The Journal of Symbolic Logic
Volume 85, Number 1, March 2020

TRUTH AND FEASIBLE REDUCIBILITY

ALI ENAYAT, MATEUSZ ŁEŁYK, AND BARTOSZ WCISŁO

Abstract. Let $\mathcal{T}$ be any of the three canonical truth theories CT^- (compositional truth without extra induction), FS^- (Friedman–Sheard truth without extra induction), or KF^- (Kripke–Feferman truth without extra induction), where the base theory of $\mathcal{T}$ is PA (Peano arithmetic). We establish the following theorem, which implies that $\mathcal{T}$ has no more than polynomial speed-up over PA.

Theorem. *$\mathcal{T}$ is feasibly reducible to* PA, *in the sense that there is a polynomial time computable function f such that for every $\mathcal{T}$-proof π of an arithmetical sentence ϕ, $f(\pi)$ is a* PA*-proof of ϕ.*

§1. Introduction . One of the celebrated results in the area of axiomatic theories of truth is the Krajewski–Kotlarski–Lachlan (KKL) theorem [14] that asserts that every countable recursively saturated model of PA (Peano arithmetic) is expandable to a model of $\mathsf{CT}^-[\mathsf{PA}]$ (compositional truth over PA with no extra induction[1]). The KKL theorem is an overtly model-theoretic result, but it is well known that it is equivalent to the conservativity of $\mathsf{CT}^-[\mathsf{PA}]$ over PA.[2] Recent proofs of the KKL theorem given by Enayat and Visser [6] (using model-theoretic techniques) and Leigh [16] (using proof-theoretic machinery) show that $\mathsf{CT}^-[\mathsf{B}]$ is conservative over B for every "base theory" B (i.e., a theory B that supports a modicum of coding machinery for handling elementary syntax). Moreover, Leigh's proof makes it clear that $\mathsf{CT}^-[\mathsf{B}]$ is *proof-theoretically reducible* to B for every recursively axiomatized base theory B, and, in particular, there is a primitive recursive function f such that for any proof π of a sentence ϕ in $\mathsf{CT}^-[\mathsf{B}]$, where ϕ is a sentence in the language of B, $f(\pi)$ is a proof of ϕ in B. Indeed, Leigh's "reducing function" f is readily seen to be a provably total function of the fragment of PRA (Primitive Recursive Arithmetic) commonly known as $\mathsf{I}\Delta_0 + \mathsf{Superexp}$.[3]

The main result of this article shows that $\mathsf{CT}^-[\mathsf{PA}]$ is *feasibly reducible to* PA, i.e., there is a polynomial time computable function f such that for any proof

Received January 31, 2019.

2010 *Mathematics Subject Classification.* Primary 03F30, 03C62, Secondary 03D15, 03H15.

Key words and phrases. axiomatic theories of truth, compositional truth, conservativity, feasible computations, polynomial simulation, reducibility, speed-up, truth predicates.

[1]This theory is referred to as $\mathsf{PA}^{\mathsf{FT}}$ in [6], CT^- in [3], $\mathsf{CT}\restriction$ in [11], and $\mathsf{CT}[\mathsf{PA}]$ in [16].

[2]This equivalence follows from two key facts: (1) every countable consistent theory has a countable recursively saturated model, and (2) countable recursively saturated models are resplendent, both of which can be verified in the subsystem ACA_0 of second-order arithmetic.

[3]$\mathsf{Superexp}$ asserts the totality of the superexponential function $\mathsf{Superexp}(n, x)$, with $\mathsf{Superexp}(0, x) = x$ and $\mathsf{Superexp}(n + 1, x) = 2^{\mathsf{Superexp}(n,x)}$. Leigh [16] refers to this function as hyper-exponentiation.

0022-4812/20/8501-0017
DOI:10.1017/jsl.2019.24

π of an arithmetical sentence ϕ in $\mathsf{CT}^-[\mathsf{PA}]$, $f(\pi)$ is a proof of ϕ in PA. The feasible reducibility of $\mathsf{CT}^-[\mathsf{PA}]$ to PA readily implies that $\mathsf{CT}^-[\mathsf{PA}]$ does not exhibit significant speed-up over PA, i.e., there is a polynomial $p(n)$ such that for any arithmetical sentence ϕ, if ϕ is provable in $\mathsf{CT}^-[\mathsf{PA}]$ by a proof of length n, then ϕ is provable in PA by a proof of length at most $p(n)$. This solves a problem posed by Enayat in 2012 [5].

The absence of significant speed up of $\mathsf{CT}^-[\mathsf{PA}]$ over PA implied by the feasible reducibility of $\mathsf{CT}^-[\mathsf{PA}]$ to PA exhibits a dramatic difference between $\mathsf{CT}^-[\mathsf{PA}]$ and $\mathsf{CT}^-[\mathsf{B}]$ for finitely axiomatized base theories B, since it is well known that for such theories B, $\mathsf{CT}^-[\mathsf{B}]$ has superexponential speed-up over B, and therefore, $\mathsf{CT}^-[\mathsf{B}]$ is not feasibly reducible to B.[4]

Our proof of the feasible reduction of $\mathsf{CT}^-[\mathsf{PA}]$ to PA includes the verification that PA proves the formal consistency of every finite subtheory of $\mathsf{CT}^-[\mathsf{PA}]$, thereby establishing that $\mathsf{CT}^-[\mathsf{PA}]$ is a reflexive theory. This result follows from Leigh's work [16]; and was also established by Enayat and Visser (unpublished) with the help of the "low basis theorem" of computability theory to arithmetize their model-theoretic proof of conservativity of $\mathsf{CT}^-[\mathsf{PA}]$ over PA. The proof presented here, however, is based on a simpler arithmetization of the Enayat–Visser construction and does not appeal to the low basis theorem; the syntactic analysis of this arithmetization forms one of the main ingredients of the proof of our main result.

We also employ the machinery developed for the proof of our main result to analyze two other prominent theories of truth, namely, $\mathsf{FS}^-[\mathsf{PA}]$ (Friedman–Sheard theory of truth over PA, with no extra induction), and $\mathsf{KF}^-[\mathsf{PA}]$ (Kripke–Feferman theory of truth over PA with no extra induction). More specifically, we show that $\mathsf{FS}^-[\mathsf{PA}]$ and $\mathsf{KF}^-[\mathsf{PA}]$ are both reflexive and feasibly reducible to PA. These results, in turn, show that both $\mathsf{FS}^-[\mathsf{PA}]$ and $\mathsf{KF}^-[\mathsf{PA}]$ are feasibly interpretable in PA and have at most polynomial speed-up over PA.

A word about the organization of the article is in order. Section 2 deals with arithmetical preliminaries and technical machinery that will be employed for establishing our principal results. Section 3 presents basic definitions and facts about the truth theories $\mathsf{CT}^-[\mathsf{PA}]$, $\mathsf{KF}^-[\mathsf{PA}]$, and $\mathsf{FS}^-[\mathsf{PA}]$, including outlines of the proofs of their conservativity over PA. The main results of the article are contained in Section 4, which contains the proofs of feasible reduction of $\mathsf{CT}^-[\mathsf{PA}]$, $\mathsf{KF}^-[\mathsf{PA}]$, and $\mathsf{FS}^-[\mathsf{PA}]$ to PA; these proofs should be viewed as refined arithmetizations of the conservativity proofs outlined in Section 3.3. In Section 4.4 we spell out the interpretability-theoretic ramifications of our work. Section 5 collects some open questions; and the Appendix (Section 6) consists of routine-but-technical proofs

[4] In [8], Corollary 8, Fischer uses a key theorem of Pudlák to verify that the theory $\mathsf{CT}^-[\mathsf{PA}]$ enriched with the axiom of internal induction (that states that all axioms of PA are true) has super-exponential speed-up over PA. However, if B is a finitely axiomatizable theory, then already CT^- can prove that all axioms of B are true and the whole speed-up argument can be repeated for $\mathsf{CT}^-[\mathsf{B}]$. In fact, the arguments of Section 5 of Fischer's aforementioned article imply that we do not even need the full compositional theory CT^- but rather its fragment PT^- with positive compositional axioms, since the truth predicate of PT^- has to satisfy the law of excluded middle on a definable cut. Since PT^- is contained in both KF^- and FS^-, we conclude that $\mathsf{KF}^-[\mathsf{B}]$ and $\mathsf{FS}^-[\mathsf{B}]$ both exhibit super-exponential speed-up over B for finitely axiomatized B.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

of certain results employed in the body of the article, as well as a glossary for the convenience of the reader.

§2. Setting the stage: arithmetical machinery. This section discusses basic notions and fundamental machinery that can be generally described as *refined arithmetization* of certain parts of proof theory and model theory that play a key role in the statements and proofs of our main results in Section 4. Note, however, that the material in Section 2.6 will be only employed in Section 4.4.

2.1. Arithmetized syntax. The first-order theory PA (Peano Arithmetic) plays a central role in our article. PA is formulated in the functional language $\{0, S, +, \times\}$; its axioms are obtained by augmenting the axioms of Robinson's Arithmetic Q with the usual induction scheme for the whole language $\mathcal{L}_{\mathsf{PA}}$ of PA. Its intended model is the result of equipping the set of natural numbers $\mathbb{N}$ with the successor, addition, and multiplication functions. We will also denote $\mathbb{N}$ by ω, typically when treating it as a set of indices for some construction. Sometimes in this article, we will be referring to $\mathbb{N}$ or ω when working *within* PA, in which case these symbols simply refer to the ambient universe.

Crucially for our purposes, PA is capable of representing syntax. This means that in PA, one can employ recursion to define notions such as "term," "formula," or "proof in PA" similarly to how these notions are defined in Zermelo–Fraenkel set theory. This is a standard topic, covered, e.g., in [12] and [10].

Definition 2.1 (Coding conventions). We make the standard assumption that terms, formulae, and proofs are all coded by numbers as follows: if λ is a string of formal symbols corresponding to a term, a formula, or a formal proof,[5] then λ is first represented as a binary string $s_\lambda = (w_0, \ldots, w_n) \in \{0,1\}^{<\omega}$, and then coded as a number $\ulcorner \lambda \urcorner$ (the Gödel-number of λ), where:

$$\ulcorner \lambda \urcorner = \sum_{i=0}^{n} 2^i (w_i + 1).$$

Furthermore, we demand that λ and s_λ are feasibly computable from each other, i.e., there are polynomial time computable functions f and g such that for all strings of formal symbols λ, $f(\lambda) = s_\lambda$ and $g(s_\lambda) = \lambda$.[6]

Next we introduce numerals. For our purposes, we will need *two kinds* of numerals for representing a natural number n, the usual "tally" numeral, denoted $\mathsf{tal}(n)$; and an efficient one, denoted $\underline{n}$, which is based on the binary expansion of n. Tally numerals are employed in the formulation of many of our results that are related to polynomial time computations; see the comments at the end of Remark 2.17 for a general explanation of how they are employed.

[5] We will assume that our proof system is one of the "textbook" proof systems (e.g., a Hilbert-style one, or one based on the Sequent Calculus), where proofs can readily be coded as strings. It is well known that any two "textbook" proof systems for first-order logic feasibly simulate each other (see, e.g., Sections 2.5 and 4 of [18] and the references therein).

[6] Indeed f and g can be chosen to be linear time functions. See Remark 2.15 for the definition of polynomial time computability.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

DEFINITION 2.2 (Numerals). Let $n \in \mathbb{N}$. The *tally numeral* representing n, denoted $\mathsf{tal}(n)$, is defined recursively by: $\mathsf{tal}(0) = 0$ and $\mathsf{tal}(n+1) = S(\mathsf{tal}(n))$. On the other hand, the *binary numeral* representing n, denoted $\underline{n}$, is the binary expansion of n written as a term in the usual language of arithmetic. More precisely, let $n = \sum_{i \leq k} \varepsilon_i 2^i$, where $\varepsilon_i \in \{0, 1\}$. We define

$$\underline{n} = \mathsf{tal}(\varepsilon_0) + \mathsf{tal}(2) \times (\mathsf{tal}(\varepsilon_1) + \mathsf{tal}(2) \times (\ldots \mathsf{tal}(\varepsilon_{k-1}) + \mathsf{tal}(2) \times \mathsf{tal}(\varepsilon_k) \ldots).$$

Thus it takes $O(n)$ symbols to represent n as $\mathsf{tal}(n)$, but only $O(\log n)$ symbols to represent n as $\underline{n}$.[7]

Throughout the article we will use certain formulae to represent various syntactic and technical notions. For the convenience of our reader, we gather here all the notation which might be possibly confusing.

DEFINITION 2.3 (Arithmetized syntax). Fix a language $\mathcal{L}$. In what follows we systematically use the official expression "is the code of," but later, in the interest of better readability, we will use the common convention of confusing syntactic objects with their codes.

- If s is the code of a binary string, then $|s|$ denotes the length of s. On the other hand, if s is the code of a *sequence* of binary strings, then we use $\mathsf{len}(s)$ to denote the length of s.
- $\mathsf{Term}_{\mathcal{L}}(x)$ asserts: "x is the code of a term of $\mathcal{L}$." For instance, $\mathsf{Term}_{\mathcal{L}_{\mathsf{PA}}}(x)$ asserts that x is the code of an arithmetical term.
- $\mathsf{ClTerm}_{\mathcal{L}}(x)$ asserts: "x is the code of a closed term (i.e., without free variables) of $\mathcal{L}$."
- $\mathsf{TermSeq}_{\mathcal{L}}(x)$ asserts: "x is the code of a sequence of terms of $\mathcal{L}$."
- $\mathsf{ClTermSeq}_{\mathcal{L}}(x)$ asserts: "x is the code of a sequence of closed terms of $\mathcal{L}$."
- $x^{\circ} = y$ asserts: "$\mathsf{ClTerm}_{\mathcal{L}}(x)$ and y is the value of the term coded by x." For instance, the following holds:

$$\ulcorner 1 + ((1+1)+0) \urcorner^{\circ} = 3.$$

- $\mathsf{Var}(x)$ asserts: "x is the code of a variable." For instance, $\mathsf{Var}(17)$ means that 17 is the code of a variable. Since without loss of generality we can assume that all first-order languages have the same set of variables, we omit the reference to a specific language in the subscript.
- $\mathsf{Form}_{\mathcal{L}}(x)$ asserts: "x is the code of a formula of $\mathcal{L}$."
- $\mathsf{FV}(x, y)$ asserts: "y is the code of a free variable of x," where x is either the code of a term or a formula.
- $\mathsf{Form}^{\leq 1}_{\mathcal{L}}(x)$ asserts: "x is the code of a formula of $\mathcal{L}$ with at most one free variable"; and $\mathsf{Form}^{1}_{\mathcal{L}}(x)$ asserts: "x is the code of a formula of $\mathcal{L}$ with exactly one free variable."
- $\mathsf{Sent}_{\mathcal{L}}(x)$ asserts: "x is the code of a sentence of $\mathcal{L}$."
- $\mathsf{FVSeq}(x, y)$ asserts: "y is the code of a sequence whose elements are (some) free variables of the term or the formula coded by x."

[7] Here the "big O" notation is defined as usual: Given functions f and g from $\mathbb{N}$ to $\mathbb{N}$, $f = O(g)$ means that for some constant M and for some $n_0 \in \mathbb{N}$, $f(n) \leq Mg(n)$ for all $n \geq n_0$.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

- The expression "α is the code of an assignment for (a formula or a term) ϕ" (also referred to as "α is a ϕ-assignment," or "α is a ϕ-valuation") means that α is the code of a function whose domain includes the free variables of ϕ. The formula $\mathsf{Asn}(x, y)$ asserts: "y is the code of an assignment for the formula or the term coded by x." We will often denote it with $y \in \mathsf{Asn}(x)$. If s is a coded set or sequence, we will write $y \in \mathsf{Asn}(s)$ to denote that y is the code of an assignment for all elements of s. We will sometimes also write $\mathsf{Asn}(x_1, \ldots, x_n, \alpha)$ or $\alpha \in \mathsf{Asn}(x_1, \ldots, x_n)$ meaning $\alpha \in \mathsf{Asn}(s)$, where s is the code for the tuple $(x_1, \ldots, x_n)$.
- $\beta \sim_v \alpha$ asserts: "α and β are the codes of assignments, v is the code of a variable, and $\alpha(w) = \beta(w)$ for all variables w, possibly except for v which belongs to the domain of β (and not necessarily to the domain of α)."

The reader could expect to see in the above list certain other predicates such as Proof or Con. Since we will need some more precise information about these formulae and their lengths, they will be only introduced in Section 2.3.

Let us introduce one more definition.

DEFINITION 2.4. We employ the following notations in relation to substitutions.

- Let $\phi(v_1, \ldots, v_n)$ be a formula with n free variables shown and let α be an assignment for ϕ. By $\phi[\alpha]$ we denote the formula in which the binary numeral (in the sense of Definition 2.2) denoting $\alpha(v_i)$ is substituted for the variable v_i.
- Similarly, if $t \in \mathsf{Term}_{\mathcal{L}}$ and $\alpha \in \mathsf{Asn}(t)$, then by $t[\alpha]$ we mean the closed term obtained by substituting the numeral $\alpha(v)$ for each free variable v in t.
- If $t \in \mathsf{Term}_{\mathcal{L}}$ and $\alpha \in \mathsf{Asn}(t)$, then by t^{α} we mean $t[\alpha]^{\circ}$. Notice that if v is a variable and $\alpha \in \mathsf{Asn}(v)$, then $v^{\alpha} = \alpha(v)$ provably in PA.
- If $x \in \mathsf{Form}_{\mathcal{L}}$ for some language $\mathcal{L}$, $v \in \mathsf{Var}$ and $t \in \mathsf{Term}_{\mathcal{L}}$, then $x[t/v] = y$ is an arithmetical formula which asserts "y is the effect of substituting in the formula x the term t for every free occurrence of the variable v."

CONVENTION 2.5. We adopt the following conventions concerning formalized syntactic notions.

- Recall from Definition 2.1 that formulae are represented by binary strings. In this context, $|\lambda|$ refers to the length of the binary string representing λ. Note that $\ulcorner\lambda\urcorner$ is of size exponential in $|\lambda|$ and $\underline{\ulcorner\lambda\urcorner}$ is of size linear in $|\lambda|$.
- We abbreviate $\underline{\ulcorner\phi\urcorner}$ as $\underline{\phi}$ for a standard formula ϕ.
- We will sometimes use the formulae defining syntactic notions as if they were denoting sets. For example, we will sometimes write "$x \in \mathsf{Form}_{\mathcal{L}_{\mathsf{PA}}}$" rather than "$\mathsf{Form}_{\mathcal{L}_{\mathsf{PA}}}(x)$."
- We will use provably functional formulae such as $\underline{x}$, x°, or $x[t/v]$ as if they were terms.
- For better readability we will sometimes skip formulae denoting syntactic operations and write the effect of the operations instead. Thus, for example, we will write $T(\neg\phi)$ to denote "There exists ψ which is the negation of the sentence ϕ and $T(\psi)$."

2.2. Arithmetized model theory. Peano arithmetic is capable of accommodating a substantial part of the model theory of countable structures. We will make constant

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

use of this fact throughout the whole article. This subsection briefly introduces the reader to this topic. The rough convention is as follows: a *theory* is a definable set of sentences. If ϕ is a formula which defines a set of (codes of) sentences, then we call *that formula* a theory.

Models come in two kinds. By a *full model* $\mathcal{M}$ we mean the elementary diagram of that model (or, actually, a formula defining the elementary diagram). It is given as a complete Henkinized theory. By a *model* $\mathcal{M}$, we mean a formula defining its domain and some relations on that domain (this does not mean that we only deal with models of relational languages, but rather, we construe the denotations of function and constant symbols as relations).

DEFINITION 2.6 (Theories, full models, models). Let $\mathcal{L}$ be a language.

- A formula ϕ defines a *theory* in $\mathcal{L}$ if for all x, $\phi(x) \rightarrow \left(x \in \mathsf{Sent}_{\mathcal{L}}\right)$ holds.
- Let $\mathcal{T}$ be a theory in $\mathcal{L}$. By a *full model of* $\mathcal{T}$, we mean a theory $\mathcal{T}' \supseteq \mathcal{T}$ in a language $\mathcal{L}'$ extending $\mathcal{L}$ with some constants (possibly trivially) such that:
  1. $\mathcal{T}'$ is complete and consistent, so for any sentence ϕ of $\mathcal{L}'$, $\phi \in \mathcal{T}'$ if and only if $\neg\phi \notin \mathcal{T}'$; and
  2. $\mathcal{T}'$ has all existential statements witnessed by constants from $\mathcal{L}'$, which means that if $\exists x \phi(x) \in \mathcal{T}'$, then for some constant c in $\mathcal{L}'$, $\phi(c) \in \mathcal{T}'$.
- By a *full model of* $\mathcal{L}$ we mean a full model of some theory $\mathcal{T}$ in $\mathcal{L}$.
- By a *model* of $\mathcal{L}$ (or simply an $\mathcal{L}$-structure), we mean a formula $\mathcal{M}$ which defines a set of (coded) sequences such that if $\mathcal{M}(s)$ holds, then the following hold:
  1. $s(0)$ is either a symbol of $\mathcal{L}$ or some fixed element d which is not a symbol of $\mathcal{L}$. The intent is as follows: $s(0) = d$ means that $s(1)$ is an element of the domain; otherwise, $s(0)$ is a relation or function symbol and the rest of tuple are the elements linked by that relation.
  2. If $s(0)$ is a relation symbol of $\mathcal{L}$, then the length of s is the arity of $s(0)$ plus one.
  3. If $s(0)$ is a function symbol of $\mathcal{L}$, then the length of s is the arity of $s(0)$ plus two. (We treat constants as functions of arity zero.)
  4. If $s(0)$ is the fixed element d, then s has length two.
  5. If $a = s(n)$ for some $n > 0$, then $\mathcal{M}(\langle d, a\rangle)$ holds. (Which means that a is in the domain of a model.)

