Th} := \{\phi \in \Sigma^*_c \mid S(\phi, \varepsilon)\}$. By our assumption, $(\mathcal{M}, S, c) \models \mathrm{Con}_{\mathrm{Th}}$. Work in $(\mathcal{M}, S, c)$. By the Arithmetized Completeness Theorem (we use the assumption that $(\mathcal{M}, S, c) \models \mathrm{PA}^*$), we have a full model $\mathcal{N} \models_{\mathrm{Sat}_{\mathcal{N}}} \mathrm{Th}$. Hence, obviously $\mathrm{Sat}_{\mathcal{N}}$ and S coincide on $s(\Sigma^*_c)$ (observe that they need not coincide on Σ^*_c).

Now assume that the equivalence holds for n. Fix $\mathcal{M}$, S, Σ^*_c as above and assume first that S is $(n+1)$-prolongable and pick $(\mathcal{N}, S', c') \models \mathrm{CS}(\Sigma^*_{c'})$ such that

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

$S \subseteq S'$ and S' is n-prolongable. By the induction assumption $(\mathcal{N}, S', c') \models \forall \phi \in s(\Sigma^*_{c'})\big(\mathrm{Pr}^S_{\mathrm{REF}^n(\mathrm{EA})}(\phi) \rightarrow S(\phi, \varepsilon)\big)$. By compositional clauses, it follows that for every $\phi(v) \in M$

$$(\mathcal{N}, S', c') \models S(\ulcorner \forall v \big(\mathrm{Pr}_{\mathrm{REF}^n(\mathrm{EA})}(\phi[v]) \rightarrow \phi(v)\big)\urcorner, \varepsilon).$$

Hence in particular, if $\psi \in M$ is any axiom of $\mathrm{EA} + \mathrm{REF}^{n+1}(\mathrm{EA})$, then $(\mathcal{N}, S') \models S(\psi, \varepsilon)$. So suppose $p \in M$ is a proof of a sentence $\phi \in s(\Sigma^*_c)$ from the axioms of $\mathrm{REF}^{n+1}(\mathrm{EA})$. Work in $(\mathcal{N}, S')$. By the previous argument, if θ is a premise of p, then $S(\theta, \varepsilon)$ holds. Since all the formulae occurring in p belong to $\Sigma^*_{c'}$, then we have $(\mathcal{N}, S') \models S(\phi, \varepsilon)$. However, S' and S coincide on $s(\Sigma^*_c)$.

Now, working in $(\mathcal{M}, S, c)$, assume $\forall \phi \in s(\Sigma^*_c)\big(\mathrm{Pr}^S_{\mathrm{REF}^{n+1}(\mathrm{EA})}(\phi) \rightarrow S(\phi, \varepsilon)\big)$. We apply Lemma 5.6 to $\mathrm{Th} := \mathrm{REF}^n(\mathrm{EA})$ and conclude that there is $(\mathcal{N}, S', c')$ as in the thesis of the lemma. By our inductive assumption applied to $(\mathcal{N}, S', c')$, it follows that S' is n-prolongable. Hence S is $n+1$-prolongable, which ends the proof. □

Corollary 5.8. *The following conditions are equivalent for a model* $(\mathcal{M}, S, c) \models \mathrm{CS}(\Sigma^*_c, S)$:

1. $(\mathcal{M}, S, c)$ *is n-prolongable.*
2. *There exists an* $(\mathcal{M}, S, c)$*-definable sequence of models* $\{(\mathcal{M}_k, S_k, c_k)\}_{k \leq n \in \omega}$ *such that* $(\mathcal{M}_0, S_0, c_0) = (\mathcal{M}, S, c)$ *and for each* $k < n$:
   (a) $\mathcal{M}_k \precneqq_e \mathcal{M}_{k+1}$ *and* $S_k \subseteq S_{k+1}$.
   (b) $(\mathcal{M}_k, S_k, c_k) \models \mathrm{CS}(\Sigma^*_{c_k})$.
   (c) $c_{k+1} \in M_{k+1} \setminus M_k$.
   (d) $(\mathcal{M}_{k+1}, S_{k+1}, c_{k+1})$ *is strongly interpretable in* $(\mathcal{M}_k, S_k, c_k)$.