In the above, a model is essentially defined as a particular kind of tuple: (definition of the domain, definition of the first relation, definition of the second relation, ...). We have formulated the above compact definition rather than defining a model as an actual tuple of formulae since we will need to allow models with infinite signatures. However, if a model is defined with a standard number of definable relations, we can easily construct a definition in the format specified above. It is also important to bear in mind that although officially in this article a full model is the same as the *elementary diagram* of that model, in practice we will refer to models in the usual way, since it is clear how to transfer statements about models to statements about their elementary diagrams and vice versa. Finally, note that in our setting, a full model can be naturally identified with a model, but not vice versa since by Tarski's

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

undefinability of truth theorem, given any arithmetical formula $\phi(x)$, PA proves that ϕ does not define the elementary diagram of the ambient model of arithmetic.

DEFINITION 2.7. Let $\mathcal{L}$ be a language. The following are to be understood in the context of Definition 2.6.

- If $\mathcal{M}$ is a full model of $\mathcal{L}$, we write $x \in M$ to say that x is a constant $\mathcal{L}$. (This means that x is an element of $\mathcal{M}$, since we implicitly assume that all full models are built on Henkin constants.) If $\mathcal{M}$ is a model, the expression $x \in M$ means that $\mathcal{M}(\langle d, x\rangle)$ holds.
- If $\mathcal{M}$ is a full model of $\mathcal{L}$, and $\phi \in \mathsf{Form}_{\mathcal{L}}$, we say that α is an $\mathcal{M}$-*assignment* (or an $\mathcal{M}$-*valuation*) for a formula ϕ if α is a (coded) finite function, whose domain contains $\mathsf{FV}(\phi)$, $\alpha \in \mathsf{Asn}(\phi)$, and for every x, $\alpha(x) \in M$. We denote this by $\alpha \in \mathsf{Asn}(\phi, \mathcal{M})$.
- If $\mathcal{M}$ is a full model of $\mathcal{L}$, $\phi \in \mathsf{Form}_{\mathcal{L}}$, and α is an $\mathcal{M}$-assignment, then the relation $\mathcal{M} \models \phi[\alpha]$ is defined simply as $\phi(\alpha(x_1), \ldots, \alpha(x_c)) \in \mathcal{M}$.
- If $\mathcal{M}$ is a model of $\mathcal{L}$, $\phi \in \mathsf{Form}_{\mathcal{L}}$, and α is an $\mathcal{M}$-assignment, then the relation $\mathcal{M} \models \phi[\alpha]$ is defined only for ϕ of standard complexity via the usual compositional conditions with quantifiers restricted to the domain of $\mathcal{M}$ and satisfaction for base relations $R \in \mathcal{L}$ (of arity c) defined as follows:

$$\mathcal{M} \models R[\alpha] \text{ iff } \mathcal{M}\big(\langle R, \alpha(v_1), \ldots, \alpha(v_c)\rangle\big).$$

  We define satisfaction for equalities of terms in an analogous fashion.
- If $\mathcal{M}$ is a full model for $\mathcal{L}$, we will write $\mathsf{ElDiag}(\mathcal{M})$ (elementary diagram of $\mathcal{M}$) instead of $\mathcal{M}$ when we want to stress that we are thinking of a theory rather than of a structure.
- If $\mathcal{M}$ is a (full) model of $\mathcal{L}$, $a_1, \ldots, a_c \in M$ and $\phi(v_1, \ldots, v_c)$ is a formula with the displayed free variables, we will write:

$$\mathcal{M} \models \phi(a_1, \ldots, a_c)$$

  meaning there exists an $\mathcal{M}$-valuation α for ϕ such that $\alpha(v_i) = a_i$ for all $i < c$ and $\mathcal{M} \models \phi[\alpha]$.
- If $\mathcal{M}$ is a (full) model of $\mathcal{L}$ and $\phi(v_1, \ldots, v_c) \in \mathsf{Form}_{\mathcal{L}}$ with all free variables displayed, then by $\phi(\mathcal{M})$ we mean the set of (tuples of) elements defined by the formula ϕ in $\mathcal{M}$. In other words, it is the set of tuples $(a_1, \ldots, a_c)$ of the elements of $\mathcal{M}$ such that $\mathcal{M} \models \phi[\alpha]$ for some (equivalently, any) $\alpha \in \mathsf{Asn}(\phi)$ such that $\alpha(v_i) = a_i, i \leq c$.

Note that we have not yet defined what it means that a model satisfies a theory. This is not an omission. Since for general (not full) models, satisfaction is defined only for standard sentences, we only define satisfaction for standard formulae. This is actually a scheme: for each formula we define what it means that a model satisfies this formula. More precisely: for each $n \in \mathbb{N}$, we define what it means that a model satisfies a formula of depth n.

On the other hand, in our article, nonfull models will play a crucial role and in some specific circumstances we are going to say that a model satisfies a theory. This will be defined in some specific cases that are of interest to us later in Definition 2.34.

Let us define some more notions which will be particularly important in further parts of our article.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

DEFINITION 2.8. Let $\mathcal{M}, \mathcal{N}$ be full models of theories in the same language. We say that $\mathcal{N}$ is an *elementary extension* of $\mathcal{M}$ if $\mathcal{M}$ is contained in $\mathcal{N}$.

Recall that officially, a full model is the same as the elementary diagram of that model, so elementary submodels in our sense correspond to elementary submodels in the usual sense. In what follows, we will sometimes follow the usual practice of conflating *elementary submodels* with *images of elementary embeddings*. This should be understood in the obvious way: a formula $\phi(x, y)$ defines an elementary embedding of the model $\mathcal{M}$ into the model $\mathcal{N}$ if it defines an injection from the elements of the model $\mathcal{M}$ into the elements of the model $\mathcal{N}$ (i.e., it defines a relation on constants such that $\phi(a, b_1)$ and $\phi(a, b_2)$ together imply $(b_1 = b_2) \in \mathcal{N}$; and $\phi(a_1, b)$ and $\phi(a_2, b)$ together imply that $(a_1 = a_2) \in \mathcal{M}$); and the image of the injection is an elementary submodel of $\mathcal{M}$ (i.e, the restriction of $\mathcal{N}$ to the language with constants representing the image of $\mathcal{M}$ is a full model).

We will denote both being an elementary submodel and being an image of an elementary embedding with

$$\mathcal{M} \preceq \mathcal{N}.$$

DEFINITION 2.9. If $\mathcal{M}, \mathcal{N}$ are full models of two languages $\mathcal{L}_{\mathcal{M}}, \mathcal{L}_{\mathcal{N}}$, respectively, and $\mathcal{L} \subseteq \mathcal{L}_{\mathcal{M}} \cap \mathcal{L}_{\mathcal{N}}$, we say that $\mathcal{M}$ is an $\mathcal{L}$-*elementary submodel* of $\mathcal{N}$, written $\mathcal{M} \preceq_{\mathcal{L}} \mathcal{N}$, if

$$\mathcal{M} \cap \mathsf{Sent}_{\mathcal{L}} \subseteq \mathcal{N} \cap \mathsf{Sent}_{\mathcal{L}},$$

where $\mathsf{Sent}_{\mathcal{L}}$ is the set of $\mathcal{L}$-sentences.

Similar to the case of full elementarity, we conflate $\mathcal{L}$-elementary submodels with images of $\mathcal{L}$-elementary embeddings.

DEFINITION 2.10. Let $\mathcal{M}, \mathcal{N}$ be models or full models in languages $\mathcal{L}_{\mathcal{M}}$, $\mathcal{L}_{\mathcal{N}}$ respectively. We say that $\mathcal{N}$ is an *expansion* of $\mathcal{M}$ if the following are satisfied:

1. $\mathcal{L}_{\mathcal{M}} \subseteq \mathcal{L}_{\mathcal{N}}$.
2. For every element $a \in N$ there is an element $b \in M$ such that $\mathcal{N} \models x = y[\alpha]$, where $\alpha(x) = a$, $\alpha(y) = b$. (That is, the domain does not change. We write it in this slightly convoluted manner, since we want the definition to work both for models and full models.)
3. For every atomic formula $\phi \in \mathcal{L}_{\mathcal{M}}$ and $\mathcal{M}$-assignment α for ϕ, $\mathcal{M} \models \phi[\alpha]$ if and only if $\mathcal{N} \models \phi[\alpha]$.

CONVENTION 2.11. Throughout the article, we will be using the following handy conventions concerning models:

- When there is no risk of confusion, we will use the same symbol for a predicate symbol and for its denotation in a given (full) model.
- We will sometimes denote (full) models as tuples, like $(\mathcal{M}, T)$. This will simply mean that $(\mathcal{M}, T)$ is an expansion of $\mathcal{M}$ with a predicate T.

We can say that PA is capable of handling basic model theory, since it is able to capture the link between models and consistent theories. In the context of our definitions, this means that in PA every consistent theory can be extended to a complete consistent theory with Henkin constants. More precisely, if we define a Δ_n-theory to be a theory defined both with a Σ_n-formula and a Π_n-formula, and

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

analogously define the notion of a Δ_{n+1}-full model, then by using the "left-most branch" proof of König's Lemma we can readily verify the following standard theorem (cf. Section 13.2 of Kaye's book [12]).

THEOREM 2.12 (Arithmetized Completeness Theorem). *For each* $n \in \mathbb{N}$, PA *proves that every consistent* Δ_n*-theory has a* Δ_{n+1}*-full model.*

2.3. Proof simulations and reductions. This section summarizes the basic definitions and tools that we will need in connection with analyzing the length and complexity of proofs. Much of this material is standard and taken from Pudlák's articles [17] and [18].[8]

- Throughout the section we identify a theory with the set of its axioms, thus theories need not be closed under deductions. This is consistent with how theories are arithmetically handled, as in Definition 2.6.

The following definition provides us with a useful distance function between formulae and theories.

DEFINITION 2.13. Given a theory $\mathcal{T}$ and an $\mathcal{L}_\mathcal{T}$-formula ϕ, $\| \phi \|_\mathcal{T}$ is defined by:

$$\| \phi \|_\mathcal{T} = \begin{cases} \text{the length of the shortest proof of } \phi, \text{ if } \mathcal{T} \vdash \phi; \\ \infty \text{ otherwise.} \end{cases}$$

We write $\mathcal{T} \vdash^n \phi$ as shorthand for $\| \phi \|_\mathcal{T} \leq n$.

DEFINITION 2.14 (Simulations, speed-up, reducibility). Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two theories and $\mathcal{F}$ a family of functions $f : \mathbb{N} \to \mathbb{N}$.

- $\mathcal{T}_1$ $\mathcal{F}$*-simulates* $\mathcal{T}_2$ means that there exists a function $f \in \mathcal{F}$ such that for every sentence $\phi \in \mathcal{L}_{\mathcal{T}_1} \cap \mathcal{L}_{\mathcal{T}_2}$, and for every $n \in \mathbb{N}$, we have:
$$\mathcal{T}_2 \vdash^n \phi \Rightarrow \mathcal{T}_1 \vdash^{f(n)} \phi.$$
- $\mathcal{T}_2$ is $\mathcal{F}$*-reducible* to $\mathcal{T}_1$ means that there exists a function $f \in \mathcal{F}$ such that for every sentence $\phi \in \mathcal{L}_{\mathcal{T}_1} \cap \mathcal{L}_{\mathcal{T}_2}$ and every $n \in \mathbb{N}$ we have:
$$n \text{ codes a } \mathcal{T}_2\text{-proof of } \phi \Rightarrow f(n) \text{ codes a } \mathcal{T}_1\text{-proof of } \phi.$$
- $\mathcal{T}_2$ has *super-*$\mathcal{F}$ *speed-up* over $\mathcal{T}_1$ means that $\mathcal{T}_1$ does not $\mathcal{F}$-simulate $\mathcal{T}_2$.
- $\mathcal{T}_2$ is *feasibly reducible* to $\mathcal{T}_1$ means that $\mathcal{T}_2$ is $\mathcal{F}$-reducible to $\mathcal{T}_1$ for the family $\mathcal{F}$ of polynomial time (hereafter P-time) computable functions (see Remark 2.15 for the definition of polynomial time computability).

Note that if $\mathcal{T}_2$ is proof-theoretically reducible to $\mathcal{T}_1$ (i.e., the conservativity of $\mathcal{T}_2$ over $\mathcal{T}_1$ is provable in Primitive Recursive Arithmetic), then $\mathcal{T}_2$ is $\mathcal{F}$-reducible to $\mathcal{T}_1$ for the family $\mathcal{F}$ of primitive recursive functions. Other typical examples in the literature of simulation/speed-up phenomena concern the cases where $\mathcal{F}$ is either the family of polynomial functions, or the family of exponential (also known as elementary) functions, which respectively correspond to polynomial simulation/super-*polynomial* speed-up; and exponential simulation/super-*exponential* speed-up.

[8] Philosophical motivations for studying lengths of proofs have been presented by Caldon and Ignjatović [1] (in a general setting) and by Fischer [8] (in the setting of axiomatic theories of truth).

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

REMARK 2.15 (P-time computable functions).

- We will refer to polynomial time computable functions as P-time computable functions. These functions are also commonly referred to as *feasibly computable* functions. The inputs and outputs of a P-time computable function f are *strings*, i.e., finite sequences of formal symbols; such a function f should have the property that there is a deterministic Turing machine which computes f in polynomial time, i.e., there is a polynomial function $p(n)$ from $\mathbb{N}$ to $\mathbb{N}$ such that for each input string s the output string $f(s)$ is computed by the Turing machine in at most $p(|s|)$-many steps, where $|s|$ is the length of s.
- *The notion of P-time computability is then lifted to functions from* $\mathbb{N}$ *to* $\mathbb{N}$ *by identifying each* $n \in \mathbb{N}$ *with the binary sequence representing* n (equivalently, n can be identified with the term $\underline{n}$ since the binary representation of n and $\underline{n}$ are feasibly computable from each other).

In light of our coding conventions stated in Definition 2.1 and Remark 2.15 the following equivalent formulation of feasible reducibility can be readily verified.

LEMMA 2.16. *The following statements are equivalent for a pair of theories* $\mathcal{T}_1$ *and* $\mathcal{T}_2$:

1. $\mathcal{T}_1$ *is feasibly reducible to* $\mathcal{T}_2$.
2. *There is a P-time computable function* f *whose inputs and outputs are strings such that for every string* s, *if* s *is the proof of* ϕ *in* $\mathcal{T}_2$, *then* $f(s)$ *is the proof of* ϕ *in* $\mathcal{T}_1$.

REMARK 2.17 (Inputs and outputs of P-time computable functions).

- Throughout the rest of the article, when a P-time computable function f is viewed abstractly, it is construed as a function whose inputs and outputs are strings, not *codes* of strings. For this reason we use the letter s, or an indexed version of s, to denote inputs of f in order to remind the reader that the inputs of f are strings.
- Typically, the inputs of f will be *particular* kinds of strings, namely, strings corresponding to terms, formulae, or proofs; and the outputs will be strings corresponding to proofs. For example an expression such as $f(\underline{m}, \mathsf{tal}(n), \phi, \mathcal{M})$, means that the input string of f consists of four substrings (separated by appropriate markers), where the first two substrings correspond to special kinds of terms (the binary numeral for m, followed by the tally numeral for n), and the last two substrings correspond to formulae (where the second formula defines a model in the ambient theory at work).
- The tally numeral $\mathsf{tal}(n)$ is employed as an input in many of our results (especially in Sections 2.4 and 2.5) pertaining to P-time computable functions *since the relevant P-time computations require the subcomputation of a sequence of length n. Note that a sequence of length n is P-time computable (indeed linear time computable) from* $\mathsf{tal}(n)$, *but it is not P-time computable from* $\underline{n}$.

The proofs of the following observations are routine:

OBSERVATION 2.18. *Let* $\mathcal{F}$ *be any family of functions from* $\mathbb{N}$ *to* $\mathbb{N}$.

- *If* $\mathcal{T}_2$ *is* $\mathcal{F}$*-reducible to* $\mathcal{T}_1$, *then* $\mathcal{T}_1$ $\mathcal{F}$*-simulates* $\mathcal{T}_2$. *Note that if* $\mathcal{T}_2$ *is feasibly reducible to* $\mathcal{T}_1$, *then* $\mathcal{T}_1$ *polynomially simulates* $\mathcal{T}_2$.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

- *If $\mathcal{F}$ is countable, then $\mathcal{T}_2$ has a super $\mathcal{F}$ speed-up over $\mathcal{T}_1$ if there exists an infinite sequence of formulae $\phi_0, \phi_1, \ldots$, provable in both $\mathcal{T}_1$ and $\mathcal{T}_2$ such that for every function $f \in \mathcal{F}$ there exists $n \in \mathbb{N}$ such that*:

$$\| \phi_n \|_{\mathcal{T}_1} > f(\| \phi_n \|_{\mathcal{T}_2}).$$

The most prominent role in the investigations in the lengths of proofs is played by consistency statements. We shall now discuss arithmetized provability. Recall that $|\underline{n}|$ denotes the length of the binary numeral $\underline{n}$ representing n.

DEFINITION 2.19 (Pudlák, [17]). Let $\mathcal{T}$ be a theory, $\phi(x_0, \ldots, x_k)$ be a formula and $R \subseteq \mathbb{N}^{k+1}$ be a relation. We say that ϕ *polynomially numerates* R in $\mathcal{T}$ if there exists a polynomial $p(x_0, \ldots, x_k)$ such that for all natural numbers $n_0, \ldots, n_k$ we have:

$$R(n_0, \ldots, n_k) \text{ iff } \| \phi(\underline{n_0}, \ldots, \underline{n_k}) \|_{\mathcal{T}} \leq p(|\underline{n_0}|, \ldots, |\underline{n_k}|).$$

THEOREM 2.20 (Pudlák, [17] Theorem 3.2). *For any consistent NP-time*[9] *theory $\mathcal{T} \supseteq \mathsf{Q}$ (where Q is Robinson's Arithmetic) and any $R \subseteq \mathbb{N}^k$ the following are equivalent*:

1. *R is an NP-time computable relation*.
2. *R is polynomially numerable in Q*.
3. *R is polynomially numerable in $\mathcal{T}$*.

We will use a modification of Pudlák's result (which actually is simpler than the original theorem) since, firstly, we want slightly stronger results concerning *feasible reducibility* between theories rather than mere facts about speed-up, and secondly, we work with relatively strong theories.

DEFINITION 2.21. Let $\mathcal{T}$ be a theory, $\phi(x_0, \ldots, x_k)$ be a formula and $R \subseteq \mathbb{N}^{k+1}$ be a relation. We say that ϕ *feasibly numerates* R in $\mathcal{T}$ if there exists a P-time computable function $f(s_0, \ldots, s_k)$ such that for all natural numbers $n_0, \ldots, n_k$, $R(n_0, \ldots, n_k)$ holds iff $f(\underline{n_0}, \ldots, \underline{n_k})$ is a $\mathcal{T}$-proof of $\phi(\underline{n_0}, \ldots, \underline{n_k})$.

In what follows, we need only the following simple fact which may be proved by a natural formalization of Turing machines in $\mathsf{I}\Delta_0 + \mathsf{Exp}$. Its proof is significantly simpler than the proof of Theorem 2.20 since we do not need to consider cuts or use cut-shortening techniques.

THEOREM 2.22. *For any P-time computable theory $\mathcal{T} \supseteq \mathsf{I}\Delta_0 + \mathsf{Exp}$, and any $R \subseteq \mathbb{N}^k$, the following are equivalent*:

1. *R is a P-time computable relation*.
2. *R is feasibly numerable in $\mathsf{I}\Delta_0 + \mathsf{Exp}$*.
3. *R is feasibly numerable in $\mathcal{T}$*.