Corollary 5.9. *The following conditions are equivalent for a model* $(\mathcal{M}, S, c) \models \mathrm{CS}(\Sigma^*_c, S)$ *and for every* $n \in \omega$:

1. $\mathcal{M} \models \mathrm{REF}^{n+1}(\mathrm{EA})$.
2. *There exists* $b \in M \setminus \omega$ *such that* S_b *is n-prolongable.*

Proof. (2) ⇒ (1) follows immediately from Lemma 5.7. We prove (1) ⇒ (2). Fix $(\mathcal{M}, S, c)$ as in the assumptions. Let Sat_k be an arithmetically definable truth predicate for Σ^*_k (as in the proof of Lemma 5.6). By induction in $(\mathcal{M}, S, c)$, we have

$$\forall \phi \in \Sigma^*_k \big(\mathrm{Sat}_k(\phi, \varepsilon) \equiv S(\phi, \varepsilon)\big)$$

for every k. Since $\mathcal{M} \models \mathrm{REF}^{n+1}(\mathrm{EA})$, then, for every k, we have

$$\mathcal{M} \models \forall \phi \in \Sigma^*_k \big(\mathrm{Pr}^{\mathrm{Sat}_k}_{\mathrm{REF}^n(\mathrm{EA})}(\phi) \rightarrow \mathrm{Sat}_k(\phi, \varepsilon)\big).$$

Hence for every $l \in \omega$, $(\mathcal{M}, S, c) \models \forall x < l \forall \phi \in \Sigma^*_x \big(\mathrm{Pr}^{S_x}_{\mathrm{REF}^n(\mathrm{EA})}(\phi) \rightarrow S_x(\phi, \varepsilon)\big)$. By the overspill principle we can find a $b > \omega$, $b \leq c$, such that

$$(\mathcal{M}, S, c) \models \forall x < b \forall \phi \in \Sigma^*_x \big(\mathrm{Pr}^{S_x}_{\mathrm{REF}^n(\mathrm{EA})}(\phi) \rightarrow S_x(\phi, \varepsilon)\big).$$

Hence S_{b-1} is n-prolongable. □

Since $\mathrm{REF}(\mathrm{EA}) = \mathrm{PA}$, the following is the most memorizable version of our main lemma:

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

Theorem 5.10. *The following conditions are equivalent for a model* $(\mathcal{M}, S, c) \models \mathrm{CS}(\Sigma^*_c)$ *and for every* $n \in \omega$:

1. $\mathcal{M} \models \mathrm{REF}^n(\mathrm{PA})$.
2. *There exists* $b \in M \setminus \omega$ *such that* S_b *is n-prolongable.*

Now, we have all that is needed to finish our conservativity proof for CT_0:

Proof of Lemma 5.3. Suppose $(\mathcal{M}, S, c) \models \mathrm{CS}(\Sigma^*_c)$ and $\mathcal{M} \models \mathrm{REF}^{\omega}(\mathrm{Th})$. Let Sat_n be the arithmetical partial truth predicate for formulae in Σ^*_n. By assumption on $\mathcal{M}$ it follows that for every n

$$\mathcal{M} \models \forall \phi \in \Sigma^*_n \big(\mathrm{Pr}^{\mathrm{Sat}_n}_{\mathrm{REF}^n(\mathrm{Th})}(\phi) \to \mathrm{Sat}_n(\phi, \varepsilon)\big).$$

By induction, for every $n \in \omega$, Sat_n and S_n coincide. Hence for every $n \in \omega$ we have

$$(\mathcal{M}, S) \models \forall x < n \; \mathrm{GR}(\Sigma^*_x, S_x, S_x + \mathrm{REF}^x(\mathrm{Th})).$$

By overspill it follows that for some $d > \omega$

$$(\mathcal{M}, S) \models \mathrm{GR}(\Sigma^*_d, S_d, S_d + \mathrm{REF}^d(\mathrm{Th})).$$

We define the chain $(\mathcal{M}_n, S_n, c_n)$ by induction. Assume that $(\mathcal{M}_k, S_k, c_k)$ has been defined and it satisfies $\mathrm{GR}(\Sigma^*_{c_k}, S_{c_k}, S_{c_k} + \mathrm{REF}^{d-k}(\mathrm{Th}))$. To get $(\mathcal{M}_{k+1}, S_{k+1}, c_{k+1})$ we apply Lemma 5.6 to $\mathrm{Th}' = \mathrm{REF}^{d-(k+1)}(\mathrm{Th})$. □

Corollary 5.11. *The arithmetical consequences of* $\mathrm{CT}_0 + \forall \phi \big(\mathrm{Th}(\phi) \to S(\phi, \varepsilon)\big)$ *and* $\mathrm{REF}^{\omega}(\mathrm{Th})$ *coincide.*

Proof. Conservativity part follows from Theorem 5.2. That $\mathrm{CT}_0 + \forall \phi \big(\mathrm{Th}(\phi) \to S(\phi, \varepsilon)\big) \vdash \mathrm{REF}^{\omega}(\mathrm{Th})$ was established in Corollary 3.11. □

It might seem that in the above proof the limit model is a very specific model of CS_0. Quite surprisingly every model of CS_0 is of this form. One of the crucial steps in the proof is worth isolating as a lemma.

Lemma 5.12. *Suppose that* $(\mathcal{M}, S) \models \mathrm{CS}_0$, $(\mathcal{M}_0, S{\restriction}_{M_0}) \subseteq (\mathcal{M}, S)$, $d_1 \in M_0$, *and* $(\mathcal{M}_0, S_{d_1}{\restriction}_{M_0}) \models \mathrm{CS}^-(\Sigma^*_{d_1})$. *Then for every* $d_0 \in M_0$ *such that* $d_1 - d_0$ *is nonstandard,* $(\mathcal{M}_0, S_{d_0}{\restriction}_{M_0}) \models \mathrm{CS}(\Sigma^*_{d_0})$.

Proof. Let us fix $\mathcal{M}, S, \mathcal{M}_0, d_1, d_1$ as above. Since $(\mathcal{M}_0, S_{d_1}{\restriction}_{M_0}) \models \mathrm{CS}^-(\Sigma^*_{d_1})$, it follows that

$$(\mathcal{M}_0, S_{d_0}{\restriction}_{M_0}) \models \mathrm{CS}^-(\Sigma^*_{d_0}).$$

We prove that induction axioms hold in $(\mathcal{M}_0, S_{d_0}{\restriction}_{M_0})$ as well. By the assumptions, it follows that $(\mathcal{M}, S_{d_1}) \models \forall \phi \in \Sigma^*_{d_1} \big(\mathrm{Pr}_{\mathrm{PA}}(\phi) \to S(\phi, \varepsilon)\big)$, hence also

$$(\mathcal{M}_0, S_{d_1}{\restriction}_{M_0}) \models \forall \phi \in \Sigma^*_{d_1} \big(\mathrm{Pr}_{\mathrm{PA}}(\phi) \to S(\phi, \varepsilon)\big).$$

Let $\mathrm{Sat}_{\Sigma_{d_0}}$ be as in the proof of Theorem 3.19 and let us abbreviate $S(\phi, \varepsilon)$ with $T(\phi)$ and $\mathrm{Sat}_{\Sigma_{d_0}}(x, \varepsilon)$ with $Tr_{d_0}(x)$. Observe that $Tr_{d_0} \in \Sigma^*_{d_0}$. It follows that

$$(\mathcal{M}_0, S_{d_1}{\restriction}_{M_0}) \models \forall \phi \in \Sigma^*_{d_0} \big(T(Tr_{d_0}(\underline{\phi^{\Sigma}})) \equiv T(\phi^{\Sigma})\big).$$