DEFINITION 2.23. Given a formula $\mathcal{T}(x)$ describing the axioms of a theory $\mathcal{T}$, we use $\mathcal{T} \restriction y$ to denote the theory whose axioms are defined by the formula $\mathcal{T} \restriction y\ (x)$, where:

$$\mathcal{T} \restriction y\ (x) := \mathcal{T}(x) \wedge (|x| \leq y)\,.$$

[9] An NP-time computable set or relation is one whose characteristic function can be computed by a non-deterministic Turing machine that runs in polynomial time.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

COROLLARY 2.24. *Suppose $\mathcal{T}$ is a P-time computable theory.*

1. *There is a binary arithmetical formula $\mathsf{Proof}_{\mathcal{T}}(x, y)$ expressing "x is a $\mathcal{T}$-proof of y", and a P-time computable function $f(s_0, s_1)$ such that for all m and n in $\mathbb{N}$ we have*:

$$\mathbb{N} \models \mathsf{Proof}_{\mathcal{T}}(m, n) \text{ iff } f(\underline{m}, \underline{n}) \text{ is a proof of } \mathsf{Proof}_{\mathcal{T}}(\underline{m}, \underline{n}) \text{ in } \mathsf{I}\Delta_0 + \mathsf{Exp}.$$

2. *There is a ternary arithmetical formula $\mathsf{Proof}_{\mathcal{T}\restriction z}(x, y)$ expressing "x is a $\mathcal{T} \restriction z$-proof of y" and a P-time computable function $g(s_0, s_1, s_2)$ such that for all m, n, and r in $\mathbb{N}$ we have*:

$$\mathbb{N} \models \mathsf{Proof}_{\mathcal{T}\restriction r}(m, n) \text{ iff } g(\underline{m}, \underline{n}, \underline{r}) \text{ is a proof of } \mathsf{Proof}_{\mathcal{T}\restriction \underline{r}}(\underline{m}, \underline{n}) \text{ in } \mathsf{I}\Delta_0 + \mathsf{Exp}.$$

Moreover, we define formulae Con and Pr in the usual manner as follows:

$$\mathsf{Pr}_{\mathcal{T}}(y) := \exists x \mathsf{Proof}_{\mathcal{T}}(x, y),$$

and

$$\mathsf{Con}_{\mathcal{T}} := \neg \mathsf{Pr}_{\mathcal{T}}(\underline{0 = 1}).$$

REMARK 2.25 (Relativized provability predicates). The content of this remark will be needed only in Section 4.4. The formalization of the provability predicate from Corollary 2.24 is of the form: "There exists an accepting computation of the Turing machine which recognizes $\mathcal{T}$-proofs". Let $\phi(x)$ be any arithmetical formula. By writing

$$\mathsf{Proof}^{\phi}_{\mathcal{T}}(x, y),$$

we mean the relativized version of the above predicate, i.e., the one in which the relevant Turing machine is supplied with an oracle given by ϕ and recognizes the theorems of $\mathcal{T} + \phi$ (whatever ϕ means). We can treat $\mathcal{T} + \phi$ as a new arithmetized theory, but then in typical cases it won't be Δ_1. This is why we decided to distinguish between the roles played by the lower and the upper indices in $\mathsf{Proof}^{\phi}_{\mathcal{T}}(x, y)$; the former will be reserved for P-time computable theories $\mathcal{T}$ which satisfy conditions (1) and (2) of Corollary 2.24, and the latter for arbitrary formulae ϕ that act as oracles. Obviously the relativized version of Corollary 2.24 need not be true, but we will only demand that the following two conditions hold:

1. There exists a P-time computable function $f(s_0, s_1, s_2)$ such that for all $n \in \mathbb{N}$ and all $\mathcal{L}_{\mathcal{T}}$-formulae $\phi(x)$, $f(\underline{n}, \phi, \mathcal{T})$ is a PA-proof of:

$$(\phi(\underline{n}) \to \mathsf{Proof}^{\phi}_{\mathcal{T}}(\underline{n}, \underline{n})). \tag{RelProv1}$$

(Note that we are identifying a proof consisting of one formula with that formula.)

2. Likewise, there exists a P-time computable function $g(s_0, s_1, s_2)$ such that $g(\phi, \psi, \mathcal{T})$ is a PA-proof of the sentence:

$$\forall x \left(\phi(x) \to \psi(x)\right) \to \forall y \forall z \left(\mathsf{Proof}^{\phi}_{\mathcal{T}}(y, z) \to \mathsf{Proof}^{\psi}_{\mathcal{T}}(y, z)\right). \tag{RelProv2}$$

This requires that $\mathsf{Proof}^{\phi}_{\mathcal{T}}(y, z)$ be constructed uniformly in ϕ, which can certainly be arranged.

In particular, $\mathsf{Proof}^{\phi}_{\mathcal{T}}(x, y)$ is of length polynomial in the lengths of ϕ and the chosen definition of $\mathcal{T}$. As usual, let:

$$\mathsf{Pr}^{\phi}_{\mathcal{T}}(y) := \exists x \mathsf{Proof}^{\phi}_{\mathcal{T}}(x, y),$$

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

and

$$\mathsf{Con}^{\phi}_{\mathcal{T}} := \neg \mathsf{Pr}^{\phi}_{\mathcal{T}}(\underline{0 = 1}).$$

The following theorem gives a canonical example of a family of sentences whose proofs grow super-exponentially. To state the theorem concisely, we use $\mathsf{Con}_{\mathcal{T}}(x)$ to express "there is no $\mathcal{T}$-proof of $\underline{0 = 1}$ whose length is below x."

THEOREM 2.26 (Pudlák, [18], Theorem 7.2.2). *Let $\mathcal{T}$ be a sufficiently strong theory. Let f be an increasing computable function, provably total in $\mathcal{T}$, whose graph has a polynomial numeration in $\mathcal{T}$. Then there exists a $\delta > 0$ such that*

$$\| \mathsf{Con}_{\mathcal{T}}(f(\underline{n})) \|_{\mathcal{T}} > f(n)^{\delta}.$$

In particular, for $f(n) := 2_n$, where $2_0 := 1$, and $2_{n+1} := 2^{2_n}$, there is some $\delta > 0$ such that

$$\| \mathsf{Con}_{\mathsf{I}\Sigma_1}(2_{\underline{n}}) \|_{\mathsf{I}\Sigma_1} > 2^{\delta}_n; \text{ and}$$
$$\| \mathsf{Con}_{\mathsf{PA}}(2_{\underline{n}}) \|_{\mathsf{PA}} > 2^{\delta}_n.$$

2.4. Feasible truth predicates. Now we turn to arithmetized partial truth predicates, which we want to apply to arbitrary sentences of a fixed complexity, where the relevant notion of complexity is the *depth* of a formula, and is therefore different from the usual notion of arithmetical complexity (i.e., Σ_n, Π_n).

DEFINITION 2.27 (Depth of a formula).

- The *depth* of a formula is the length of the longest path in its syntactic tree, which is allowed to contain arbitrary terms as leaves (for example: the depth of $(0 = 0) \wedge \forall x \neg (SSS(x) = 0)$ is 3).
- $\mathsf{dp}(x, y)$ denotes the arithmetical formula asserting that the depth of a formula x is at most y. We will also write it as $\mathsf{dp}(x) \leq y$.

Next we recall the notion of a sequential theory, which provides the general setting for Pudlák's result.

DEFINITION 2.28 (Pudlák, [17]). A theory $\mathcal{T}$ is *sequential*, if Robinson's arithmetic Q is interpretable in $\mathcal{T}$ relativized to some formula $N(x)$ of $\mathcal{L}_{\mathcal{T}}$ and there exists a formula $(x)_t$ (of two variables x, t) that defines in $\mathcal{T}$ a *total* function (in both variables) and such that $\mathcal{T}$ proves:

$$\forall x, y, t \exists z \ \Big(N(t) \rightarrow \forall s < t \ \ \big((x)_s = (z)_s \wedge (z)_t = y \big) \Big).$$

Pudlák showed that a sequential theory supports a sequence of partial satisfaction predicates $\{\mathsf{Sat}_n\}_{n \in \mathbb{N}}$ such that $\mathcal{T}$ can "polynomially" verify that Sat_n is compositional for formulae of depth at most n; and moreover $\mathcal{T}$ can "polynomially" verify that Sat_n satisfies uniform Tarski biconditionals for all formulae of length at most n. This is made precise in Theorem 2.29, and refined in Theorem 2.30.

THEOREM 2.29 (Pudlák, [18], Theorem 3.3.1). *Let $\mathcal{T}$ be a sequential theory. There is a family $\{\mathsf{Sat}_n(x, y)\}_{n \in \mathbb{N}}$ of $\mathcal{L}_{\mathcal{T}}$-formulae and a polynomial $r_1(n)$ such that for each $n \in \mathbb{N}$, $\mathcal{T} \vdash^{r_1(n)} \theta_n$, where θ_n expresses*

"$\mathsf{Sat}_n(x, y)$ satisfies Tarski's compositional conditions for all formulae x of depth at most n, and for all x-valuations y."

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

Moreover, there is a polynomial $r_2(n)$ such that for every $\phi(x_1, \ldots, x_k)$ of length *at most n, $\mathcal{T} \vdash^{r_2(n)} \theta'_n$, where*

$$\theta'_n := \forall \alpha \in \mathsf{Asn}(\phi)\ \left(\mathsf{Sat}_n(\underline{\phi}, \alpha) \equiv \phi[\alpha]\right).$$

An inspection of Pudlák's proof of the above theorem makes it clear that in fact the proofs of polynomial length whose existence is asserted in the above theorem can be feasibly calculated. This means that Theorem 2.29 can be slightly strengthened as follows.

THEOREM 2.30. *Let $\mathcal{T}$ be a sequential theory. There is a family $\{\mathsf{Sat}_n(x, y)\}_{n\in\mathbb{N}}$ of $\mathcal{L}_{\mathcal{T}}$-formulae and a P-time computable function $f(s)$ such that for each $n \in \mathbb{N}$, $f(\mathsf{tal}(n))$ is a $\mathcal{T}$-proof of the sentence θ_n, where θ_n is as in Theorem* 2.29. *Moreover, there is a P-time computable function $g(s_0, s_1)$ such that for every $\phi(x_1, \ldots, x_k)$ of* length *less than n, $g(\phi, \mathsf{tal}(n))$ is a $\mathcal{T}$-proof of θ'_n, where θ'_n is as in Theorem* 2.29.

- In what follows, $\mathsf{Tr}_n(x)$ abbreviates $\mathsf{Sat}_n(x, \varnothing)$.

OBSERVATION 2.31. *There exists a P-time computable function $f(s_0, s_1)$ such that for every $n \in \mathbb{N}$ and for each $k \leq n$, $f(\mathsf{tal}(n), \mathsf{tal}(k))$ is a* PA*-proof of*:

$$\forall \phi \in \mathsf{Sent}_{\mathcal{L}_{\mathsf{PA}}}\ \left(\mathsf{dp}(\phi) \leq \underline{k} \rightarrow \mathsf{Tr}_n(\phi) \equiv \mathsf{Tr}_k(\phi)\right).$$

The proof uses induction (in PA) on the complexity of ϕ and provable Tarski biconditionals for Tr_l predicates.

Later in the article, we will also need the relativized version of the Theorem 2.30 (we state it only for PA). If $\mathcal{M}$ is a Δ_k-model (not necessarily a full one) for a language $\mathcal{L}'$ with finitely many fresh nonarithmetical relational symbols, then by $\mathcal{M}$-relativized Tarski's conditions for a formula $\Phi(x, y)$ we mean the usual statement that $\Phi(x, y)$ satisfies Tarski's compositional truth conditions in which the condition for atomic formulae is:

$$\forall \alpha \in \mathsf{Asn}(\phi, \mathcal{M})\ \ \Phi(\ulcorner R(s_0, \ldots, s_{n-1})\urcorner, \alpha) \equiv \mathcal{M} \models R(s_0, \ldots, s_{n-1})[\alpha]$$

for an arbitrary relation R in $\mathcal{L}'$, and the condition for the existential quantifier is given by:

$$\forall v \in \mathsf{Var}\, \forall \phi(v) \in \mathsf{Form}_{\mathcal{L}'}\, \forall \alpha \in \mathsf{Asn}(\phi, \mathcal{M})\ \Phi(\exists v \phi, \alpha) \equiv \exists \beta \sim_v \alpha\ (\beta \in \mathsf{Asn}(\phi, \mathcal{M}) \wedge \Phi(\phi, \beta)).$$

COROLLARY 2.32. *Let $k \in \mathbb{N}$ and suppose that* PA *proves that $\mathcal{M}$ is a Δ_k-model of a language $\mathcal{L}$. There is a family $\{\mathsf{Sat}_n^{\mathcal{M}}(x, y)\}_{n\in\mathbb{N}}$ of $\mathcal{L}_{\mathsf{PA}}$-formulae and a P-time computable function $f(s)$ such that for each $n \in \mathbb{N}$, $f(\mathsf{tal}(n))$ is a* PA*-proof of the sentence θ_n, where θ_n expresses*:

> *"$\mathsf{Sat}_n^{\mathcal{M}}(x, y)$ satisfies $\mathcal{M}$-relativized Tarski's compositional conditions for all formulae x of depth at most n, and for all $\mathcal{M}$-valuations y for x."*

Moreover, there is a P-time computable function $g(s_0, s_1)$ such that for every $\phi(x_1, \ldots, x_k)$ of length *less than n, $g(\phi, \mathsf{tal}(n))$ is a* PA*-proof of θ'_n, where*:

$$\theta'_n := \forall \alpha \in \mathsf{Asn}(\phi, \mathcal{M})\ \ \left(\mathsf{Sat}_n^{\mathcal{M}}(\underline{\phi}, \alpha) \equiv \mathcal{M} \models \phi[\alpha]\right).$$

The family $\mathsf{Sat}_n^{\mathcal{M}}(x, y)$ can be defined essentially by relativizing $\mathsf{Sat}_n(x, y)$ predicates from Theorem 2.29. Since the definition of $\mathcal{M}$ does not depend on n, the length of $\mathsf{Sat}_n^{\mathcal{M}}(x, y)$ will be polynomial in n.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

As above, we will write $\mathsf{Tr}_n^{\mathcal{M}}$ to denote satisfaction under the empty valuation. Occasionally, we will use the handy notational convention described below.

CONVENTION 2.33. If $\phi(v_1, \ldots, v_n) \in \mathsf{Form}$ and n is standard, we will be writing $\mathsf{Sat}_k(\underline{\phi}, a_1, \ldots, a_n)$ to denote:

$$\mathsf{Sat}_k(\underline{\phi}, \alpha),$$

where $\alpha \in \mathsf{Asn}(\phi)$ is some valuation which assigns a_i to the variable v_i for $1 \leq i \leq n$.

Let us end this subsection with a definition of satisfaction of theories for a larger class of models.

DEFINITION 2.34. Let $n \in \mathbb{N}$. Let $\mathcal{M} \models \mathsf{B}$ be a (full) model over a language $\mathcal{L}$ and let $(\mathcal{M}, T)$ be its expansion to an $\mathcal{L}'$-structure. Suppose that $\mathcal{T}$ is a theory over the language $\mathcal{L}'$ such that $\mathcal{T} \setminus \mathsf{B}$ consists only of sentences of depth $\leq n$. We will write:

$$(\mathcal{M}, T) \models \mathcal{T},$$

if $\mathsf{Tr}_n^{(\mathcal{M},T)}(\phi)$ holds for all $\phi \in \mathcal{T} \setminus \mathsf{Sent}_{\mathcal{L}}$.

The above definition is actually a scheme. We define separately for all standard n what it means for an expanded structure to satisfy a theory whose axioms have depth bounded by n. Note that whenever $(\mathcal{M}, T)$ is an expansion of a full model $\mathcal{M} \models \mathsf{B}$ and $\mathcal{T}$ extends B with finitely many standard sentences $\phi_1, \ldots, \phi_n$, the condition $(\mathcal{M}, T) \models \mathcal{T}$ means simply that $(\mathcal{M}, T) \models \phi_i$ for $i \leq n$.

2.5. Polynomial simulations and feasible reductions. Let us fix $\mathcal{T} \supseteq \mathsf{PA}$ such that $\mathcal{T}$ conservatively extends PA. It turns out that in order to verify that $\mathcal{T}$ is feasibly reducible to PA it is sufficient to demonstrate that the formalized conservativity statements for finite fragments of $\mathcal{T}$ over sufficiently large finite fragments of PA are feasibly provable in PA. Theorem 2.36 below makes this precise in a very general manner.[10]

- Note that we continue to identify a theory $\mathcal{T}$ with its set of axioms, and not with the deductive closure of its axioms.

We begin with a definition that will allow us to state our results in a succinct manner.

DEFINITION 2.35. Let $\{\psi_n\}_{n\in\mathbb{N}}$ be a family of arithmetical formulae.

- $\{\psi_n\}_{n\in\mathbb{N}}$ is said to be *polynomially* PA-*provable* if there is a polynomial $p(n)$ such that for each $n \in \mathbb{N}$, $\mathsf{PA} \vdash^{p(n)} \psi_n$.
- $\{\psi_n\}_{n\in\mathbb{N}}$ is said to be *feasibly* PA-*provable*, if there is a P-time computable function $f(s)$ such that for each $n \in \mathbb{N}$, $f(\mathsf{tal}(n))$ is a PA-proof of ψ_n.

The first part of Theorem 2.36 provides a sufficient condition for an extension $\mathcal{T}$ of PA to by polynomially simulated by PA; the "moreover" part, in turn, provides a sufficient condition for $\mathcal{T}$ to be feasibly reducible to PA. Theorem 2.36 below is

[10] We are grateful to Fedor Pakhomov who pointed out to us that this is the most direct way of proving our main results. Our previous proofs employed the conceptually more transparent—but technically more demanding—framework of feasible interpretations, as explained in Section 4.4. The results in this section can be readily generalized by replacing PA with a theory $\mathcal{T}_1$ extending PA (in the same language) with a P-time computable set of axioms, and replacing $\mathcal{T}$ with an extension $\mathcal{T}_2$ of $\mathcal{T}_1$.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

used to derive all the other results in this section. Note that the only results of this section that will be explicitly invoked in the remaining sections of this article are Corollary 2.40 and Observation 2.42, when $\mathcal{T}$ is chosen as CT^-, KF^-, and FS^-.

In the statement of the theorem below, recall from Definition 2.23 that $\mathcal{T} \restriction n$ refers to the sentences of $\mathcal{T}$ whose length is at most n.

THEOREM 2.36. *Let $\mathcal{T}$ be an NP-time computable theory extending* PA*, and suppose that there is a polynomial $q(n)$ such that the family $\{\psi_n\}_{n\in\mathbb{N}}$ is polynomially* PA*-provable, where*:

$$\psi_n := \forall \phi \in \mathsf{Sent}_{\mathcal{L}_{\mathsf{PA}}} \left(\big(\mathsf{dp}(\phi) \leq \underline{n} \wedge \mathsf{Pr}_{\mathcal{T}\restriction \underline{n}}(\phi)\big) \rightarrow \mathsf{Pr}_{\mathsf{PA}\restriction \underline{q(n)}}(\phi) \right).$$

Then PA *polynomially simulates $\mathcal{T}$. Moreover, if $\mathcal{T}$ is P-time computable and $\{\psi_n\}_{n\in\mathbb{N}}$ is feasibly* PA*-provable, then $\mathcal{T}$ is feasibly reducible to* PA.

Recall that $\mathsf{dp}(\phi)$ is the height of the syntactic tree of ϕ and that we use this symbol for the arithmetic formula representing this function. The proof of Theorem 2.36 will be facilitated by the following lemma which shows that PA is *feasibly strongly reflexive*.

LEMMA 2.37. *There is a P-time computable function $f(s_0, s_1)$ such that for every $n, k \in \mathbb{N}$, $f(\mathsf{tal}(n), \mathsf{tal}(k))$ is a* PA*-proof of*

$$\forall \phi \in \mathsf{Sent}_{\mathcal{L}_{\mathsf{PA}}} \big(\big(\mathsf{dp}(\phi) \leq \underline{k} \wedge \mathsf{Pr}_{\mathsf{PA}\restriction \underline{n}}(\phi)\big) \rightarrow \mathsf{Tr}_k(\phi) \big).$$

PROOF OF LEMMA 2.37. The proof follows the usual pattern, but we have to check that each transformation at work is feasible. Here we provide the general outline; the details are verified carefully in Section 6.1 of the Appendix, where the propositions invoked in the proof are presented. Assume first that $n \leq k$. Working in PA, we first prove cut-elimination for First-Order Logic (this is a single sentence independent of n). Then we show that every axiom of PA of length $\leq n$ is true. For finitely many axioms of Robinson's Q, this is done independently of n. For induction axioms of length at most n we use Proposition 6.4, and for logical axioms we use Proposition 6.3. Next we apply cut-elimination over First-Order Logic to show that for a sentence ϕ of depth $\leq k$ if $\mathsf{Pr}_{\mathsf{PA}\restriction n}(\phi)$, then there is a cut-free proof of a sequent

$$\Gamma \longrightarrow \phi,$$

where Γ contains only axioms of $\mathsf{PA} \restriction n$. By the subformula property, in such a proof every formula is of depth bounded by k. Then, using induction on the number of proof lines in a proof using only formulae of depth at most k, we show that if

$$\Gamma \longrightarrow \Delta$$

is provable, then we have

$$\forall \alpha \big(\alpha \in \mathsf{Asn}(\Gamma \cup \Delta) \wedge \forall x \in \Gamma \ \ \mathsf{Sat}_k(x, \alpha) \rightarrow \exists y \in \Delta \ \ \mathsf{Sat}_k(y, \alpha)\big),$$

where $\alpha \in \mathsf{Asn}(\Gamma \cup \Delta)$ abbreviates $\forall x \in \Gamma \cup \Delta \ \ \alpha \in \mathsf{Asn}(x)$, as in Definition 2.3. Since we already know that all axioms of $\mathsf{PA} \restriction n$ are true, we conclude that ϕ is true.

If $k \leq n$, then it is sufficient to carry out the above proof substituting n for k everywhere and use Observation 2.31. Note that all the transformations above are uniform in n, k, hence in particular they give rise to a P-time computable function f as claimed by the lemma. ⊣

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

Proof of Theorem 2.36. Assume $\mathcal{T} \vdash^n \phi$, where ϕ is an $\mathcal{L}_{\mathsf{PA}}$-formula. Then clearly the length and, consequently, the depth of ϕ is at most n, and there is a $\mathcal{T} \restriction n$-proof π of ϕ of length at most n. Let k be the code for π. By the properties of the provability predicate and Theorem 2.20 there exists a polynomial $p(x, y, z)$ such that whenever π, k, n, ϕ are as above, then

$$\mathsf{PA} \vdash^{p(|\underline{n}|,|\underline{k}|,|\underline{\phi}|)} \mathsf{Proof}_{\mathcal{T}\restriction \underline{n}}(\underline{k}, \underline{\phi}).$$

Since $|\underline{k}|$ and $|\underline{\phi}|$ are bounded above by a polynomial n, $p(|\underline{n}|, |\underline{k}|, |\underline{\phi}|)$ is bounded above by a polynomial $r_1(n)$, and thus we conclude that $\mathsf{PA} \vdash^{r_1(n)} \mathsf{Pr}_{\mathcal{T}\restriction \underline{n}}(\underline{\phi})$. It is also clear that there is a polynomial $r_2(n)$ such that PA proves $\mathsf{dp}(\underline{\phi}) \leq \underline{n}$ via a proof of length at most $r_2(n)$.