Since S_{d_0} coincides with S_{d_1} on $(\Sigma^*_{d_0})^{\mathcal{M}_0}$, then

$$(\mathcal{M}_0, S_{d_0}\restriction_{M_0}) \models \forall \phi \in \Sigma^*_{d_0}\big(T(Tr_{d_0}(\underline{\phi^\Sigma})) \equiv T(\phi^\Sigma)\big).$$

Since $T(Tr_{d_0}(\underline{x^\Sigma}))$ defines and $S_{d_0}\restriction_{M_0}$ in $(\mathcal{M}_0, S_{d_1}\restriction_{M_0})$ it is sufficient to show that

$$(\mathcal{M}_0, S_{d_1}\restriction_{M_0}) \models \eta[T(Tr_{d_0}(\underline{x^\Sigma}))/P], \qquad (*)$$

where η is an arbitrary instance of an induction axiom for a fresh predicate letter P, in a semi-relational form. Fix η. Since $\mathcal{M}_0 \models \mathrm{Pr}_{\mathrm{PA}}(\eta[Tr_{d_0}(\underline{x^\Sigma})/P])$ and $\eta[Tr_{d_0}(\underline{x^\Sigma})/P] \in \Sigma^*_{d_1}$, we have

$$(\mathcal{M}_0, S_{d_1}\restriction_{M_0}) \models T\left(\eta[Tr_{d_0}(\underline{x^\Sigma})/P]\right).$$

$(*)$ follows as in [19] and Theorem 3.19. □

THEOREM 5.13. *Suppose* $(\mathcal{M}, S) \models \mathrm{CS}_0$ *has cofinality* κ. *Then there is a chain* $\{(\mathcal{M}_\alpha, S_\alpha, c_\alpha)\}_{\alpha\in\kappa}$ *such that* $\bigcup_{\alpha\in\kappa}\mathcal{M}_\alpha = \mathcal{M}$, $\bigcup_{\alpha\in\kappa} S_\alpha$ *and for every* $\alpha < \beta < \kappa$:

1. $\mathcal{M}_\alpha \preceq_e \mathcal{M}_\beta$ *and* $S_\alpha \subseteq S_\beta$.
2. $(\mathcal{M}_\beta, S_\beta) \models \mathrm{CS}(\Sigma^*_{c_\beta})$.
3. $c_\beta \in M_\beta \setminus M_\alpha$.

We note that the above chain need not be continuous.

PROOF. Fix $(\mathcal{M}, S) \models \mathrm{CS}_0$ and a cofinal sequence $\{d_\alpha\}_{\alpha\in\kappa}$. In the base step we build $(\mathcal{M}_0, S_0, d_0) \models \mathrm{CS}(\Sigma^*_{d_0})$. Consider the formula $\theta'(x, y, z)$

$$\exists c\big(\mathrm{Seq}(c) \wedge \mathrm{len}(c) = x \wedge c_0 = z \wedge \forall i < x(2^{c_i} < c_{i+1}) \wedge \theta(x, y, c)\big),$$

where $\theta(x, y, c)$ is

$$\forall \phi \in \Sigma^*_y \forall i < x \forall \alpha < c_i \Big(\phi < c_i \wedge \alpha \in \mathrm{Asn}(\exists v\phi) \wedge S(\exists v\phi, \alpha) \\ \rightarrow \exists \beta < c_{i+1}\big(\beta \sim_v \alpha \wedge S(\phi, \beta\restriction_\phi)\big)\Big).$$

Thus $\theta'(x, y, z)$ expresses that there is a witness-bounding sequence c of length x and starting from z, which works for those formulae of Σ^*_y complexity which are below some of the elements of the sequence. Let e be such that $e - d_0$ is nonstandard. We reason in $(\mathcal{M}, S_e) \models \mathrm{CS}(\Sigma^*_e)$ and by a straightforward induction conclude that

$$\forall x \theta'(x, e, e).$$

We let a be an arbitrary nonstandard number and we fix c witnessing $\theta'(a, e, e)$. We define