Therefore by invoking the assumption that ψ_n is PA-provable via a proof of length at most $p(n)$ for some polynomial $p(n)$, there is a polynomial $r_3(n)$ such that there is a PA-proof of length at most $r_3(n)$ of $\mathsf{Pr}_{\mathsf{PA}\restriction \underline{q(n)}}(\underline{\phi})$. Coupled with Lemma 2.37, this shows that there is a PA-proof of $\mathsf{Tr}_{q(n)}(\underline{\phi})$ whose length is at most $r_4(n)$ for some polynomial $r_4(n)$. So by feasibly provable T-biconditionals (Theorem 2.29) we finally obtain a polynomial $r_5(n)$ such that there is a PA-proof ϕ of length at most $r_5(n)$. This completes the proof that PA polynomially simulates $\mathcal{T}$.

To prove the "moreover" part, assume that $\mathcal{T}$ is P-time computable. Then with the help of part (2) of Corollary 2.24 and Lemma 2.37, the NP-time computable transformation implicitly described in the above proof of polynomial simulation of $\mathcal{T}$ by PA can be turned into an explicit P-time computable function f that feasibly transforms any $\mathcal{T}$-proof π of ϕ into a PA-proof $f(\pi)$ of ϕ. ⊣

Corollary 2.38. *Let $\mathcal{T}$ be an NP-time computable theory extending* PA. *Suppose that there is a $k \in \mathbb{N}$ and a polynomial $q(n)$ such that the family $\{\theta_n\}_{n\in\mathbb{N}}$ is* PA*-polynomially provable, where θ_n expresses*:

"Every Δ_2-full model $\mathcal{M}$ of PA $\restriction q(n)$ *has an elementary extension to a Δ_k-full model $\mathcal{N}$ which can be expanded to a Δ_k-full model of $\mathcal{T} \restriction n$".*

Then PA *polynomially simulates $\mathcal{T}$. Moreover, if $\mathcal{T}$ is P-time computable and $\{\theta_n\}_{n\in\mathbb{N}}$ is* PA*-feasibly provable, then $\mathcal{T}$ is feasibly reducible to* PA.

Let us make one remark before we proceed to the proof of Corollary 2.38. Recall from Section 2.2 that in the current article we treat full models as specific arithmetically definable sets of sentences. *Note that although we cannot quantify over models in general, we can do this for models of fixed quantifier complexity using arithmetical partial satisfaction predicates.*

Proof of Corollary 2.38. Fix an NP-computable $\mathcal{T}$ extending PA and let ψ_n be as in Theorem 2.36. We will show that the assumption of Corollary 2.38 about $\mathcal{T}$ implies the assumption of Theorem 2.36 about $\mathcal{T}$ by informally describing a PA-proof π of ψ_n whose length is bounded by a polynomial in n. Our proof will make it clear that if $\mathcal{T}$ is P-time computable, then π can be feasibly computed from the input $\mathsf{tal}(n)$, thereby establishing the "moreover" part of Corollary 2.38.

Let k, $p(n)$, $q(n)$, be such that for all $n \in \mathbb{N}$,

$$\mathsf{PA} \vdash^{p(n)} \text{"Every } \Delta_2\text{-full model } \mathcal{M} \text{ of } \mathsf{PA} \restriction \underline{q(n)} \text{ has an elementary extension } \mathcal{N} \text{ which can be expanded to } \Delta_k\text{-full model of } \mathcal{T} \restriction \underline{n}\text{"}.$$

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

Recall that θ_n denotes the above sentence in quotation marks. Fix n. Start the PA-proof by proving θ_n using a subproof of length $p(n)$. Arguing in PA, fix a formula ϕ of depth $\leq n$ and assume:

$$\neg \mathsf{Pr}_{\mathsf{PA}\restriction \underline{q(n)}}(\phi).$$

It immediately follows that $\mathsf{PA} \restriction \underline{q(n)} + \neg\phi$ is consistent. We verify that this is a Δ_1-theory (the length of this PA-verification is bounded above by a polynomial in the definition of $\mathsf{PA} \restriction \underline{q(n)} + \neg\phi$, hence it is bounded above by a polynomial in n).[11] Then we prove the Arithmetized Completeness Theorem (Theorem 2.12) for Δ_1-theories; this is independent of n and gives us a Δ_2-full model $\mathcal{M}$ of $\mathsf{PA} \restriction \underline{q(n)} + \neg\phi$. Now, by θ_n, this model has an elementary extension to a full model $\mathcal{N}$ which can be expanded to a full model $\mathcal{N}^+$ of $\mathcal{T} \restriction \underline{n}$. By elementarity, $\mathcal{T} \restriction \underline{n} + \neg\phi$ holds in $\mathcal{N}^+$, and is therefore consistent. This, in turn, gives us:

$$\neg \mathsf{Pr}_{\mathcal{T}\restriction \underline{n}}(\phi),$$

which completes the informal description of a PA-proof of ψ_n whose length is bounded above by a polynomial in n. ⊣

The next result is a strengthening of Corollary 2.38; note that as opposed to the sentence θ_n of Corollary 2.38, the sentence θ'_n of Corollary 2.39 does not include the demand that the Δ_k-model of $\mathcal{T} \restriction n$ be a *full* model.

Corollary 2.39. *Let $\mathcal{T}$ be an NP-time computable theory extending* PA, *and suppose that there exist $k \in \mathbb{N}$ and a polynomial $q(n)$ such that the family $\{\theta'_n\}_{n\in\mathbb{N}}$ is* PA*-polynomially provable, where θ'_n expresses*:

"Every Δ_2-full model $\mathcal{M}$ of $\mathsf{PA} \restriction q(n)$ has an elementary extension to a Δ_k-full model $\mathcal{N}$ which can be expanded to a Δ_k-model of $\mathcal{T} \restriction n$."

Then PA *polynomially simulates $\mathcal{T}$. Moreover, if $\mathcal{T}$ is P-time computable and $\{\theta'_n\}_{n\in\mathbb{N}}$ is* PA*-feasibly provable, then $\mathcal{T}$ is feasibly reducible to* PA.

Proof. Fix an NP-time computable theory $\mathcal{T}$ extending PA, and let k, $q(n)$, and θ'_n be as in the assumptions of Corollary 2.39. We will informally describe a PA-proof π of θ_n whose length is bounded by a polynomial in n. Moreover, if $\mathcal{T}$ is P-time computable, it will be routine to verify that π can be feasibly computed from the input $\mathsf{tal}(n)$, thereby establishing the "moreover" part of Corollary 2.39.

Work in PA. Fix an arbitrary Δ_2-full model $\mathcal{M}$ of $\mathsf{PA} \restriction \underline{q(n)}$. Then there exists a Δ_k-full model $\mathcal{N}$ elementarily extending $\mathcal{M}$ and there is an expansion $\mathcal{N}^+$ of $\mathcal{N}$ such that $\mathcal{N}^+ \models \mathcal{T} \restriction \underline{n}$. We claim that the Δ_2-theory

$$\Phi := \mathsf{ElDiag}(\mathcal{M}) \cup \mathcal{T} \restriction \underline{n}$$

is consistent. Thanks to Corollary 2.38, this will finish the proof, since by ACT (Arithmetized Completeness Theorem) we will get a Δ_3-full model of this theory (the length of this subproof is polynomial in n as the proof of ACT is independent of n).

To verify our claim that Φ is consistent, take an arbitrary proof π of a sentence ϕ in Φ, prove cut-elimination for first-order logic (the length of this proof is independent

[11] Note that ϕ doesn't contribute to the length of this verification at all. It is a variable and the length of the verification is estimated from the outside.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

of n), and conclude that there exists a proof π' of ϕ in Φ with the subformula property. It follows that every formula in this proof is either an arithmetical formula or is a subformula of an additional axiom of $\mathcal{T} \restriction \underline{n}$. In particular the depth of the nonarithmetical formulae which occur in this proof is bounded by n. Define:

$$\mathsf{Sat}(x, y) := x \in \mathsf{Form}_{\mathcal{L}_{\mathsf{PA}}} \wedge \mathcal{M} \models x[y] \vee \big(x \notin \mathsf{Form}_{\mathcal{L}_{\mathsf{PA}}} \wedge \mathsf{Sat}_n^{\mathcal{N}^+}(x, y)\big),$$

where $\mathsf{Sat}_n^{\mathcal{N}^+}(x, y)$ is a feasible relativized truth predicate from Corollary 2.32. By induction on the length of π' show that if $\Gamma \longrightarrow \Delta$ occurs in π', then for every α we have:

$$\big(\forall x \in \Gamma \ \ \mathsf{Sat}(x, \alpha)\big) \to \big(\exists x \in \Delta \ \ \mathsf{Sat}(x, \alpha)\big).$$

It follows that π' cannot be a proof of the empty sequent, hence Φ is consistent. ⊣

COROLLARY 2.40. *Suppose that $\mathcal{T}$ is a finite extension of* PA *of the form* $\mathsf{PA} + \phi$ *and $k \in \mathbb{N}$. Assume that* $\mathsf{PA} \vdash \psi$, *where ψ is the sentence expressing*:

"If B *is any finite fragment of* PA, *then every Δ_2-full model $\mathcal{M}$ of* B *has an elementary extension to a Δ_k-full model $\mathcal{N}$ which has an expansion to a Δ_k-model of* $\mathsf{B} + \phi$*".*

Then $\mathcal{T}$ is feasibly reducible to PA.

PROOF. This is an immediate consequence of the "moreover" clause of Corollary 2.39 with $q(n) := n$, since given n, the additional PA-verification that $\mathsf{PA} \restriction n$ is a finite (i.e., coded) fragment of PA can be done feasibly in the input $\mathsf{tal}(n)$. ⊣

We will close this subsection with two simple observations which may be obtained by inspection of the proof of Theorem 2.36 and the proof of subsequent corollaries. They provide slightly different sufficient conditions for feasible reducibility. Observation 2.42 will be useful in Section 4.3.

OBSERVATION 2.41. *Let $\mathcal{T}$ be a P-time theory. Suppose that $\mathcal{T}$* is PA*-provably feasibly strongly reflexive, i.e., there exists a P-time computable function $h(s_0, s_1)$ such that for each $n, k \in \mathbb{N}$, $h(\mathsf{tal}(n), \mathsf{tal}(k))$ is a* PA*-proof of*:

$$\forall \phi \in \mathsf{Sent}_{\mathcal{L}_{\mathsf{PA}}}\big(\mathsf{dp}(\phi) \leq \underline{k} \wedge \mathsf{Pr}_{\mathcal{T} \restriction \underline{n}}(\phi) \to \mathsf{Tr}_k(\phi)\big). \qquad (*)$$

Then $\mathcal{T}$ is feasibly reducible to PA,

OBSERVATION 2.42. *Suppose that $\mathcal{T}$ satisfies the assumptions of Corollary* 2.39 *(with the "moreover" part) or Corollary* 2.40. *Then $\mathcal{T}$ is* PA*-provably feasibly strongly reflexive*.

PROOF. By (very direct) inspection of the proofs of Corollaries 2.39 and 2.40, we see that if $\mathcal{T}$ satisfies the assumptions of any of these statements, then there exist a a polynomial $q(n)$ and a P-time computable function $f(s)$ such that for each $n \in \mathbb{N}$, $f(\mathsf{tal}(n))$ is a PA-proof of:

$$\forall \phi \in \mathsf{Sent}_{\mathcal{L}_{\mathsf{PA}}}\left(\mathsf{dp}(\phi) \leq \underline{n} \wedge \mathsf{Pr}_{\mathcal{T} \restriction \underline{n}}(\phi) \to \mathsf{Pr}_{\mathsf{PA} \restriction \underline{q(n)}}(\phi)\right). \qquad (1)$$

It follows that $\mathcal{T}$ is PA-provably feasibly strongly reflexive. Indeed, fix the above mentioned polynomial $q(n)$, and function $f(s)$, and let $g(s_0, s_1)$ be a function witnessing the feasible strong reflexivity of PA.[12] The desired function $h(\mathsf{tal}(n), \mathsf{tal}(k))$ can be

[12] Recall that this was the property introduced in Lemma 2.37.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

now be informally described as follows: First compute $f(\mathsf{tal}(n))$, then compute $g(\mathsf{tal}(q(n)), \mathsf{tal}(k))$, i.e., the proof of

$$\forall \phi \in \mathsf{Sent}_{\mathcal{L}_{\mathsf{PA}}}\big(\mathsf{dp}(\phi) \leq \underline{k} \wedge \mathsf{Pr}_{\mathsf{PA}\restriction q(n)}(\phi) \rightarrow \mathsf{Tr}_k(\phi)\big).$$

Finally, the proof of (*) is then readily obtained after performing a fixed number of steps depending on whether $n > k$ or $n \leq k$. ⊣

2.6. Feasible interpretability and speed-up. This section presents certain refinements of the notion of interpretability that will be only used in Section 4.4. The rudiments of interpetability theory can be found in Chapter III of [10].

We begin with an observation of Albert Visser, found in the proof of Proposition 6.4 of [8].[13]

THEOREM 2.43. *If there is a polynomial interpretation* (*see Definition* 2.44 *below*) *of* CT^- *in* PA, *then* PA *polynomially simulates* CT^- *with respect to* Π_1*-sentences.*

A key ingredient in the proof of the above theorem is that any interpretation of CT^- in PA is Π_1-correct, i.e., for every Π_1-sentence ϕ of arithmetic, we have:

$$\mathsf{CT}^- \vdash \phi^I \rightarrow \phi,$$

where I is an interpretation of CT^- in PA. The definitions below provide the conceptual tools for establishing that PA polynomially simulates CT^- for *all* arithmetical sentences, thereby generalizing Theorem 2.43.

- Since interpretations are given by translation maps, the notation $I : \mathcal{T}_2 \rightarrow \mathcal{T}_1$ abbreviates the assertion that the translation map I yields an interpretation of $\mathcal{T}_2$ in $\mathcal{T}_1$.

DEFINITION 2.44. Let $\mathcal{T}_1$, $\mathcal{T}_2$ be two theories each of which has an NP-time computable set of axioms, with $\mathcal{T}_1 \subseteq \mathcal{T}_2$ (for the applications in this article, $\mathcal{T}_1$ is PA, and $\mathcal{T}_2$ is any of the truth theories CT^-, KF^-, FS^- over the base theory PA).

1. Given $n \in \mathbb{N}$, an interpretation $I : \mathcal{T}_2 \rightarrow \mathcal{T}_1$ is *n-correct* if for each $\mathcal{L}_{\mathcal{T}_1}$-formula ϕ of length at most n we have:
$$\mathcal{T}_1 \vdash (\phi^I \rightarrow \phi).$$
2. An interpretation $I : \mathcal{T}_2 \rightarrow \mathcal{T}_1$ is *polynomial*[14] if there is a polynomial $p(n)$ such that for every $\mathcal{L}_{\mathcal{T}_2}$-formula ϕ and every $n \in \mathbb{N}$ we have:
$$\mathcal{T}_2 \vdash^n \phi \Rightarrow \mathcal{T}_1 \vdash^{p(n)} \phi^I.$$
3. A family of interpretations $\{I_n\}_{n\in\mathbb{N}} : \mathcal{T}_2 \rightarrow \mathcal{T}_1$ is *polynomially neat* if it is uniformly polynomial and uniformly n-correct for each $n \in \mathbb{N}$, more precisely, if the following two conditions hold:

[13]The proof presented for Proposition 6.4 of [8] only establishes Theorem 2.43, and not the stronger version of Theorem 2.43 that asserts that PA polynomially simulates CT^- with respect to Π_1-sentences, since the proof presented conflates the notions of interpretability and polynomial interpretability. Note that there are plenty of interpretations that are not polynomial interpretations, e.g., let $s\mathsf{PA} := \mathsf{PA} + \{\mathsf{Con}_{\mathsf{PA}}(2_{\underline{n}}) \mid n \in \mathbb{N}\}$. Then the identity interpretation witnesses the relative interpretability of $s\mathsf{PA}$ in PA, but in light of Theorem 2.26 and the remark following it, the former theory has super-exponential speed-up over the latter for Π_1-sentences.

[14]Polynomial interpretations in the sense of this definition are called "feasible interpretations" in Verbrugge's doctoral thesis [21]. As shown in Theorem 6.4.2 of [21], there is a sentence θ such that $\mathsf{PA} + \theta$ is interpretable in PA, and yet there is no polynomial interpretation of $\mathsf{PA} + \theta$ in PA.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

(a) There is polynomial $p(k)$ such that for every $k \in \mathbb{N}$ and every $\mathcal{L}_{\mathcal{T}_1}$-formula ϕ of length at most k,
$$\mathcal{T}_1 \vdash^{p(k)} (\phi^{I_k} \to \phi).$$
(b) There is a polynomial $q(n,k)$ that witnesses that I_k is a polynomial interpretation of $\mathcal{T}_2$ in $\mathcal{T}_1$, i.e., for every $k, n \in \mathbb{N}$, and for every $\mathcal{L}_{\mathcal{T}_2}$-formula ϕ,
$$\mathcal{T}_2 \vdash^n \phi \Rightarrow \mathcal{T}_1 \vdash^{q(n,k)} \phi^{I_k}.$$

4. $I : \mathcal{T}_2 \to \mathcal{T}_1$ is a *feasible interpretation* if there is a P-time computable function $f(s)$ such that for all $\mathcal{L}_{\mathcal{T}_2}$-formulae ϕ, if π is a $\mathcal{T}_2$-proof of ϕ, then $f(\pi)$ is $\mathcal{T}_1$-proof of ϕ^I.
5. A family of interpretations $\{I_n\}_{n\in\mathbb{N}} : \mathcal{T}_2 \to \mathcal{T}_1$ is *feasibly neat* if there exist P-time computable functions $f(s_0, s_1)$ and $g(s_0, s_1)$ such that the following two conditions hold:
   (a) For every $k \in \mathbb{N}$, and every $\mathcal{L}_{\mathcal{T}_1}$-formula ϕ of length at most k, $g(\mathsf{tal}(k), \phi)$ is a $\mathcal{T}_1$-proof of:
   $$\phi^{I_k} \to \phi.$$
   (b) For every $k \in \mathbb{N}$, and every $\mathcal{T}_2$-proof π of an $\mathcal{L}_{\mathcal{T}_2}$-formula ϕ, $f(\mathsf{tal}(k), \pi)$ is a $\mathcal{T}_1$-proof of ϕ^{I_k}.

REMARK 2.45. In the context of theories with no additional rules of reasoning, condition (b) in the definition of polynomially neat interpretations (item 3 above) can equivalently be replaced with the following one:

(b)′ For every $k \in \mathbb{N}$,
$$\mathcal{T}_1 \vdash^{q(n,k)} \phi^{I_k}$$
for every axiom ϕ of $\mathcal{T}_2$ (including the logical axioms for $\mathcal{L}_{\mathcal{T}_2}$) of length at most n.

(Condition (b) of the definition of feasibly neat interpretations also lends itself to an analogous reformulation.) However, we prefer (b) over (b)′ as it can be used in the context of theories such as FS^- which is closed under two additional rules of reasoning: NEC and CONEC.

By unravelling the relevant definitions we obtain the following useful proposition that provides sufficient conditions for polynomial simulability and feasible reducibility which are conceptually different from the ones presented in Section 2.5.

PROPOSITION 2.46. *If there exists a polynomially neat family of interpretations* $\{I_n\}_{n\in\mathbb{N}} : \mathcal{T}_2 \to \mathcal{T}_1$, *then* $\mathcal{T}_1$ *polynomially simulates* $\mathcal{T}_2$. *Moreover, if* $\{I_n\}_{n\in\mathbb{N}}$ *is a feasibly neat family of interpretations, then* $\mathcal{T}_2$ *is feasibly reducible to* $\mathcal{T}_1$.

PROOF. Fix a polynomially neat family of interpretations $\{I_n\}_{n\in\mathbb{N}}$ and let $p(n)$ and $q(n,k)$ be the pair of polynomials witnessing this. Suppose that $\mathcal{T}_2 \vdash^n \phi$ for some $\mathcal{L}_{\mathcal{T}_1}$-formula ϕ. Then clearly the length of ϕ is at most n. Therefore by condition (a) of polynomial neatness we have:
$$\mathcal{T}_1 \vdash^{p(n)} \phi^{I_n} \to \phi.$$

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

On the other hand, condition (b) of polynomial neatness implies:

$$\mathcal{T}_1 \vdash^{q(n,n)} \phi^{I_n}.$$

This makes it evident that $\mathcal{T}_1 \vdash^{O(p(n)+q(n,n)+n)} \phi$, thus completing the proof of polynomial simulation of $\mathcal{T}_2$ by $\mathcal{T}_1$. The proof of the "moreover" part is fully analogous. ⊣

§3. Dramatis personæ: typed and untyped theories of truth. In this section, B denotes a "base theory" for a theory of truth, i.e., a theory with a modicum of arithmetic capable of handling syntax. For example, any theory extending $\mathsf{I}\Delta_0 + \mathsf{Exp}$ will do. T denotes a fresh unary predicate that is not in the language of B. $\mathcal{L}_{\mathsf{B}}$ denotes the language of B and $\mathcal{L}_T$ denotes the language of B enriched with the predicate T. For simplicity assume that the signature of $\mathcal{L}_{\mathsf{B}}$ extends the arithmetical signature with finitely many relational symbols.