$$M_0 := \sup\{c_i \mid i \in \omega\}.$$

Clearly if $b_1, b_2 < c_i$, for some i, then $b_1 + b_2, b_1 \cdot b_2 < c_{i+1}$ by the assumption on c. Hence $\mathcal{M}_0 \subseteq \mathcal{M}$. We check that $(\mathcal{M}_0, S_e\restriction_{M_0}) \models \mathrm{CS}^-(\Sigma^*_e)$. The unique non-trivial step is the one for quantifiers: fix any formula $\phi(v, \bar{w}) \in \Sigma^*_e \cap M_0$, $\alpha \in M_0$ and assume that $(\mathcal{M}_0, S_e\restriction_{M_0}) \models S(\exists v\phi(v, \bar{w}), \alpha)$ and $\alpha, \phi < c_i$. Then

$$(\mathcal{M}, S) \models S(\exists v\phi(v, \bar{w}), \alpha)$$

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

Hence by the properties of c there is $\beta < c_{i+1}$, $\beta \sim_v \alpha$ such that

$$(\mathcal{M}_0, S_e) \models S(\phi, \beta).$$

To conclude that $S_0 := S_{d_0}$ is fully inductive, we apply Corollary 5.12 for $\mathcal{M}' = \mathcal{M}_0$, $d_0 = d_0$, and $d_1 = e$.

In the successor step assume that $(\mathcal{M}_\alpha, S_\alpha, c_\alpha)$ has been constructed and w.l.o.g. assume that $d_{\alpha+1} \notin M_\alpha$. We pick $e \in M$ such that $e - d_{\alpha+1}$ is nonstandard and repeat the reasoning from the base step with $d_{\alpha+1}$ replacing d_0. In the limit step, we assume that for every $\alpha < \beta$, $(\mathcal{M}_\alpha, S_\alpha, c_\alpha)$ has been constructed. We take $(\mathcal{M}'_\beta, S'_\beta) = \bigcup_{\alpha<\beta}(\mathcal{M}_\alpha, S_\alpha)$ and assume (by the regularity of κ) that γ is the least such that $d_\gamma \notin M'_\beta$. We put $c_\beta = d_\gamma$ and repeat the reasoning from the base step with c_β replacing d_0. □

The construction from Lemma 5.3 enables us to extend the result from Section 4.

THEOREM 5.14. *$\mathrm{CT}_0 + \Sigma_1(\mathcal{L}_T)$-REF$(\mathrm{UTB} + T)$ is $\Pi_1(\mathcal{L}_T)$ conservative over* CT_0.

PROOF. Fix $(\mathcal{M}, T) \models \mathrm{CT}_0$. We shall find $(\mathcal{M}, T) \subseteq (\mathcal{N}, T') \models \mathrm{CT}_0 + \Sigma_1(\mathcal{L}_T)$-REF$(\mathrm{UTB} + T)$, which suffices to end the proof. Firstly, turn $(\mathcal{M}, T)$ into a model $(\mathcal{M}, S) \models \mathrm{CS}_0$ in the canonical way. Secondly, assume that there exists $\{(\mathcal{M}_i, S_i, c_i)\}_{i\in\omega}$ such that $(\mathcal{M}_0, S_0) = (\mathcal{M}, S)$, the rest of the chain is as in the thesis of Lemma 5.3 and for each $i > 0$, $(\mathcal{M}_{i+1}, S_{i+1}, c_{i+1})$ is strongly interpreted in its predecessor $(\mathcal{M}_i, S_i, c_i)$. Let Sat_{i+1} denote the satisfaction relation witnessing the interpretability of $(\mathcal{M}_{i+1}, S_{i+1}, c_{i+1})$ in $(\mathcal{M}_i, S_i, c_i)$. Put $(\mathcal{M}_\infty, S_\infty) = \bigcup_{i\in\omega}(\mathcal{M}_i, S_i)$ and define $T_\infty := \{\phi \in M_\infty \mid S(\phi, \varepsilon)\}$. By the proof of Theorem 5.2, $(\mathcal{M}_\infty, S_\infty) \models \mathrm{CS}_0$, and hence $(\mathcal{M}_\infty, T_\infty) \models \mathrm{CT}_0$, so it is sufficient to show that

$$(\mathcal{M}_\infty, T_\infty) \models \Sigma_1(\mathcal{L}_T)\text{-REF}(\mathrm{UTB} + T).$$