In this article, we will be dealing with theories of truth conservative over their base theories. We say that a theory $\mathcal{T}$ in the language $\mathcal{L}_{\mathcal{T}}$ is conservative over $\mathsf{B} \subseteq \mathcal{T}$ if for every sentence $\phi \in \mathcal{L}_{\mathsf{B}}$ we have:

$$\mathsf{B} \vdash \phi \text{ iff } \mathcal{T} \vdash \phi.$$

In our case, this means that adding the truth predicate and some axioms governing its behaviour does not allow us to prove new arithmetical sentences.

Below, we discuss some prominent examples of truth theories. The standard reference to the subject is Halbach's book [11].

3.1. CT^-**.**

DEFINITION 3.1. $\mathsf{CT}^-[\mathsf{B}]$ is the theory extending a theory B with the following axioms:

- $\mathsf{CT1}$. $\forall s, t \in \mathsf{ClTerm}_{\mathcal{L}_{\mathsf{B}}}\ T(s = t) \equiv (s^\circ = t^\circ)$.
- $\mathsf{CT2}$. $\forall s_1, \dots, s_n \in \mathsf{ClTerm}_{\mathcal{L}_{\mathsf{B}}}\ T\big(R(s_1, \dots, s_n)\big) \equiv R(s_1{}^\circ, \dots, s_n{}^\circ)$, for every relation symbol R of $\mathcal{L}_{\mathsf{B}}$.
- $\mathsf{CT3}$. $\forall \phi, \psi \in \mathsf{Sent}_{\mathcal{L}_{\mathsf{B}}}\ T(\phi \vee \psi) \equiv T(\phi) \vee T(\psi)$.
- $\mathsf{CT4}$. $\forall \phi \in \mathsf{Sent}_{\mathcal{L}_{\mathsf{B}}}\ T(\neg\phi) \equiv \neg T(\phi)$.
- $\mathsf{CT5}$. $\forall \phi \in \mathsf{Form}^{\leq 1}_{\mathcal{L}_{\mathsf{B}}} \forall v \in \mathsf{Var}\ T(\exists v\ \phi) \equiv \exists x T(\phi(\underline{x}))$.
- $\mathsf{CT6}$. $\forall \phi(\bar{x}) \in \mathsf{Form}_{\mathcal{L}_{\mathsf{B}}} \forall \bar{s}, \bar{t} \in \mathsf{ClTermSeq}_{\mathcal{L}_{\mathsf{B}}}\ \big(\bar{s^\circ} = \bar{t^\circ} \to T(\phi[\bar{s}/\bar{x}]) \equiv T(\phi[\bar{t}/\bar{x}])\big)$.

The last condition is sometimes called *generalized regularity*, or *generalized term-extensionality*. It resembles the well-known extensionality rule from deductive calculi for first-order logic, i.e.,

$$\frac{s_0 = t_0, \dots, s_k = t_k, \phi(s_0, \dots, s_k)}{\phi(t_0, \dots, t_k)}.$$

We include it since without it the quantifier axiom for CT^- behaves in an unnatural way.[15] For example, for $\mathsf{B} = \mathsf{PA}$ we have:

$$\mathsf{PA} + \big(\mathsf{CT1} \wedge \mathsf{CT2} \wedge \mathsf{CT3} \wedge \mathsf{CT4} \wedge \mathsf{CT5}\big) \nvdash T(\forall x\ \phi(x)) \to \forall t (T\phi(t)).$$

[15] It behaves decently already after adding the ungeneralized version of CT6 for single terms.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

Obviously one can simply interchange the quantifier axiom with the following:

$$T(\exists y\ \phi(y)) \equiv \exists t \in \mathsf{ClTerm}_{\mathcal{L}_{\mathsf{PA}}} T(\phi(t)).$$

But then, without regularity, the following implication becomes unprovable:

$$\forall x T(\phi(\underline{x})) \rightarrow T(\forall y\ \phi(y)).$$

With regularity both quantifier axioms are easily seen to be equivalent.

The above version of $\mathsf{CT}^-[\mathsf{PA}]$ was claimed to be conservative over PA in [22]. However, no proof of this fact was provided and only a hint was provided to the effect that it requires a slight modification of the Enayat–Visser construction (see [6]). This modification, however, adds a layer of technical difficulty, so in the current version we prove feasible conservativity of $\mathsf{CT}^-[\mathsf{PA}]$ in full detail. A detailed proof of the conservativity of this theory is provided also in [13].

3.2. KF^- **and** FS^-**.** The idea behind the untyped notion of truth is that the truth predicate can be meaningfully applied also to sentences containing it, to the effect that we could e.g., judge

$$T(\ulcorner 0 = 0 \urcorner)$$

to be true. For this reason untyped truth predicates are also referred to as *self-applicative*. In this setting the following additional axiom seems desirable:

$$\forall s \in \mathsf{ClTerm}_{\mathcal{L}_{\mathsf{B}}} \forall \phi \in \mathsf{Sent}_{\mathcal{L}_T}\ \big(s^\circ = \phi \rightarrow T(T(s)) \equiv T(\phi)\big), \tag{TRP}$$

where "TRP" abbreviates "TRansParency". Obviously if one wants to have a compositional theory of self-applicable truth, one cannot simply take (TRP), the axioms CT1 through CT6, and let the quantifiers range over all formulae of $\mathcal{L}_T$, since the resulting theory would be inconsistent by Tarski's Theorem. The next two truth theories which we shall investigate exhibit two different directions that the search for a natural theory of untyped truth leads to. In the first one (Kripke–Feferman) the axiom CT3 is rejected and somewhat compensated. In the second one (Friedman–Sheard) the transparency axiom is missing.

DEFINITION 3.2. $\mathsf{KF}^-[\mathsf{B}]$ is the $\mathcal{L}_T$-theory extending B with the following axioms:

KF1. $\forall s, t \in \mathsf{ClTerm}_{\mathcal{L}_{\mathsf{B}}}\ T(s = t) \equiv (s^\circ = t^\circ)$.
KF2. $\forall s, t \in \mathsf{ClTerm}_{\mathcal{L}_{\mathsf{B}}}\ T(s \neq t) \equiv (s^\circ \neq t^\circ)$.
KF3. $\forall s_1, \ldots, s_n \in \mathsf{ClTerm}_{\mathcal{L}_{\mathsf{B}}}\ T\big(R(s_1, \ldots, s_n)\big) \equiv R(s_1{}^\circ, \ldots, s_n{}^\circ)$, for every relation symbol R of $\mathcal{L}_{\mathsf{B}}$.
KF4. $\forall s_1, \ldots, s_n \in \mathsf{ClTerm}_{\mathcal{L}_{\mathsf{B}}}\ T\big(\neg R(s_1, \ldots, s_n)\big) \equiv \neg R(s_1{}^\circ, \ldots, s_n{}^\circ)$, for every relation symbol R of $\mathcal{L}_{\mathsf{B}}$.
KF5. $\forall \phi \in \mathsf{Sent}_{\mathcal{L}_T}\ T(\neg\neg\phi) \equiv T(\phi)$.
KF6. $\forall \phi, \psi \in \mathsf{Sent}_{\mathcal{L}_T}\ T(\phi \vee \psi) \equiv T(\phi) \vee T(\psi)$.
KF7. $\forall \phi, \psi \in \mathsf{Sent}_{\mathcal{L}_T}\ T(\neg(\phi \vee \psi)) \equiv T(\neg\phi) \wedge T(\neg\psi)$.
KF8. $\forall y \in \mathsf{Var}\, \forall \phi \in \mathsf{Form}^{\leq 1}_{\mathcal{L}_T}\ T(\exists y\ \phi(y)) \equiv \exists x T(\phi(\underline{x}))$.
KF9. $\forall v \in \mathsf{Var}\, \forall \phi \in \mathsf{Form}^{\leq 1}_{\mathcal{L}_T}\ T(\neg\exists v\ \phi(v)) \equiv \forall x T(\neg\phi(\underline{x}))$.
KF10. $\forall \bar{s}, \bar{t} \in \mathsf{ClTermSeq}_{\mathcal{L}_{\mathsf{B}}} \forall \phi(\bar{x}) \in \mathsf{Form}_{\mathcal{L}_T}\ \Big(\bar{s}^\circ = \bar{t}^\circ \rightarrow T(\phi(\bar{s})) \equiv T(\phi(\bar{t}))\Big)$.
KF11. $\forall \phi \in \mathsf{Sent}_{\mathcal{L}_T}\, \forall t \in \mathsf{Term}_{\mathcal{L}_{\mathsf{B}}}\ \Big(t^\circ = \phi \rightarrow T(T(t)) \equiv T(\phi)\Big)$.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

KF12. $\forall \phi \in \mathsf{Sent}_{\mathcal{L}_T}\ \forall t \in \mathsf{Term}_{\mathcal{L}_\mathsf{B}}\ \Big(t^\circ = \phi \to T(\neg T(t)) \equiv T(\neg\phi)\Big)$.

KF, a theory obtained by augmenting $\mathsf{KF}^-[\mathsf{PA}]$ with the full induction scheme for formulae with the truth predicate, was introduced by Feferman in [7] as an axiomatization of a theory of truth proposed by Kripke in [15].

KF^- represents an attempt to define a reasonably behaved self-applicable truth predicate guided by the following intuition: we try to mark the sentences which are definitely true. We start with the set of true arithmetical equations. Then we proceed in stages, e.g., whenever ϕ and ψ are definitely true, we mark $\phi \wedge \psi$ as definitely true. Whenever ϕ is definitely true, we mark $T(\phi)$ as definitely true. Whenever $\neg\phi(\underline{x})$ is definitely true for all x, we mark $\neg\exists x \phi(\underline{x})$ as definitely true.

Thus in the process we only enlarge the set of true sentences until it reaches a fixed point. KF^- axiomatizes properties of fixed points obtained in such a way.

The desirable feature of $\mathsf{KF}^-[\mathsf{B}]$ is that it satisfies the TRP axiom. However, the idempotence of the truth predicate fails rather spectacularly in a different place. It turns out that adding both derivation rules:

$$\frac{\phi}{T(\phi)}\ (\mathrm{NEC}) \qquad\qquad \frac{T(\phi)}{\phi}\ (\mathrm{CONEC})$$

to $\mathsf{KF}^-[\mathsf{B}]$ at the same time yields this theory inconsistent, as indicated by Lemma 15.20 of [11] (the Lemma is stated for the full KF, but the induction axioms are not used in the proof). Moreover, the rule (NEC) is inconsistent with the following axiom of consistency which says that no sentence is both true and false:

$$\forall \phi \in \mathsf{Sent}_{\mathcal{L}_T}\ \neg\big(T(\phi) \wedge T(\neg\phi)\big)\,.$$

Dually, the rule (CONEC) is inconsistent with the axiom of completeness which states that every sentence is either true or false.

The other standard candidate for a well-behaved theory of self-referential truth is Friedman–Sheard's theory FS. Note that FS comes equipped with two extra rules of inference NEC and CONEC.

DEFINITION 3.3. $\mathsf{FS}^-[\mathsf{B}]$ is the $\mathcal{L}_T$-theory extending B with the following axioms, and with the extra rules of inference NEC and CONEC.

FS1. $\forall s, t \in \mathsf{ClTerm}_{\mathcal{L}_\mathsf{B}}\ \ T(s = t) \equiv \big(s^\circ = t^\circ\big)$.

FS2. $\forall s_1, \ldots, s_n \in \mathsf{ClTerm}_{\mathcal{L}_\mathsf{B}}\ \ T\big(R(s_1, \ldots, s_n)\big) \equiv R(s_1{}^\circ, \ldots, s_n{}^\circ)$, for every relation symbol R of $\mathcal{L}_\mathsf{B}$.

FS3. $\forall \phi \in \mathsf{Sent}_{\mathcal{L}_T}\ \ T(\neg\phi) \equiv \neg T(\phi)$.

FS4. $\forall \phi, \psi \in \mathsf{Sent}_{\mathcal{L}_T}\ \ T(\phi \vee \psi) \equiv T(\phi) \vee T(\psi)$.

FS5. $\forall v \in \mathsf{Var}\ \forall \phi \in \mathsf{Form}^{\leq 1}_{\mathcal{L}_T}\ \ T(\exists v\ \phi) \equiv \exists x T(\phi(\underline{x}))$.

FS6. $\forall \bar{s}, \bar{t} \in \mathsf{ClTermSeq}_{\mathcal{L}_\mathsf{B}} \forall \phi(\bar{x}) \in \mathsf{Form}_{\mathcal{L}_T}\ \Big(\bar{s}^\circ = \bar{t}^\circ \to T(\phi(\bar{s})) \equiv T(\phi(\bar{t}))\Big)$.

Note that in none of the above theories we extend the induction scheme to the full $\mathcal{L}_T$. As usual we write simply FS^- to abbreviate $\mathsf{FS}^-[\mathsf{PA}]$.

A set of axioms which is deductively equivalent to the above was first introduced in [9]. The above list of axioms is taken from [11] with a minor variation: we supplemented the normal axiomatization with FS6 for reasons analogous to the ones for CT^-.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

At first sight, $\mathsf{FS}^-[\mathsf{B}]$ seems to be much more natural than $\mathsf{KF}^-[\mathsf{B}]$. The presence of NEC and CONEC rules somewhat compensates for the lack of the transparency axiom making the theory *symmetric*: for every $\phi \in \mathcal{L}_T$ it holds that

$$\mathsf{FS}^-[\mathsf{B}] \vdash \phi \text{ iff } \mathsf{FS}^-[\mathsf{B}] \vdash T(\phi).$$

This heavily contrasts with the case of $\mathsf{KF}^-[\mathsf{B}]$. However, this symmetric feature turns out to be very pricey, as McGee's well-known theorem shows

THEOREM 3.4 (McGee, [20]). $\mathsf{FS}^-[\mathsf{B}]$ *is* ω*-inconsistent*.

Moreover, the fully inductive versions of both theories differ dramatically in strength, when evaluated over PA: KF[PA] can define ε_0 levels of the ramified truth hierarchy (i.e., $\mathsf{RT}_{<\alpha}$ for every $\alpha < \varepsilon_0$. See [11] for details), while the strength of FS[PA] is exhausted by ω-many such levels.

Both $\mathsf{KF}^-[\mathsf{B}]$ and $\mathsf{FS}^-[\mathsf{B}]$ are conservative extensions of B, as indicated by the following theorems, the proofs of which will be sketched in the next Subsection.

THEOREM 3.5 (Cantini, [2]). $\mathsf{KF}^-[\mathsf{B}]$ *is a conservative extension of* B.

The above theorem has been proved by Cantini for PA, but his proof works essentially in the same way for all base theories B with a modicum of arithmetic. Conservativity of FS^- follows from the work of Halbach. He showed that FS with full induction is reducible to the system $\mathsf{RT}_{<\omega}$ with full induction and a stratified family of compositional truth predicates. His proof, however, does not rely on induction in the considered theories or on the specific choice of the base theory. Therefore, essentially the same argument shows that $\mathsf{FS}^-[\mathsf{B}]$ is reducible to $\mathsf{RT}^-_{<\omega}[\mathsf{B}]$ for a wide choice of base theories B. Conservativity of $\mathsf{RT}^-_{<\omega}[\mathsf{B}]$ can in turn be shown by using known proofs of conservativity for CT^-, so, in a sense, it was "in the air."[16] We will provide more details (including the definition of $\mathsf{RT}^-_{<\omega}$) in Section 3.3.3.

THEOREM 3.6 (Essentially due to Halbach). $\mathsf{FS}^-[\mathsf{B}]$ *is a conservative extension of* B.

3.3. Conservativity of truth theories. The main goal of this article is to establish that certain truth theories over PA are feasibly reducible to PA. This involves certain elaborate technical arguments in each case. However, what these proofs have in common is that they all rely on the results from Section 2.5 since they follow the same general pattern: Suppose that $\mathcal{T}$ is a theory of truth over PA that is conservative over PA. Moreover, assume that the conservativity proof in fact can be formalized in PA and that it is uniform in the sense that the proof works equally well for PA and its large enough finitely axiomatized fragments B containing $\mathsf{I}\Delta_0 + \mathsf{Exp}$. Then $\mathcal{T}$ can be shown to be feasibly reducible to PA. Let us recall the precise formulation of this fact (it was formulated as Corollary 2.40):

Suppose that $\mathcal{T}$ *is a finite extension of* PA *of the form* $\mathsf{PA} + \phi$ *and* $k \in \mathbb{N}$. *Assume that* $\mathsf{PA} \vdash \psi$, *where* ψ *is the sentence expressing*:

"If B *is any finite fragment of* PA, *then every* Δ_2*-full model* $\mathcal{M}$ *of* B *has an elementary extension to a* Δ_k*-full model* $\mathcal{N}$ *which has an expansion to a* Δ_k*-model of* $\mathsf{B} + \phi$."

[16] However, we know of no published proof of this result.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

Then $\mathcal{T}$ is feasibly reducible to PA.

The proofs of our feasible reducibility results will in each case consist of an appropriate arithmetization in PA of a known conservativity proof of $\mathcal{T}$ over fragments of PA. Therefore, we are forced to pay close attention to the specific features of the arithmetical implementation of the conservativity proofs, which is bound to obscure the main idea of the proof of feasible reduction. Therefore, to provide some help to the reader, we present outlines of the relevant conservativity proofs in this section.

3.3.1. Conservativity of CT^-. Fix any fragment B of PA extending $\mathsf{I}\Delta_0 + \mathsf{Exp}$. In this section, we sketch the proof of the conservativity of $\mathsf{CT}^-[\mathsf{B}]$ over B. We will base our proof on the model-theoretic argument given by Enayat and Visser in [6].

By the completeness theorem for first-order logic, it suffices to show that every model $\mathcal{M}$ of B has an elementary extension that expands to a model of $\mathsf{CT}^-[\mathsf{B}]$. To this end, fix a model $\mathcal{M}$ of B. We will construct $(\mathcal{N}, T) \models \mathsf{CT}^-[\mathsf{B}]$, where $\mathcal{M} \preceq \mathcal{N}$, by first constructing an ω- chain of models:

$$(\mathcal{M}_0, \varnothing) \subseteq (\mathcal{M}_1, S_1) \subseteq (\mathcal{M}_2, S_2) \subseteq \cdots$$

such that $\mathcal{M}_0 = \mathcal{M}$, $\mathcal{M}_i \preceq \mathcal{M}_{i+1}$, and the subsets S_i are partially defined satisfaction predicates in the sense *that each S_{i+1} is only required to satisfy compositional conditions for formulae from $\mathcal{M}_i$, but the valuations are allowed to come from $\mathcal{M}_{i+1}$*. For example, we require that if $\phi \vee \psi$ is an arithmetical formula from $\mathcal{M}_i$, then:

$$(\mathcal{M}_{i+1}, S_{i+1}) \models \forall \alpha \in \mathsf{Asn}(\phi, \psi)\ \ S_{i+1}(\phi \vee \psi, \alpha) \equiv S_{i+1}(\phi, \alpha) \vee S_{i+1}(\psi, \alpha),$$

and that if $\exists v \phi$ is an arithmetical formula from $\mathcal{M}_i$, then:

$$(\mathcal{M}_{i+1}, S_{i+1}) \models \forall \alpha \in \mathsf{Asn}(\exists v \phi)\ \ S_{i+1}(\exists v \phi, \alpha) \equiv \exists \alpha' \sim_v \alpha\ \ S_{i+1}(\phi, \alpha').$$

To recap: We demand that S_{i+1} behaves compositionally for formulae belonging to $\mathcal{M}_i$, including nonstandard ones, but all valuations from $\mathcal{M}_{i+1}$ are allowed. We also require that S_{i+1} agrees with S_i on formulae from $\mathcal{M}_{i-1}$. Note that if ϕ is in $\mathcal{M}_i$, then a direct subformula of ϕ is also in $\mathcal{M}_i$. Finally, we require that each S_{i+1} is *extensional* on formulae from $\mathcal{M}_i$ for all valuations in $\mathcal{M}_{i+1}$, i.e., if ϕ and ψ are arithmetical formulae in $\mathcal{M}_i$, and α and β are valuations in $\mathcal{M}_{i+1}$ such that $\phi[\alpha] = \psi[\beta]$ (in the notation of Definition 2.4), then $(\phi, \alpha) \in S_{i+1}$ iff $(\psi, \beta) \in S_{i+1}$.[17]

To build the $(i+1)$-st member $(\mathcal{M}_{i+1}, S_{i+1})$ of the ω-chain, suppose we have already built $(\mathcal{M}_i, S_i)$. We will formulate a particular set Γ of sentences (called *the Enayat–Visser theory* in Section 4.1) which has the property that any model of Γ can serve as $(\mathcal{M}_{i+1}, S_{i+1})$. Γ is formulated in the language obtained by augmenting the language of arithmetic with a new predicate S[18] as well as constants for each element from $\mathcal{M}_i$. Γ consists of (1) the elementary diagram of $\mathcal{M}_i$, (2) sentences

[17] Note that CT6 guarantees that the satisfaction predicate S_T induced by the truth predicate T is extensional, where $(\phi, \alpha) \in S_T$ iff $\phi[\alpha] \in T$.

[18] The language of Γ in the proof presented in [6] does not employ the binary relation symbol S, but rather, the family of unary relation symbols U_ϕ, as ϕ ranges over the $\mathcal{L}_{\mathsf{B}}$-formulae in the sense of $\mathcal{M}$. Each such predicate defines a set of satisfying assignments for a specific ϕ in $\mathcal{M}$. The two approaches are readily seen to be equivalent, since the language of Γ has access to constants for each element in the universe of discourse of $\mathcal{M}$. In Lemma 4.3 of Section 4.1 we stick to the use of U_ϕ's, which we find conceptually simpler.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

stipulating compositional conditions for S in a pointwise manner (i.e., one formula at a time), (3) sentences ensuring in a pointwise manner that S extends S_i, and (4) sentences asserting in a pointwise manner that S is extensional. By compactness, Γ has a model if for each finite subset Γ_0 of Γ, there is a subset S^* of $\mathcal{M}_i$ such that $(\mathcal{M}_i, S^*)$ satisfies Γ_0. As shown in [6], when the language of arithmetic is purely relational (i.e., when addition and multiplication are construed as ternary relations) this turns out to be possible with a straightforward recursion using an appropriate notion of "rank" for arithmetical formulae of $\mathcal{M}_i$ that appear in Γ_0. In the presence of function symbols in the language of arithmetic a slightly fancier rank function can be introduced to get the job done, as demonstrated by Cieśliński [3]. The proof we will present in Section 4.1 will employ a more complicated rank function than the one used by Cieśliński since we need to handle the generalized regularity axiom $\mathsf{CT6}$, which is included among the axioms of CT^- in Section 4.1, but is not included in the axioms of CT^- in [3].

Having built the desired ω-chain, let S^*_{i+1} consist of $(\phi, \alpha) \in S_{i+1}$ such that ϕ is in $\mathcal{M}_i$, and consider the union $(\mathcal{N}, S)$ of models $(\mathcal{M}_{i+1}, S^*_{i+1})$ for $i \in \omega$. In order to check that the resulting union satisfies the compositional axioms (for all formulae and assignments), we take an arbitrary formula ϕ, its direct subformulae, and some fixed valuation α for ϕ. We check that it satisfied the compositional conditions in the model $(\mathcal{M}_{i+1}, S^*_{i+1})$, where ϕ and its direct subformulae were present in $\mathcal{M}_i$, and that the compositional conditions were preserved along the construction.

Finally, we obtain the desired model $(\mathcal{N}, T)$ of CT^- by defining $T \subsetneq N$ as follows:

$$\phi \in T \equiv \big(\phi \in \mathsf{Sent}_{\mathcal{L}_{\mathsf{PA}}}(\mathcal{N}) \wedge (\phi, \varnothing) \in S\big).$$

This concludes the sketch of the proof. A detailed argument will be presented in Section 4.1.

The above proof does not overtly formalize in PA. The obstacle is as follows: when we speak in PA of full models $(\mathcal{M}_i, S_i)$, we really speak of formulae defining elementary diagrams of $(\mathcal{M}_i, S_i)$. The defining formulae for the full models can in general be more and more complex as we iterate the construction, so even though each standard initial segment of the chain is overtly definable, there might be no formally correct way of defining the whole chain.

There are a couple of ways to circumvent the obstacle. One is to use the low basis version of the arithmetical completeness theorem; this method was used in the privately circulated 2012 manuscript of Enayat and Visser. The route undertaken in this article is the simplest we know of : we build the ω-chain indirectly by a detour through appropriate first-order theories. More specifically, we will show, reasoning in PA, that for any natural number x, the theory $\mathcal{T}_x$ (formulated in an extension of the language of B with finitely many new predicate symbols) is consistent, where $\mathcal{T}_x$ says:

"There is a *finite* chain of models $(\mathcal{M}_0, S_0) \subseteq (\mathcal{M}_1, S_1) \subseteq \cdots \subseteq (\mathcal{M}_x, S_x)$ satisfying the conditions from the Enayat-Visser construction."

This will be done by formalizing the inductive step in the Enayat–Visser construction, i.e., by showing that for all numbers x, if $\mathcal{T}_x$ is consistent, then $\mathcal{T}_{x+1}$ is consistent as well. The consistency of $\mathcal{T}_x$ is a Π_1-statement, so PA will be able to verify that for any x the theory $\mathcal{T}_x$ is consistent. This in turn will be enough to show that the theory $\mathcal{T}_\omega$ is consistent, and hence has a model, where $\mathcal{T}_\omega$ says:

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

"There is an *infinite* chain of models $(\mathcal{M}_0, S_0) \subseteq (\mathcal{M}_1, S_1) \subseteq \cdots$ satisfying the conditions from the Enayat-Visser construction."

From a model of $\mathcal{T}_\omega$ we will be able to define the whole chain in a uniform way and, consequently, its sum, which will give us a model of $\mathsf{CT}^-[\mathsf{B}]$ (not a full model though). The details involve a number of intricate and technical considerations; they are presented in the next section.

3.3.2. Conservativity of KF^-. In this subsection we will outline the proof of conservativity of $\mathsf{KF}^-[\mathsf{B}]$, where B is a fragment of PA extending $\mathsf{I}\Delta_0 + \mathsf{Exp}$. Our proof of conservativity resembles Cantini's variation [2] of Kripke's original fixed point argument [15]. A similar construction appears in [4].

We can construct a truth predicate over $\mathbb{N}$ as a fixed point of an operator that takes a subset T_α of $\mathbb{N}$, thought of as the set of sentences (possibly containing the truth predicate) which can be already identified as true at a given stage of the construction, and replaces it with $T_{\alpha+1} \supseteq T_\alpha$ in the following way:

- If ϕ is a true atomic or negated atomic formula, then $\phi \in T_{\alpha+1}$.
- If $\phi \in T_\alpha$, then $\phi \in T_{\alpha+1}$.
- If $\phi \in T_\alpha$, and $\phi = t^\circ$ for a term t, then $T(t) \in T_{\alpha+1}$.
- If $\neg\phi \in T_\alpha$, and $(\neg\phi) = t^\circ$, then $\neg T(t) \in T_{\alpha+1}$.
- If $\phi \in T_\alpha$, then $\neg\neg\phi \in T_{\alpha+1}$.
- If $\phi \in T_\alpha$ or $\psi \in T_\alpha$, then $\phi \vee \psi \in T_{\alpha+1}$.
- If $\neg\phi \in T_\alpha$ and $\neg\psi \in T_\alpha$, then $\neg(\phi \vee \psi) \in T_{\alpha+1}$.
- If $\phi(\underline{x}) \in T_\alpha$, then $\exists v\phi(v) \in T_{\alpha+1}$.
- If $\neg\phi(\underline{x}) \in T_\alpha$ for all x, then $\neg\exists v\phi(v) \in T_{\alpha+1}$.

If λ is a limit ordinal, we set $T_\lambda = \bigcup_{\alpha<\lambda} T_\alpha$. In the above construction, we enlarge the set T_α of sentences which are definitely true with a set of sentences which are definitely true if we interpret the truth predicate as the set T_α. Since at each stage, we only keep enlarging our set, the construction will reach its fixed point. This construction yields a truth predicate over the standard model $\mathbb{N}$ of arithmetic, thus yielding a model of KF. Moreover, the method of construction readily carries over to arbitrary models to show that *every* model of B expands to a model of $\mathsf{KF}^-[\mathsf{B}]$, which immediately implies the conservativity of $\mathsf{KF}^-[\mathsf{B}]$ over its base theory. However, the outlined argument relies on the higher order principle: "Every positive operator on subsets of $\mathbb{N}$ reaches a fixed point," which is clearly not available in PA. As it turns out, a rather simple fix to this problem can be formulated as follows.

Given a model $\mathcal{M} \models \mathsf{B}$, let $\mathcal{N}$ be a *recursively saturated* elementary extension of $\mathcal{M}_0$. Notice that for $n \in \omega$, the n-th set obtained in the inductive procedure described above, T_n, is arithmetically definable in $\mathcal{M}$ (let us call the defining formula Θ_n). By definability of T_n and recursive saturation of $\mathcal{M}$, we can deduce that already T_ω is a truth predicate satisfying axioms of $\mathsf{KF}^-[\mathsf{B}]$. Essentially, this relies on the fact that in recursively saturated models, $\phi(\underline{x}) \in T_\omega$ holds for all $x \in M$ if and only if $\phi(\underline{x}) \in T_k$ holds for some $k \in \omega$ and all $x \in M$.[19]

[19] A very similar argument has been presented in [4] in the proof that any recursively saturated model of PA can be expanded to a model of PT^- with internal induction for total formulae. It seems that this reasoning appears originally in [2], where Cantini proved conservativity of KF^- with internal induction for total formulae over PA.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

It turns out that the above argument can be implemented in PA, i.e., within PA we can elementarily extend any full model $\mathcal{M}$ of B to a recursively saturated full model $\mathcal{N}$, and then we can build the predicate T as the union of all sets defined in $\mathcal{N}$ with formulae Θ_n mentioned above. The details will be given in Section 4.2.

3.3.3. Conservativity of FS^-. The proof of conservativity of FS^- over PA is analogous to the one showing the upper bounds on the proof-theoretic strength of its fully inductive version, FS[PA]. As an intermediate step we pass through a theory of iterated compositional truth predicates of length ω, $\mathsf{RT}^-_{<\omega}$.

DEFINITION 3.7. For each $n \in \mathbb{N}$, $\mathsf{RT}^-_{<n+1}[\mathsf{B}]$ is the extension of B in the language $\mathcal{L}_{<n+1}$ extending $\mathcal{L}_{\mathsf{PA}}$ with $n+1$ new predicate symbols $\{T_0, \ldots, T_n\}$ (we stipulate that $\mathcal{L}_{<0} = \mathcal{L}_{\mathsf{PA}}$ and $\mathsf{RT}^-_{<0} = \mathsf{PA}$) satisfying the following axioms for all $k < n+1$:

RT1. $\forall s, t \in \mathsf{ClTerm}_{\mathcal{L}_{\mathsf{PA}}}\ T_k(s = t) \equiv (s^\circ = t^\circ)$.
RT2. $\forall \phi \in \mathsf{Sent}_{\mathcal{L}_{<k}}\ T_k(\neg\phi) \equiv \neg T_k(\phi)$.
RT3. $\forall \phi, \psi \in \mathsf{Sent}_{\mathcal{L}_{<k}}\ T_k(\phi \vee \psi) \equiv T_k(\phi) \vee T_k(\psi)$.
RT4. $\forall \phi(x) \in \mathsf{Form}^{\leq 1}_{\mathcal{L}_{<k}} \forall v \in \mathsf{Var}\ T_k(\exists v \phi) \equiv \exists x\ T_k(\phi(\underline{x}))$.
RT5. $\forall \bar{s}, \bar{t} \in \mathsf{ClTermSeq}_{\mathcal{L}_{\mathsf{PA}}} \forall \phi(x) \in \mathsf{Form}^{\leq 1}_{\mathcal{L}_{<k}}\ \big(\bar{s}^\circ = \bar{t}^\circ \to T_k(\phi(\bar{s})) \equiv T_k(\phi(\bar{t}))\big)$.
RT6. $\bigwedge_{i<k} \forall s \in \mathsf{ClTerm}_{\mathcal{L}_{\mathsf{PA}}}\ \big(s^\circ \in \mathsf{Sent}_{\mathcal{L}_{<i}} \to T_k(T_i(s)) \equiv T_i(s^\circ)\big)$.
RT7. $\forall i < k \forall s \in \mathsf{ClTerm}_{\mathcal{L}_{\mathsf{PA}}}\ \big(s^\circ \in \mathsf{Sent}_{\mathcal{L}_{<i}} \to T_k(T_i(s)) \equiv T_k(T_i(s^\circ))\big)$.

Define $\mathsf{RT}^-_{<\omega}[\mathsf{PA}] := \bigcup_{n\in\omega} \mathsf{RT}^-_{<n}[\mathsf{PA}]$.

REMARK 3.8. We assume that the initially chosen coding is extended in such a way that the length of T_n (the n-th truth predicate) is logarithmic in n (in fact, polynomial will do, so this logarithmic bound is not that important).

As in the case of FS^-, $\mathsf{RT}^-_{<n}$ and $\mathsf{RT}^-_{<\omega}$ abbreviate $\mathsf{RT}^-_{<n}[\mathsf{PA}]$ and $\mathsf{RT}^-_{<\omega}[\mathsf{PA}]$ respectively. Note that similar to other theories studied in this article, in $\mathsf{RT}^-_{<\omega}$ we do not extend the scheme of induction to formulae with the truth predicate.

Now let B be our base theory. We shall demonstrate that the problem of conservativity of $\mathsf{FS}^-[\mathsf{B}]$ over B can be reduced to the analogous problem of conservativity of $\mathsf{RT}^-_{<\omega}[\mathsf{B}]$ over B. Recall that an interpretation I is an *ω-interpretation* if for every arithmetical sentence ϕ we have:

$$\phi^I = \phi.$$

In order to perform such a reduction it suffices to show that every "finite piece" of FS^- can be ω-interpreted in $\mathsf{RT}^-_{<\omega}$. In this context "an n-piece" means "a sentence which can be deduced from B and axioms FS1–FS6 (note that in this context FS2 is missing) using at most n applications of NEC and CONEC rules." We shall denote it with $\mathsf{FS}^-_n[\mathsf{B}]$. Thus ϕ is in $\mathsf{FS}^-_1[\mathsf{B}]$ if it can be deduced using one application of the NEC rule or one application of the CONEC rule. (But not both. Our definition differs from the original one given by Halbach.) Now the following holds:

LEMMA 3.9 (Essentially Halbach, [11], Theorem 14.31). *For each $n \in \mathbb{N}$, $\mathsf{FS}^-_n[\mathsf{B}]$ is ω-interpretable in $\mathsf{RT}^-_{<2n+1}[\mathsf{B}]$.*

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

PROOF. Define a family $\{g_n\}_{n\in\mathbb{N}}$ of primitive recursive functions as follows:

$$g_n(k) = \begin{cases} k & \text{if } k = (s = t), \\ \ulcorner 0 = 1 \urcorner & \text{if } k \notin \mathsf{Sent}_{\mathcal{L}_T} \text{ or } k = T(t) \text{ and } n = 0, \\ T_{n-1}(g_{n-1}(t)) & \text{if } k = T(t) \text{ and } n > 0, \\ \neg g_n(\phi) & \text{if } k = \neg\phi, \\ g_n(\phi) \vee g_n(\psi) & \text{if } k = \phi \vee \psi, \\ \exists x\, g_n(\phi) & \text{if } k = \exists x\, \phi. \end{cases}$$

Where $T_n(g_n(t))$ abbreviates:

$$\forall x \big(g_n(t) = x \to T(x)\big),$$

and $g_n(x) = y$ is a natural Δ_0-formula which represents g_n in $\mathsf{I}\Delta_0 + \mathsf{Exp}$. We shall check that for every n, g_{n+1} is an ω-interpretation of $\mathsf{FS}_n^-[\mathsf{B}]$ in $\mathsf{RT}^-_{<2n+1}[\mathsf{B}]$. It is evident that each g_n acts as the identity function on arithmetical sentences. Moreover, for every $\phi \in \mathcal{L}_T$ and each n, $g_n(\phi)$ is a sentence of $\mathcal{L}_{<n}$ (that is, it contains truth predicates with indices at most $n-1$) and this fact is provable in B. Hence if ϕ is any axiom from FS1 through FS6 and $0 < k \leq n$, then:

$$\mathsf{RT}^-_{<n}[\mathsf{B}] \vdash g_k(\phi). \tag{$*$}$$

Now, following the lines of Halbach's argument, we fix n and then use induction on i, where $0 \leq i \leq n$, we show that for every $i \leq n$ and every $j \in \{i+1, \ldots, 2n+1-i\}$[20] we have:

$$\forall \psi \;\; \mathsf{FS}_i^-[\mathsf{B}] \vdash \psi \Longrightarrow \mathsf{RT}^-_{<2n+1}[\mathsf{B}] \vdash g_j(\psi).$$

Note that $(*)$ witnesses that the above holds for $i = 0$. Now inductively assume that the above holds for some nonzero $i < n$, and fix $j \in \{i+2, \ldots, 2n+1-(i+1)\}$.

Fix a proof π of ψ in $\mathsf{FS}^-_{i+1}[\mathsf{B}]$. Arguing by induction assume that the last rule used in π is either NEC or CONEC. In both cases we will use the fact that for all $k \leq l < m$, and every $\phi \in \mathcal{L}_{<k}$, $\mathsf{RT}^-_{<m}[\mathsf{B}]$ proves:

$$T_l(\phi) \equiv \phi. \tag{$**$}$$

If ψ is obtained by NEC, then $\psi = T(\theta)$ and by our induction assumption we know that $\mathsf{RT}^-_{<2n+1}[\mathsf{B}] \vdash g_{j-1}(\theta)$. Since $g_{j-1}(\theta) \in \mathsf{Sent}_{<j-1}$, by $(**)$ we obtain $\mathsf{RT}^-_{<2n+1}[\mathsf{B}] \vdash T_{j-1}(g_{j-1}(\theta))$. The last sentence is by definition equal to $g_j(T(\theta))$, thus concluding the verification of this case. If ψ is obtained by CONEC, then we argue dually, using g_{j+1} applied to $T(\psi)$. ⊣

In the rest of this section we sketch the proof of conservativity of $\mathsf{RT}^-_{<\omega}[\mathsf{B}]$ over B based on the Enayat–Visser construction. For starters, let us note that it suffices to construct, for an arbitrary model $\mathcal{M} \models \mathsf{B}$, a chain of models $(\mathcal{M}_i)_{i<\omega}$ satisfying the following properties:

1. $\mathcal{M}_0 = \mathcal{M}$;
2. $\mathcal{M}_i \models \mathsf{RT}^-_{<i}$; and
3. $\mathcal{M}_i \preceq_{\mathcal{L}_{<i}} \mathcal{M}_{i+1}$.

Then $\bigcup_{i\in\mathbb{N}} \mathcal{M}_i$ will be an elementary extension of $\mathcal{M}$ satisfying $\mathsf{RT}^-_{<\omega}[\mathsf{B}]$. To get $\mathcal{M}_{i+1}$ we basically start the Enayat–Visser construction (as sketched in Section

[20] The fact that this range of j shrinks in the induction process is needed to deal with CONEC.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

3.3.1) on $\mathcal{M}_i$ for the base language $\mathcal{L}_{<i}$. More precisely, we build an ω-chain of models $(\mathcal{M}_i^j, S_j)_{j\in\mathbb{N}}$ such that:

1. $\mathcal{M}_i^0 = \mathcal{M}_i$ and $S_0 = \varnothing$;
2. $\mathcal{M}_i^j \preceq_{\mathcal{L}_{<i}} \mathcal{M}_i^{j+1}$;
3. $S_j \subseteq S_{j+1}$; and
4. S_{j+1} is a satisfaction class for $\mathsf{Form}_{\mathcal{L}_{<i}}(\mathcal{M}_i^j)$ with respect to all valuations from $\mathcal{M}_i^{j+1}$.

Satisfying the above requirements would suffice to guarantee that in the limit model axioms RT1 through RT6 will hold. However, to account for RT7 we have to improve our satisfaction classes S_j slightly. This can be done by requiring that S_{j+1} makes true all the statements ϕ such that

$$\mathcal{M}_i^j \models T_l(\phi),$$

for $l \leq i$ and $\phi \in \mathsf{Sent}_{\mathcal{L}_{<l}}^{\mathcal{M}_i^j}$ (i.e., $\langle \phi, \alpha\rangle \in S_{j+1}$ for any such ϕ and for every assignment $\alpha \in \mathcal{M}_i^{j+1}$). This, in turn, requires only a tiny modification of the original Enayat–Visser proof. Details will be presented in Section 4.3.

§**4. The main act: feasible reductions of truth theories.** This section contains the principal results of this article. The first three subsections are devoted, respectively, to feasible reductions of $\mathsf{CT}^-[\mathsf{PA}]$, $\mathsf{KF}^-[\mathsf{PA}]$, and $\mathsf{FS}^-[\mathsf{PA}]$ to PA. The last section, on the other hand, presents an interpretability-theoretic perspective of our work.

4.1. Feasible reduction of $\mathsf{CT}^-[\mathsf{PA}]$ to PA. This section is devoted to the proof of the following result:

THEOREM 4.1. $\mathsf{CT}^-[\mathsf{PA}]$ *is feasibly reducible to* PA.

An immediate corollary of Theorem 4.1 is that $\mathsf{CT}^-[\mathsf{PA}]$ does not have super-polynomial speed-up over PA. The proof of a special case of this corollary for Π_1-sentences of arithmetic was presented by Fischer [8], based on an outline suggested by Visser, but as pointed out in a footnote in Section 2.6 the presented proof lacks an important detail.

Our proof of Theorem 4.1 will be based on the verification of the veracity of the assumption of Corollary 2.40 for $k = 4$ and $\mathcal{T} = \mathsf{CT}^-[\mathsf{PA}]$. In fact, for the purposes of our work in Section 4.3, we shall do slightly better and prove a more general result (Lemma 4.3) for which we need the definition below.

DEFINITION 4.2. A theory B is *good* if B extends $\mathsf{I}\Delta_0 + \mathsf{Exp}$ (hence B can serve as a base theory), and B is formulated in a language $\mathcal{L}_{\mathsf{B}}$ that extends $\mathcal{L}_{\mathsf{PA}}$ with at most finitely many new relation symbols (in particular, the terms of a good theory are only the arithmetical terms).

LEMMA 4.3. *For each $l \in \mathbb{N}$ the sentence ϕ_l is provable in* PA, *where ϕ_l is the sentence expressing*: *"If* B *is a good Δ_1-theory, then every Δ_l-full model of* B *has an elementary extension to a Δ_{l+2}-model of* $\mathsf{CT}^-[\mathsf{B}]$."