Suppose that for some $\phi(x) \in \Sigma_1(\mathcal{L}_T)$ and $c \in M$, $(\mathcal{M}_\infty, T_\infty) \models \mathrm{Pr}^T_{\mathrm{UTB}}(\phi(\underline{c}))$. Let p be the witnessing proof and fix any $i \in \omega$ such that $p < c_i$. Hence $(\mathcal{M}_i, S_i) \models \mathrm{Pr}^S_{\mathrm{UTB}}(\phi(\underline{c}))$. We reason in $(\mathcal{M}_i, S_i)$. Since every next model in the chain is strongly interpretable in the previous one, we have (in $(\mathcal{M}_i, S_i)$)

$$(\mathcal{M}_{i+1}, S_{i+1}) \models_{\mathrm{Sat}_{i+1}} \forall v(\mathcal{M}_{i+2}, S_{i+2}, c_{i+2}) \models_{\mathrm{Sat}_{i+2}} \mathrm{CS}(\Sigma^*_{c_{i+2}}) \wedge c_{i+2} > \underline{v}.$$

Consider (in $(\mathcal{M}_i, S_i)$) the model $(\mathcal{M}_{i+1}, S_{i+2}\restriction_{M_{i+1}})$. This model is interpretable in $(\mathcal{M}_{i+1}, S_{i+1})$; hence by Proposition 2.20 it is a full model (in the sense of $(\mathcal{M}_i, S_i)$; i.e., it is strongly interpretable in $(\mathcal{M}_i, S_i)$). Let Sat'_{i+1} denote the respective satisfaction class. We shall show that

$$(\mathcal{M}_{i+1}, S_{i+2}\restriction_{M_{i+1}}) \models \phi(\underline{c}).$$

This suffices to end the proof, since $\phi(c)$ is a $\Sigma_1(\mathcal{L}_T)$ formula and $(\mathcal{M}_{i+1}, S_{i+2}\restriction_{M_{i+1}}) \subseteq_e (\mathcal{M}_\infty, T_\infty)$. Reasoning in $(\mathcal{M}_i, S_i)$ we see that since $S_{i+2}\restriction_{M_{i+1}}$ is definable in (M_{i+1}, S_{i+1}), which satisfies full induction, also $(\mathcal{M}_{i+1}, S_{i+2}\restriction_{M_{i+1}})$ satisfies full induction. Moreover, since $S_{i+1} \subseteq S_{i+2}\restriction_{M_{i+1}}$ we know that for every e $(\in M_i)$ $(\mathcal{M}_{i+1}, S_{i+2}\restriction_{M_{i+1}}) \models_{\mathrm{Sat}'_{i+1}} \mathrm{CS}(\Sigma^*_e)$. Hence,

$$(\mathcal{M}_{i+1}, S_{i+2}\restriction_{M_{i+1}}) \models_{\mathrm{Sat}'_{i+1}} \mathrm{UTB}.$$

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press

Moreover, if ψ is an assumption of proof p and an arithmetical sentence, then $\psi \in T_\omega$; hence, by the choice of i, $\psi \in S_i$. It follows that $(\mathcal{M}_{i+1}, S_{i+2}\restriction_{M_{i+1}}) \models_{\mathrm{Sat}'_{i+1}} \psi$, since Sat'_{i+1} coincides with Sat_{i+1} on sentences from M_i. We conclude, still working in $(\mathcal{M}_i, S_i)$, that Sat'_{i+1} makes every premise of p true in $(\mathcal{M}_{i+1}, S_{i+2}\restriction_{M_{i+1}})$. So p's conclusion, $\phi(c)$, must be deemed true in $(\mathcal{M}_{i+1}, S_{i+2}\restriction_{M_{i+1}})$ by Sat'_{i+1}. Since $\phi(c)$ is a standard formula with a parameter, we can conclude that $(\mathcal{M}_{i+1}, S_{i+2}\restriction_{M_{i+1}}) \models \phi(\underline{c})$.