- In the case of $\mathsf{CT}^-[\mathsf{PA}]$ we will need Lemma 4.3 only for $l = 2$ and for $\mathcal{L}_{\mathsf{B}} = \mathcal{L}_{\mathsf{PA}}$. The more general version will be needed to handle the feasible reduction of $\mathsf{FS}^-[\mathsf{PA}]$ to PA. The proof of Lemma 4.3 will consist of a formalization of

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

the ω-chain Enayat–Visser construction inside PA whose sketch was given in Section 3.3. As in the Enayat–Visser conservativity, we shall make a detour through partial satisfaction classes. We now present the preliminaries needed in the proof of Lemma 4.3; the central of which is Lemma 4.8.

CONVENTION 4.4. If P is an arbitrary unary predicate and $\phi(x)$ an arbitrary formula with one free variable, then we write $\phi \restriction P$ for the formula $\phi(x) \wedge P(x)$.

DEFINITION 4.5 ($\mathsf{CS}^- \restriction P$). Let B be a theory in a finite language $\mathcal{L}_\mathsf{B}$ extending $\mathsf{I}\Sigma_1$ and P be a fresh unary predicate. $\mathsf{CS}^- \restriction P[\mathsf{B}]$ is the theory of *P-restricted, extensional satisfaction class for* $\mathcal{L}_\mathsf{B}$ formulated in the language $\mathcal{L}_S = \mathcal{L}_\mathsf{B} \cup \{S\} \cup \{P\}$ and extending B with the following axioms:

1. $\forall x, y \big(S(x,y) \to x \in \mathsf{Form}_{\mathcal{L}_\mathsf{B}} \restriction P \wedge y \in \mathsf{Asn}(x)\big)$.
2. $\forall s_0 \ldots \forall s_n \in \mathsf{Term}_{\mathcal{L}_\mathsf{B}} \restriction P \forall \alpha \in \mathsf{Asn}(s_0, \ldots, s_n) \big(S(R(s_0, \ldots, s_n), \alpha) \equiv R(s_0^\alpha, \ldots, s_n^\alpha)\big)$, where R is a relation symbol in $\mathcal{L}_\mathsf{B}$.
3. $\forall \phi, \psi \in \mathsf{Form}_{\mathcal{L}_\mathsf{B}} \restriction P \, \forall \alpha \in \mathsf{Asn}(\phi, \psi) \ \big(S(\phi \vee \psi, \alpha) \equiv S(\phi, \alpha) \vee S(\psi, \alpha)\big)$.
4. $\forall \phi \in \mathsf{Form}_{\mathcal{L}_\mathsf{B}} \restriction P \, \forall \alpha \in \mathsf{Asn}(\phi) \ \big(S(\neg\phi, \alpha) \equiv \neg S(\phi, \alpha)\big)$.
5. $\forall \phi \in \mathsf{Form}_{\mathcal{L}_\mathsf{B}} \restriction P \ \forall v \in \mathsf{Var} \restriction P \ \forall \alpha \in \mathsf{Asn}(\exists v \phi) \ \big(S(\exists v \phi, \alpha) \equiv \exists \beta \sim_v \alpha, \beta \in \mathsf{Asn}(\phi) \ \ S(\phi, \beta)\big)$,
6. Generalized regularity limited to P:

$$\forall \phi \in \mathsf{Form}_{\mathcal{L}_\mathsf{B}} \restriction P \, \forall \vec{v} \in \mathsf{FVSeq}(\phi) \restriction P \, \forall \vec{s}, \vec{t} \in \mathsf{ClTermSeq}_{\mathcal{L}_\mathsf{B}} \restriction P$$
$$\forall \alpha \in \mathsf{Asn}(\phi[\vec{s}/\vec{v}], \phi[\vec{t}/\vec{v}]) \, \big(\vec{s}^\alpha = \vec{t}^\alpha \to \big(S(\phi[\vec{s}/\vec{v}], \alpha) \equiv S(\phi[\vec{t}/\vec{v}], \alpha)\big)\big) \, .$$

Note that in the above definition we do not restrict the range of assignments (denoted by variable α in the definition). In effect, we do not assume that the assignments come from the restricted set. This is crucial for our purposes.

DEFINITION 4.6 (P-restricted extensional satisfaction class). If $\mathcal{M} \models \mathsf{B}$ and $P \subseteq M$, $S \subseteq M^2$ is such that $(\mathcal{M}, S, P) \models \mathsf{CS}^- \restriction P[\mathsf{B}]$, then S is called a *P-restricted extensional satisfaction class for* $\mathcal{L}_\mathsf{B}$ *on* $\mathcal{M}$. If S is "$x = x$"-restricted, it is called *full*. Note that this definition is meaningful even if $(\mathcal{M}, S, P)$ is not a full model since $\mathsf{CS}^- \restriction P[\mathsf{B}]$ is a finite extension of B (recall Definition 2.34).

CONVENTION 4.7. Below we always assume that P is either empty or defines in $\mathcal{M}$ a universe of an elementary submodel of $\mathcal{M}$. Under this assumption, P is closed under the direct subformula relation, which we denote with $\lhd$. More precisely:

$$(\mathcal{M}, P) \models \forall \phi, \psi \in \mathsf{Form}_{\mathcal{L}_\mathsf{B}} \Big(\big(P(\phi) \wedge \psi \lhd \phi\big) \to P(\psi) \Big).$$

The distinctive feature of the Enayat–Visser technique of building truth classes is that one creates a well-behaved *satisfaction* class via a union of chain argument. The chain, in turn, is recursively constructed by applying a key lemma countably many times. Let us now state and prove the arithmetized version of this key lemma. We call the reader's attention to the asymmetry in the above lemma: we start with a model $(\mathcal{M}, S, P)$ but finish with a model $\mathcal{N}$ and two of its subsets S^* and M. This will be compensated for in our recursive construction.

LEMMA 4.8 (Arithmetized Enayat–Visser Lemma). *The universal generalization of the formula $\theta_1 \to \theta_2$ is provable in* PA *for every $l \in \mathbb{N}$, where θ_1 expresses*:

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

"B *is a good theory*; $(\mathcal{M}, S, P)$ *is a* Δ_l*-full model, where* $\mathcal{M}$ *is a full* $\mathcal{L}_{\mathsf{B}}$*-model of* B*; and* S *is a* P*-restricted extensional satisfaction class for* $\mathcal{L}_{\mathsf{B}}$*."*

and θ_2 *expresses*:

"There exist a Δ_{l+1}*-full* $\mathcal{L}_{\mathsf{B}}$*-model* $\mathcal{N}$*, and a* Δ_{l+1}*-set* $S^* \subseteq N^2$ *such that* $\mathcal{M} \preceq \mathcal{N}$*;* S^* *is an* M*-restricted extensional satisfaction class for* $\mathcal{L}_{\mathsf{B}}$*; and* $S \subseteq S^*$*."*

Proof. We work in PA. Let $(\mathcal{M}, S, P)$ be as in θ_1. We follow the lines of the standard Enayat–Visser proof from [6], but we perform it inside PA. Moreover, we have the additional technical complication caused by adding the generalized regularity axiom to $\mathsf{CT}^-[\mathsf{B}]$. Let us define the language $\mathcal{L}_{\mathsf{EV}}$:

$$\mathcal{L}_{\mathsf{EV}} = \mathcal{L}_{\mathsf{B}} \cup \{c \mid c \in M\} \cup \{U_\phi(x) \mid \phi \in \mathsf{Form}_{\mathcal{L}_{\mathsf{B}}}(\mathcal{M})\}.$$

Next, we define *the Enayat–Visser theory* Γ for $(\mathcal{M}, S, P)$ as the union of the following sets of $\mathcal{L}_{\mathsf{EV}}$-formulae:

$\{\phi(a_1, \ldots, a_n) \mid \phi \in \mathcal{L}_{\mathsf{B}}, a_i \in M, \mathcal{M} \models \phi(a_1, \ldots, a_n)\}$ $(= \mathsf{ElDiag}(\mathcal{M}))$,

$\{U_\phi(\alpha) \mid \langle \phi, \alpha \rangle \in S, \phi \in P\}$,

$\{\forall \alpha \in \mathsf{Asn}(s_0, \ldots, s_n) \ \ (U_{R(s_0,\ldots,s_n)}(\alpha) \equiv R(s_0^\alpha, \ldots, s_n^\alpha)) \mid R \in \mathcal{L}_{\mathsf{B}},$
$s_0, \ldots, s_n \in \mathsf{Term}_{\mathcal{L}_{\mathsf{B}}}(\mathcal{M})\}$,

$\{\forall \alpha \in \mathsf{Asn}(\psi) \ \ (U_\psi(\alpha) \equiv U_\phi(\alpha) \vee U_\theta(\alpha)) \mid \mathcal{M} \models \psi \in \mathsf{Sent}_{\mathcal{L}_{\mathsf{B}}} \wedge (\psi = (\phi \vee \theta))\}$,

$\{\forall \alpha \in \mathsf{Asn}(\psi) \ \ (U_\psi(\alpha) \equiv \neg U_\phi(\alpha)) \mid \mathcal{M} \models \psi \in \mathsf{Sent}_{\mathcal{L}_{\mathsf{B}}} \wedge \psi = (\neg\phi)\}$,

$\{\forall \alpha \in \mathsf{Asn}(\psi) \ \ (U_\psi(\alpha) \equiv \exists \beta \sim_v \alpha \ \ U_\phi(\beta)) \mid \mathcal{M} \models \psi \in \mathsf{Sent}_{\mathcal{L}_{\mathsf{B}}} \wedge \psi = (\exists v \phi)\}$,

$\{\forall \alpha, \beta \in \mathsf{Asn}(\psi, s, t) \ \ (\bar{s}^\alpha = \bar{t}^\beta \rightarrow (U_\psi(\alpha) \equiv U_\phi(\beta))) \mid$
$\mathcal{M} \models \bar{s}, \bar{t} \in \mathsf{TermSeq}_{\mathcal{L}_{\mathsf{B}}} \wedge \exists \theta \in \mathsf{Form}_{\mathcal{L}_{\mathsf{B}}} \exists \bar{v} \in \mathsf{FVSeq}(\theta) (\psi = \theta[\bar{s}/\bar{v}] \wedge \phi = \theta[\bar{t}/\bar{v}])\}$.

- Note that once we verify that Γ is consistent, then by ACT (Theorem 2.12) there exists a Δ_{l+1}-(full) model $\mathcal{N}^+$ of Γ, where

$$\mathcal{N}^+ = (\mathcal{N}, c, U_\phi)_{c \in M, \ \phi \in \mathsf{Form}_{\mathcal{L}_{\mathsf{B}}}(\mathcal{M})}.$$

Then, putting:

$$S^* = \{\langle x, y \rangle \mid x \in \mathsf{Form}_{\mathcal{L}_{\mathsf{B}}}(\mathcal{M}) \wedge \mathcal{N} \models U_x(y)\},$$

we can easily check that $(\mathcal{N}, S^*)$ satisfies the properties claimed by θ_2. *The rest of the proof will therefore concentrate on demonstrating the consistency of* Γ.

To prove the consistency of Γ, by compactness it suffices to verify that every finite fragment F of Γ (in the sense of PA) has a model. For each predicate U_ϕ which occurs in F we will find a formula $\theta_\phi(x) \in \mathcal{L}_{\mathsf{B}}$ such that

$$(\mathcal{M}, S, P) \models F[\theta_\phi / U_\phi]_{U_\phi \in F},$$

where $F[\theta_\phi / U_\phi]_{U_\phi \in F}$ denotes the theory resulting from F by replacing each occurrence of U_ϕ with the corresponding formula θ_ϕ. Note that the above makes perfect sense, since $(\mathcal{M}, S, P)$ is a full model. This clearly would guarantee that F is consistent. Moreover from now on we do not need to bother with the sentences from $\mathsf{ElDiag}(\mathcal{M})$, since they obviously hold in $\mathcal{M}$.

As in the original Enayat–Visser proof we will construct θ_ϕ for $\phi \in F$ by recursion on the appropriately defined rank. Note that we have more work to do here than

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

in the proofs given by Enayat and Visser [6] (since in their set-up, the language of arithmetic is purely relational), and by Cieśliński [3] (since in his set-up CT^- does not include our generalized regularity axiom CT6). The bulk of the remaining part of the proof will be devoted to the development of a rank function which allows us to handle CT6.

Let c be the set of formulae ϕ such that the predicate U_ϕ occurs in a formula in F. Let b be an arbitrary coded set of formulae of $\mathcal{L}_{\mathsf{B}}$. We put $\mathsf{rank}^b(\phi) \geq x$ iff there exists a sequence y such that the following three conditions hold (in the last condition $\lhd$ denotes the relation of being an immediate subformula):

1. $\mathsf{len}(y) = x + 1$ and $(y)_x = \{\phi\}$.
2. For all $i < x + 1$ $(y)_i \subseteq b$.
3. For all $i < x$ for all $\theta, \theta \in (y)_{i+1}$ iff for all ψ such that $\mathcal{M} \models \psi \lhd \theta$, $\psi \in (y)_i$.

We say that $\mathsf{rank}^b(\phi) = x$ if x is the greatest y such that $\mathsf{rank}^b(\phi) \geq y$. This definition makes sense, since if $\mathsf{rank}^b(\phi) \geq x$, then $x \leq \mathsf{card}(b)$ (where $\mathsf{card}(b)$ denotes the cardinality of b). For example, if $b = \{0 = 0, 0 = 0 \vee 1 = 1\}$, then the $\mathsf{rank}^b(0 = 0 \vee 1 = 1) = 0$, since $1 = 1 \notin b$.

The intuition behind the above definition is that $\mathsf{rank}^c(\phi)$ is the complexity of ϕ, where any formula ϕ such that at least one immediate subformula of ϕ does not belong to c is treated as an atom. The idea is that for any formula ϕ of rank zero, the defining set $\theta_\phi(x)$ for U_ϕ can be chosen as the formula $(\phi, x) \in S$ if $\phi \in P$; and if $\phi \notin P$, then we choose the formula $x \neq x$ (or some other arbitrary formula) as $\theta_\phi(x)$. Then an obvious recursion can be used to define U_ψ's for formulae ψ of higher rank in terms of the definitions of $U_{\psi'}$ for formulae ψ' of lower rank than ψ.

Note that if we follow the recursive procedure described above, then all the compositional axioms (i.e., counterparts of axioms for atomic formulae, disjunction, negation and quantifier) from F will be satisfied. However, we have one immediate problem: it can happen that (an instance of) the axiom of regularity for ϕ and ψ is in F, but ϕ and ψ get different ranks. In such a situation the standard procedure does not seem to guarantee that θ_ϕ and θ_ψ (i.e., formulae which interpret U_ϕ and U_ψ in $(\mathcal{M}, S, P)$) will satisfy the regularity axiom. To simplify the notation let us define $\phi \approx_F \psi$ if the following is in F:

$$\forall \alpha \in \mathsf{Asn}(\psi, \bar{s}) \forall \beta \in \mathsf{Asn}(\phi, \bar{t}) \quad \left(\bar{s}^\alpha = \bar{t}^\beta \rightarrow \left(U_\psi(\alpha) \equiv U_\phi(\beta)\right)\right).$$

Note that if F includes the instance of the axiom of regularity for ϕ and ψ then the following holds in $\mathcal{M}$ (this follows by the definition of Γ):

$$\bar{s}, \bar{t} \in \mathsf{TermSeq}_{\mathcal{L}_{\mathsf{B}}} \wedge \exists \theta \in \mathsf{Form}_{\mathcal{L}_{\mathsf{B}}} \exists \bar{v} \in \mathsf{FVSeq}(\theta) \quad (\psi = \theta[\bar{s}/\bar{v}] \wedge \phi = \theta[\bar{t}/\bar{v}]).$$

A solution to our puzzle is to complete c, obtaining $\widehat{c}$, to assure that we have for all ϕ, ψ

$$\phi \approx_F \psi \Rightarrow \mathsf{rank}^{\widehat{c}}(\phi) = \mathsf{rank}^{\widehat{c}}(\psi).$$

It is convenient to extend $\approx_F$ a little bit to make it an equivalence relation. We say that ξ is the *term trivialization* of ϕ, and write $\xi = \widehat{\phi}$ if the following four conditions hold:

1. For every occurrence t of a term in ξ, if all occurrences of variables in t are free, then t is a free occurrence of a variable.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

2. No variable occurs in ξ as both bounded and free, and no variable occurs as free more than once.
3. For some ρ, a function with domain $\mathsf{FV}(\theta)$ and values in $\mathsf{Term}_{\mathcal{L}_\mathsf{B}}$, the equality $\xi[\rho] = \phi$ holds. (In this context $\xi[\rho]$ denotes the result of a formal substitution of terms for free variables of ξ, where $\mathsf{Term}_{\mathcal{L}_\mathsf{B}}$ contains also terms with free variables).
4. The indices of free variables of ξ are chosen in a canonical way (for example according to the tree-ordering of the syntactical tree of ξ. This is only needed to guarantee uniqueness).

The idea behind $\widehat{\phi}$ is that if for some term substitution ρ and some formula ψ we have

$$\phi[\rho] = \psi,$$

then, $\widehat{\phi} = \widehat{\psi}$ and there are unique term substitutions γ_1, γ_2 such that:

$$\widehat{\phi}[\gamma_1] = \phi \text{ and } \widehat{\phi}[\gamma_2] = \psi.$$

We write $\phi \approx^{\mathcal{M}} \psi$ if $\mathcal{M} \models \widehat{\phi} = \widehat{\psi}$.[21] Obviously $\approx^{\mathcal{M}}$ is an equivalence relation. Moreover, $\approx^{\mathcal{M}}$ is a congruence with respect to the direct subformula relation $\lhd$, i.e., the following lemma holds. For its proof consult the appendix.

LEMMA 4.9 (Congruence Lemma). *For all ϕ, ϕ', ψ' the following holds*:

$$\big(\phi \lhd \phi' \wedge \phi' \approx^{\mathcal{M}} \psi'\big) \Rightarrow \exists \psi \ \ \big(\psi \lhd \psi' \wedge \psi \approx^{\mathcal{M}} \phi\big). \tag{C}$$

By induction it follows that the congruence lemma holds for $\lhd_a$ in place of $\lhd$, where $\psi \lhd_a \psi'$ means that for some $j \leq a$ and some formulae $\psi = \phi_0, \ldots, \phi_j = \psi'$, $\phi_i \lhd \phi_{i+1}$ holds for each $i < j$.

Finally, observe that for every ϕ, $U_{\widehat{\phi}}$ and U_ϕ are mutually interdefinable. Indeed, fix ϕ and $\gamma : \mathsf{FV}(\widehat{\phi}) \to \mathsf{Term}_{\mathcal{L}_\mathsf{B}}$ such that $\widehat{\phi}[\gamma] = \phi$. Then, having $U_{\widehat{\phi}}$, we define U_ϕ with the condition:

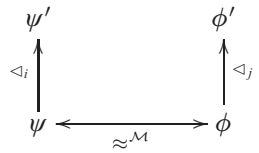
$$\alpha \in U_\phi \iff \exists \beta \in U_{\widehat{\phi}} \, \forall v \in \mathsf{FV}(\widehat{\phi}) \ \ \beta(v) = \gamma(v)^\alpha. \tag{$U_{\widehat{\phi}} \to U_\phi$}$$

Similarly, having U_ϕ we define $U_{\widehat{\phi}}$ with the condition:

$$\beta \in U_{\widehat{\phi}} \iff \exists \alpha \in U_\phi \, \forall v \in \mathsf{FV}(\widehat{\phi}) \ \ \beta(v) = \gamma(v)^\alpha. \tag{$U_\phi \to U_{\widehat{\phi}}$}$$

Now, define $\widehat{c}$ to be the *completion* of c if for all ψ, $\psi \in \widehat{c}$ iff there exists $i, j \leq m$ and $\psi', \phi, \phi' \in c$ such that:

1. $\mathcal{M} \models \psi \lhd_i \psi' \wedge \phi \lhd_j \phi'$; and
2. $\phi \approx^{\mathcal{M}} \psi$.



[21] The idea of using such term trivializations was directly inspired by the work of Graham Leigh [16].

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

Let us observe that with the current definition of $\widehat{c}$ for all $\phi, \psi \in c$ the following holds:

$$\phi \approx^{\mathcal{M}} \psi \Rightarrow \mathsf{rank}^{\widehat{c}}(\phi) = \mathsf{rank}^{\widehat{c}}(\psi).$$

Indeed, suppose this is not the case. Then, assuming without loss of generality that

$$\mathsf{rank}^{\widehat{c}}(\phi) > i = \mathsf{rank}^{\widehat{c}}(\psi),$$

there exists $\phi' \in \widehat{c}$ such that $\phi' \lhd_{i+1} \phi$ but no formula from $\widehat{c}$ is the $i+1$-st direct subformula of ψ. By the congruence lemma (and induction) there exists ψ' such that $\psi' \lhd_{i+1} \psi$ and $\psi' \approx^{\mathcal{M}} \phi'$. Since $\phi' \in \widehat{c}$, there are $\theta, \theta', \phi'' \in c$ such that for some $j, k \leq m$ $\theta \lhd_j \theta'$ and $\phi' \lhd_k \phi''$ and $\phi' \approx^{\mathcal{M}} \theta$ (possibly $\phi'' = \phi = \theta'$ and $\theta = \phi'$—when $\phi' \in c$). Since $\approx^{\mathcal{M}}$ is an equivalence relation, $\psi' \approx^{\mathcal{M}} \theta$. Now, by the definition of $\widehat{c}$ we obtain that $\psi' \in \widehat{c}$, a contradiction.