Now, we show how to justify the existence of the chain $\{(\mathcal{M}_i, S_i, c_i)\}_{i\in\omega}$. Let $\mathrm{B}\Sigma_1(\mathcal{L}_T)$ denote the extension of CT_0 with Σ_1 collection scheme for the language with the truth predicate. As shown in [20], $\mathrm{B}\Sigma_1(\mathcal{L}_T)$ is Π_2 conservative over CT_0. So we can assume that the above model $(\mathcal{M}, T)$ is a countable recursively saturated (in the extended language) model of $\mathrm{CT}^- + \mathrm{B}\Sigma_1(\mathcal{L}_T)$. By the classical result of Wilkie–Paris [26] there exists a proper end-extension $(\mathcal{M}', T') \models \mathrm{CT}_0$ of $(\mathcal{M}, T)$.[15] Hence it is sufficient to start the construction of the chain from $(\mathcal{M}', T'_a)$, where $a \in M' \setminus M$ and then proceed as in the proof of Lemma 5.3. □

§6. Two open problems.

We conclude our paper with two open problems:

Question 1. Does the statement of Theorem 5.13 remain true if we require for every $\alpha < \beta$, $\mathcal{M}_\beta$ is strongly interpretable in $\mathcal{M}_\alpha$?

Question 2. Can we strengthen Theorem 5.14 by showing that $\Sigma_1(\mathcal{L}_T)$-$\mathrm{REF}(\mathrm{UTB} + T)$ is in fact provable in $\mathrm{CT}_0 + \mathrm{EA}$?

Let us stress that, by the proof of Theorem 5.14, the positive answer to Question 1 implies that the answer to Question 2 is positive as well.

Acknowledgements. The paper summarizes the developments obtained during the years 2015–2019, some of them being included in the author's Ph.D. thesis [18]. Special thanks go to thesis' supervisor Cezary Cieśliński, who also brought our attention to the problem of the scope of arithmetical consequences of CT_0. We thank Ali Enayat and Albert Visser for many important insights which turned out to be crucial in solving problems presented in this paper. Last but not least, we are extremely grateful to Bartosz Wcisło whose constant encouragement and support during all these years were essential for bringing this work to the daylight.

The current version of the paper benefited a lot from remarks and suggestions by Ali Enayat.

We are grateful to Lev Beklemishev and Fedor Pakhomov for bringing our attention to the open problem on the status of Σ_1-uniform reflection over $\mathrm{UTB}^- + \mathrm{EA}$.

This research was supported by an NCN Maestro grant 2019/34/A/HS1/00399, "Epistemic and semantic commitments of foundational theories."

REFERENCES

[1] L. D. Beklemishev, *Reflection principles and provability algebras in formal arithmetic*. ***Russian Mathematical Surveys***, vol. 60 (2005), no. 2, pp. 197–268.

[15] Originally Wilkie–Paris theorem is formulated for $\mathrm{I}\Delta_0 + \exp$; however their proof easily relativizes and works for models of $\mathrm{I}\Delta_0(X) + \exp$, where X is a fresh predicate.

[2] L. D. Beklemishev and F. N. Pakhomov, *Reflection algebras and conservation results for theories of iterated truth*. ***Annals of Pure and Applied Logic***, vol. 173 (2022), no. 5, p. 103093.

[3] C. Cieśliński, *Deflationary truth and pathologies*. ***The Journal of Philosophical Logic***, vol. 39 (2010), no. 3, pp. 325–337.

[4] ———, *Truth, conservativeness and provability*. ***Mind***, vol. 119 (2010), pp. 409–422.

[5] ———, ***The Epistemic Lightness of Truth: Deflationism and Its Logic***, Cambridge University Press, Cambridge, 2017.

[6] A. Enayat and F. Pakhomov, *Truth, disjunction, and induction*. ***Archive for Mathematical Logic***, vol. 58 (2019), nos. 5–6, pp. 753–766.

[7] A. Enayat and A. Visser, *New constructions of satisfaction classes*, ***Unifying the Philosophy of Truth*** (T. Achourioti, H. Galinon, J. M. Fernández, and K. Fujimoto, editors), Springer, Berlin, 2015, pp. 321–335.

[8] B. Grabmayr, *On the invariance of Gödel's second theorem with regard to numberings*. ***The Review of Symbolic Logic***, vol. 14 (2021), pp. 51–84.

[9] P. Hájek and P. Pudlák, ***Metamathematics of First-Order Arithmetic***, Springer, Berlin, 1993.

[10] V. Halbach, ***Axiomatic Theories of Truth***, Cambridge University Press, Cambridge, 2011.

[11] V. Halbach and A. Visser, *Self-reference in arithmetic I*. ***The Review of Symbolic Logic***, vol. 7 (2014), no. 4, pp. 671–691.

[12] R. Kaye, ***Models of Peano Arithmetic***, Clarendon Press, Oxford.

[13] R. Kaye and H. Kotlarski, *On models constructed by means of the arithmetized completeness theorem*. ***Mathematical Logic Quarterly***, vol. 46 (2000), no. 4, pp. 505–516.

[14] R. Kossak and J. Schmerl, ***The Structure of Models of Peano Arithmetic***, Oxford University Press, Oxford, 2007.

[15] H. Kotlarski, *Bounded induction and satisfaction classes*. ***Zeitschrift für mathematische Logik und Grundlagen der Mathematik***, vol. 32 (1986), pp. 531–544.

[16] H. Kotlarski, S. Krajewski, and A. Lachlan, *Construction of satisfaction classes for nonstandard models*. ***Canadian Mathematical Bulletin***, vol. 24 (1981), pp. 283–293.

[17] G. Leigh, *Conservativity for theories of compositional truth via cut elimination*, this Journal, vol. 80 (2015), no. 3, pp. 845–865.

[18] M. Łełyk, *Axiomatic truth theories, bounded induction and reflection principles*, 2017. Available at http://www.depotuw.ceon.pl/handle/item/2266.

[19] M. Łełyk and B. Wcisło, *Models of positive truth*. ***The Review of Symbolic Logic***, vol. 12 (2019), no. 1, pp. 144–172.

[20] ———, *Local collection and end-extensions of models of compositional truth*. ***Annals of Pure and Applied Logic***, vol. 172 (2021), no. 6, p. 102941.

[21] U. R. Schmerl, *A fine structure generated by reflection formulas over primitive recursive arithmetic*, ***Logic Colloquium '78***, Studies in Logic and the Foundations of Mathematics, vol. 97, Elsevier, Amsterdam, 1979, pp. 335–350.

[22] S. T. Smith, *Nonstandard definability*. ***Annals of Pure and Applied Logic***, vol. 42 (1989), no. 1, pp. 21–43.

[23] C. Smoryński, *ω-consistency and reflection*, ***Colloque International de Logique (Clermont–Ferrand, 1975)***, Colloque International CNRS, vol. 249, CNRS, Paris, 1977, pp. 167–181.

[24] G. Takeuti, ***Proof Theory***, North-Holland, Amsterdam; Elsevier, New York, 1975.

[25] B. Wcisło and M. Łełyk, *Notes on bounded induction for the compositional truth predicate*. ***The Review of Symbolic Logic***, vol. 10 (2017), no. 3, pp. 1–26.

[26] A. Wilkie and J. Paris, *On the existence of end extensions of models of bounded induction*, ***Logic, Methodology and Philosophy of Science VIII*** (J. E. Fenstad, I. T. Frolov, and R. Hilpinen, editors), Studies in Logic and the Foundations of Mathematics, vol. 126, Elsevier, Amsterdam, 1989, pp. 143–161.

FACULTY OF PHILOSOPHY UNIVERSITY OF WARSAW
WARSAW, POLAND
E-mail: mlelyk@uw.edu.pl

https://doi.org/10.1017/jsl.2022.39 Published online by Cambridge University Press