For every x, let $F \restriction x$ denote the fragment of F consisting of axioms for U_ϕ predicates for ϕ of $\mathsf{rank}^{\widehat{c}}$ at most x and recall that if $\{\theta_\phi\}$ is a family of formulae with one free variable indexed with ϕ such that $\mathsf{rank}^{\widehat{c}}(\phi) \leq x$, then

$$F \restriction x[\theta_\phi / U_\phi]_{\mathsf{rank}^{\widehat{c}}(\phi) \leq x}$$

denotes the theory resulting from $F \restriction x$ by replacing every occurrence of U_ϕ with the formula θ_ϕ. Let $\zeta(x)$ be the formula asserting that there exists a unique family of $\mathcal{L}_{\mathsf{B}} \cup \{S\}$-formulae $\{\theta_\phi\}_{\mathsf{rank}^{\widehat{c}}(\phi) \leq x}$ indexed with formulae of $\mathsf{rank}^{\widehat{c}} \leq x$ such that the following conditions hold:

1. For every ϕ, if $\mathsf{rank}^{\widehat{c}}(\phi) = 0$, then
   (a) if $\mathcal{M} \models \exists t_1, \ldots, t_a \in \mathsf{Term}_{\mathcal{L}_{\mathsf{B}}} \phi = R(t_0, \ldots, t_a)$, then $\theta_\phi(x) = R(t_0^x, \ldots, t_a^x)$; and
   (b) if ϕ is from P, then $\theta_\phi(x) = S(\phi, x)$; and
   (c) if for some $\psi \in P$, $\phi \approx^{\mathcal{M}} \psi$, then U_ϕ is defined from U_ψ using $(U_{\widehat{\phi}} \to U_\phi)$ and $(U_\phi \to U_{\widehat{\phi}})$;
   (d) otherwise put $\theta_\phi(x) = (x \neq x)$.
2. $(\mathcal{M}, S, P) \models F \restriction x[\theta_\phi / U_\phi]_{\mathsf{rank}^{\widehat{c}}(\phi) \leq x}$.

Now we verify $\forall x \zeta(x)$ by a routine induction. This makes it clear that every finite fragment F of the Enayat–Visser theory Γ is consistent, which as pointed our earlier, can be readily coupled with Theorem 2.12 to complete the proof of Lemma 4.8. ⊣

- After the above preliminaries, we are finally ready to present the proof of Lemma 4.3.

Proof of Lemma 4.3. Working in PA, fix a good Δ_1-theory B, $l \in \mathbb{N}$ and a Δ_l-full model $\mathcal{M}$ of B. Next, still working in PA we shall construct an unbounded Δ_{l+1}-chain of Δ_{l+1}-full models

$$(\mathcal{M}_0, S_0), (\mathcal{M}_1, S_1, M_0), \ldots, (\mathcal{M}_x, S_x, M_{x-1}), \ldots$$

such that:

R1. $\mathcal{M} \preceq \mathcal{M}_0$;

and for each y we have:

R2. $\mathcal{M}_y \preceq \mathcal{M}_{y+1}$;

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

R3. $S_0 = \varnothing$ and S_{y+1} is an M_y-restricted satisfaction class for $\mathcal{L}_{\mathsf{B}}$ and
R4. $S_y \subseteq S_{y+1}$.

In particular each triple $(\mathcal{M}_x, S_x, M_{x-1})$ will have a fixed Δ_{l+1}-complexity. Let us assume that such a chain has been constructed and $\mathcal{M}_x(y)$ and $S_x(y)$ are formulae defining the sequences of respective $\mathcal{L}_{\mathsf{B}}$-full models and restricted satisfaction classes. For example it holds that $\mathcal{M}_x(y)$ iff y is the definition of the x-th full model (recall that officially full models are identified with their elementary diagrams). Then (in PA) we define the limit model with the formulae:

$$\mathcal{M}_\infty(z) := \exists x \exists y \in \mathsf{Form}^1_{\mathcal{L}_{\mathsf{PA}}} \left(\mathcal{M}_x(y) \wedge \mathsf{Sat}_{l+1}(y, z)\right),$$
$$S_\infty(z) := \exists x \exists y \in \mathsf{Form}^1_{\mathcal{L}_{\mathsf{PA}}} \left(S_x(y) \wedge \mathsf{Sat}_{l+1}(y, z)\right),$$

where $\mathsf{Sat}_{l+1}(x, y)$ denotes the canonical satisfaction predicate for Σ_{l+1}-formulae.[22] Note that $\mathcal{M}_\infty$ is really a full $\mathcal{L}_{\mathsf{B}}$-model, since the chain is elementary with respect to $\mathcal{L}_{\mathsf{B}}$-formulae and each $\mathcal{M}_x$ is a full model for $\mathcal{L}_{\mathsf{B}}$.

The rest of the argument follows along the lines of the Enayat–Visser proof: we check that S_∞ is a full satisfaction class on $\mathcal{M}_\infty$, hence $(\mathcal{M}_\infty, S_\infty)$ is a Δ_{l+2}-model of CT^-.

Let us now construct the promised chain of models: reasoning in PA, we first define a sequence of increasing theories $\langle \mathcal{T}_m : m \in \mathbb{N} \rangle$. Intuitively speaking, for each m, $\mathcal{T}_m$ describes a structure $\mathcal{K}_m = \langle (\mathcal{M}_i, S_i) : i \leq m \rangle$ and the family $\{(\mathcal{M}_i, S_i, M_{i-1})\}_{i \leq m}$ satisfies conditions R1–R4 for boundedly many numbers. In other words, $\{(\mathcal{M}_i, S_i, M_{i-1})\}_{i \leq m}$ is the initial segment of our desired chain consisting of the first $m + 1$ models.

We now give a precise description of $\mathcal{T}_m$. The nonlogical symbols of $\mathcal{T}_m$ consist of the symbols in $\mathcal{L}_{\mathsf{B}}$, together with constant symbols for every element of M, unary predicate symbols $\{\mathrm{M}_i : i \leq m\}$, and binary predicate symbols $\{\mathrm{S}_i : i \leq m\}$.

Convention 4.10. If ϕ is any formula (in the sense of PA), and $\mathrm{M}(x)$ is any of the M_i's then we write ϕ^{M} to denote the relativization of ϕ to M. This means that we syntactically replace all quantifiers $\exists x \alpha(x)$ with $\exists x \left(\mathrm{M}(x) \wedge \alpha(x)\right)$, all quantifiers $\forall x \alpha(x)$ with $\forall x \left(\mathrm{M}(x) \rightarrow \alpha(x)\right)$ and adding to ϕ a conjunct $\bigwedge_{x_i \in \mathsf{FV}(\phi)} \mathrm{M}(x_i)$.

The official translations of R1 through R4 above are as follows:

- Condition R1 is translated as $\left\{\phi^{\mathrm{M}_0} \mid \phi \in \mathsf{ElDiag}(\mathcal{M})\right\}$.
- Condition R2 is translated as

$$\left\{\forall x_0 \ldots \forall x_a \left(\phi(x_0, \ldots, x_a)^{\mathrm{M}_i} \rightarrow \phi(x_0, \ldots, x_a)^{\mathrm{M}_{i+1}}\right) \mid i < m, \phi(x_0 \ldots, x_a) \in \mathsf{Form}_{\mathcal{L}_{\mathsf{B}}}\right\}.$$

- Condition R3 is expressed by the conjunction of the universal closures of the following finitely many axioms 1i-6i, $0 \leq i \leq m$, which directly correspond to the ones from Definition 4.5 (we stipulate that $\phi^{\mathcal{M}_{-1}}(x)$ is always the formula $x \neq x$):

1i. $S_i(x, y) \rightarrow \left(\mathsf{Form}^{\mathrm{M}_{i-1}}_{\mathcal{L}_{\mathsf{B}}}(x) \wedge \mathsf{Asn}^{\mathrm{M}_i}(x, y)\right)$.

2i. $\left(\mathsf{TermSeq}^{\mathrm{M}_{i-1}}_{\mathcal{L}_{\mathsf{B}}}(\bar{s}) \wedge \left(x = R(\bar{s})\right)^{\mathrm{M}_{i-1}} \wedge \mathsf{Asn}^{\mathrm{M}_i}(x, \alpha)\right) \rightarrow \left(S_i(x, \alpha) \equiv (R(\bar{s}^\alpha))^{\mathrm{M}_i}\right)$.

[22] Recall that by Convention 2.33, $\mathsf{Sat}_{l+1}(y, z)$ means $\mathsf{Sat}_{l+1}(y, \zeta)$, where ζ is a valuation which assigns z to the only variable of y.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

3i. $\Big(\mathsf{Form}_{\mathcal{L}_\mathsf{B}}^{\mathrm{M}_{i-1}}(x) \wedge (x = \neg y)^{\mathrm{M}_{i-1}} \wedge \mathsf{Asn}^{\mathrm{M}_i}(x,\alpha)\Big) \to \Big((S_i(x,\alpha) \equiv \neg S_i(y,\alpha)\Big)$.

4i. $\Big(\mathsf{Form}_{\mathcal{L}_\mathsf{B}}^{\mathrm{M}_{i-1}}(x) \wedge (x = y_1 \vee y_2)^{\mathrm{M}_{i-1}} \wedge \mathsf{Asn}^{\mathrm{M}_i}(x,\alpha)\Big) \to$
$\Big(\mathrm{S}_i(x,\alpha) \equiv \big(\mathrm{S}_i\,(y_1,\alpha) \vee \mathrm{S}_i\,(y_2,\alpha)\big)\Big)$.

5i. $\Big(\mathsf{Form}_{\mathcal{L}_\mathsf{B}}^{\mathrm{M}_{i-1}}(x) \wedge \big(\exists v\ \ (\mathsf{Var}(v) \wedge x = \exists v\ y)\big)^{\mathrm{M}_i} \wedge \mathsf{Asn}^{\mathrm{M}_i}(x,\alpha)\Big) \to$
$\Big(\mathrm{S}_i(x,\alpha) \equiv \exists \alpha'\,\big((\alpha' \sim_v \alpha)^{\mathrm{M}_i} \wedge \mathrm{S}_i(y,\alpha')\big)\Big)$.

6i. $\Big(\mathsf{Form}_{\mathcal{L}_\mathsf{B}}^{\mathrm{M}_{i-1}}(x) \wedge \mathsf{VarSeq}^{\mathrm{M}_{i-1}}(\vec{v}) \wedge \mathsf{TermSeq}_{\mathcal{L}_\mathsf{B}}^{\mathrm{M}_{i-1}}(\vec{s}) \wedge \mathsf{TermSeq}_{\mathcal{L}_\mathsf{B}}^{\mathrm{M}_{i-1}}(\vec{t})$
$\wedge\ \mathsf{Asn}^{\mathrm{M}_i}(x,\vec{s},\vec{t},\alpha)\Big) \to \Big(\ \big((y_1 = x[\vec{s}/\vec{v}])^{\mathrm{M}_i} \wedge (y_2 = x[\vec{t}/\vec{v}])^{\mathrm{M}_i}$
$\wedge\ (\vec{s}^{\,\alpha} = \vec{t}^{\,\alpha})^{\mathrm{M}_i}\Big) \to \Big(S_i(y_1,\alpha) \equiv S_i(y_2,\alpha)\Big)\Big)$.

- Condition R4 is expressed by the following finite set of sentences:
$$\big\{\forall x \forall \alpha\big(\big(\mathrm{S}_i(x,\alpha) \to \mathrm{S}_{i+1}(x,\alpha)\big) : i < m\big\}.$$

We can now use induction on m to show that $\forall m\ \mathsf{Con}(\mathcal{T}_m)$:

Base case. Recall that $\mathcal{M}$ is a fixed Δ_l-full model of B. Let $S_0 = \varnothing$. Then since S_0 is definable in $\mathcal{M}$, the elementary diagram of $\mathcal{K}_0 := (\mathcal{M}, S_0)$ is also definable. This makes it clear that $\mathsf{Con}(\mathcal{T}_0)$ holds.

Inductive step. Fix m and suppose that $\mathsf{Con}(\mathcal{T}_m)$ holds. Then by Theorem 2.12, there is a full model $\mathcal{K}_m$ of $\mathcal{T}_m$ satisfying R1 through R4 above whose elementary diagram is Δ_{1+l}-definable.

Let $\mathcal{L}_{\mathcal{T}_m}$ be the language of $\mathcal{T}_m$, and let $\mathcal{M}_m$ be the *reduct* of the structure $\mathcal{K}_m$ to the language $\mathcal{L}_\mathsf{B}$ in which the universe of discourse is the $\mathcal{K}_m$-interpretation of M_m. For example, since $\mathcal{L}_{\mathcal{T}_1} = \{\mathrm{M}_1, \mathrm{M}_0, +, \cdot, \mathrm{S}_0, \mathrm{S}_1\}$, a model $\mathcal{K}_1$ of $\mathcal{T}_1$ will be a structure of the form:
$$(K_1, M_1, M_0, \oplus, \odot, S_0, S_1),$$
where $M_i = \mathrm{M}_i^{\mathcal{K}}$, $S_i = \mathrm{S}_i^{\mathcal{K}}$, and K_1 is the domain of discourse of $\mathcal{K}_1$. In this case, $\mathcal{M}_1 = (M_1, \oplus, \odot)$.

So in general $\mathcal{M}_m$ is of the form $(M_m, \oplus, \odot)$.[23] Observe that $\mathcal{M}_m$ is a full model. Typically, its domain is smaller than the domain of $\mathcal{K}_m$.

To this model apply Lemma 4.8 for $\mathcal{M} = \mathcal{M}_m$, $S = S_m$ and $P = M_{m-1}$. We are given $\mathcal{N}$, a Δ_{l+2}-full model for $\mathcal{L}_\mathsf{B}$, and a Δ_{l+2}-set S' such that S' is an M_m-restricted satisfaction class and $\mathcal{M}_m \preceq \mathcal{N}$. Now we "glue" this model to the end of the chain given by $\mathcal{K}_m$. More precisely, we define a model $\mathcal{K}_{m+1}$ for $\mathcal{L}_{m+1}$ in the following way. The universe of $\mathcal{K}_{m+1}$ is the union of the universes of $\mathcal{K}_m$ and $\mathcal{N}$ (without loss of generality, renaming the elements of $N \setminus M_m$ if necessary, we assume that $K_m \cap N = M_m$). M_{m+1} is interpreted as N, S_{m+1} as S' and $+$ and $\cdot$ are interpreted on elements from N as they were in $\mathcal{N}$. For $0 \leq i \leq m$ M_i and S_i are interpreted as in $\mathcal{K}_m$. Thus we have obtained a structure which contains an elementary chain of models of B, with $\mathcal{N}$ being the top one and possibly some extra elements in the domain of $K_m \setminus N$.

[23] Recall the conventions from Section 2.2. Although officially full models are elementary diagrams, we refer to them as though they were usual structures, as it is routine to translate statements about complete Henkinized theories into statements about structures.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

Also note that for a structure defined in this manner we do not have an elementary diagram at our disposal, hence an argument is needed to show that $\mathsf{Con}(\mathcal{T}_{m+1})$ holds. We argue as in the proof of Corollary 2.39. Note that if π is a purported proof of a contradiction from the axioms of $\mathcal{T}_{m+1}$, and π has the subformula property, then only the following four types of sentences can occur in π:

A. formulae of the form ϕ^{M_0} for $\phi \in \mathcal{L}_{\mathsf{B}}$.
B. subformulae of sentences of the form

$$\forall x_0 \ldots \forall x_a \left(\phi(x_0, \ldots, x_a)^{\mathrm{M}_i} \to \phi(x_0, \ldots, x_a)^{\mathrm{M}_{i+1}}\right).$$

for $\phi(x_0, \ldots, x_1) \in \mathsf{Form}_{\mathcal{L}_{\mathsf{B}}}$, $i < m+1$.
C. subformulae of sentences in the list 1i - 6i, where $i \leq m+1$.
D. subformulae of sentences of (the formalization) of condition R4.

The complexity of each formula from C and D is bounded by a fixed standard number. This is not the case of formulae from A or B. However, to decide every such sentence we can use $\mathsf{ElDiag}(\mathcal{K}_m)$ and $\mathsf{ElDiag}(\mathcal{N})$ and this is clearly sufficient (all formulae from B are in the universal closure of boolean closure of formulae of type ϕ^{M_i} for $i \leq m+1$). All in all, we can define a Σ_n-truth predicate for $\mathcal{K}_{m+1}$, for sufficiently large n, which would work for all formulae from the proof π. It follows that π cannot be a proof of a contradiction. This ends the inductive step and we can conclude that $\forall m\ \mathsf{Con}(\mathcal{T}_m)$ holds.

We shall now define the promised chain of models as a full model of the limit of $\mathcal{T}_m$'s. Define:

$$\mathcal{T}_\infty := \bigcup \{\mathcal{T}_c \mid c \in \mathbb{N}\}.$$

Here $\mathbb{N}$ is treated internally, i.e., it simply denotes the universe. T_∞ is a consistent theory of complexity Δ_l (it is computable in $\mathsf{ElDiag}(\mathcal{M})$). It follows that it has a Δ_{l+1}-full model $\mathcal{K}_\infty$. This model gives rise to the Δ_{l+1}-chain of Δ_{l+1}-full models $(\mathcal{M}_x, S_x, M_{x-1})_{x\in\mathbb{N}}$, which can be defined as follows:

$$\begin{aligned} M_x(y) &:= \mathcal{K}_\infty \models \mathrm{M}_x(y), \\ \mathcal{M}_x \models \phi &:= \mathcal{K}_\infty \models \phi^{\mathcal{M}_x}, \\ S_x(y,z) &:= \mathcal{K}_\infty \models \mathrm{S}_x(y,z). \end{aligned}$$

The construction guarantees that under such a definition, the chain $(\mathcal{M}_x, S_x, M_{x-1})_{x\in\mathbb{N}}$ satisfies the requirements R1 through R4. This concludes the proof of Lemma 4.3. ⊣

4.2. Feasible reduction of $\mathsf{KF}^-[\mathsf{PA}]$ to PA. In this subsection we will establish:

THEOREM 4.11. $\mathsf{KF}^-[\mathsf{PA}]$ *is feasibly reducible to* PA.

We will prove the above theorem by demonstrating that the assumption of Corollary 2.40 holds with the choice of $\mathcal{T} = \mathsf{KF}^-[\mathsf{PA}]$ and $k = 4$, i.e., we will prove:

LEMMA 4.12. PA *proves the sentence expressing*:

"If B *is any finite fragment of* PA, *then every* Δ_2*-full model* $\mathcal{M}$ *of* B *has an elementary extension to a* Δ_4*-full model* $\mathcal{N}$ *which has an expansion to a* Δ_4*-model of* $\mathsf{KF}^-[\mathsf{B}]$*."*

Before proving Lemma 4.12, we will first show that PA can formalize the proof of the existence of recursively saturated models over a recursive language.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

LEMMA 4.13. *For any* $k \in \mathbb{N}$ PA *proves the sentence expressing*:

"If $\mathcal{M}$ *is a* Δ_k*-full model for some* Δ_1*-language* $\mathcal{L}$*, then there exists a* Δ_{k+1}*-full model* $\mathcal{N}$ *that elementarily extends* $\mathcal{M}$ *such that* $\mathcal{N}$ *is recursively saturated."*

Let us first make sense of the above lemma. Recall (from Definition 2.6) that by a *full model* $\mathcal{M}$ over a language $\mathcal{L}$, we mean the elementary diagram of that model, that is, a complete consistent Henkinized theory. Also recall that a model $\mathcal{M}$ is said to be *recursively saturated* if for every Turing machine with code e, every finite sequence $a_1, \ldots, a_r$ of elements of $\mathcal{M}$, and every finite sequence $\phi_1(x, \bar{y}), \ldots, \phi_k(x, \bar{y})$ of formulae whose Gödel numbers are accepted by the Turing machine with code e, there is an element a in $\mathcal{M}$ such that:

$$\mathcal{M} \models \bigwedge_{i \leq k} \phi_i(a, a_1, \ldots, a_r),$$

then there exists d in $\mathcal{M}$ such that for every $\phi \in \mathcal{L}$ which is accepted by the Turing machine with code e,

$$\mathcal{M} \models \phi(d, a_1, \ldots, a_r).$$

The above definition is well known. We cite it here to assure the reader that it really can be spelled out in PA. The lemma itself was noted by Simpson (cf. [19], Lemma IX.4.2). We demonstrate it here for the convenience of the reader.

PROOF OF LEMMA 4.13. We reason in PA. Let $\mathcal{L}^*$ be the result of augmenting the language $\mathcal{L}$ of $\mathcal{M}$ with constants $c_{i,j}, i, j \in \mathbb{N}$. Let (ϕ_i^*) be a recursive (i.e., Δ_1) enumeration of all sentences of the language $\mathcal{L}^*$. Let $\mathcal{M} \models \mathsf{B}$ be a full model and let $\mathsf{ElDiag}(\mathcal{M})^*$ be the theory whose axioms consist of the elementary diagram of $\mathcal{M}$ (which, according to our official definition from Section 2.2 is the full model $\mathcal{M}$ itself), together with all Henkin sentences (in the language with the new constants) and all sentences of the following shape:

$$\bigwedge_{i \leq N} \Big(\exists x \big(\phi_1^*(x, \bar{y}) \wedge \cdots \wedge \phi_k^*(x, \bar{y}) \big) \rightarrow \phi_1^*(c_{i,e}, \bar{y}) \wedge \cdots \wedge \phi_k^*(c_{i,e}, \bar{y}) \Big),$$

where $N \in \mathbb{N}$, and all the constants of $\mathcal{L}^*_{\mathsf{PA}} \setminus \mathcal{L}_{\mathsf{PA}}$ occurring in the formulae $\phi_1^*, \ldots, \phi_k^*$ are of the form $c_{j,l}$ for $j < i$, and the machine with the code e accepts sentences $\phi_1^*, \ldots, \phi_k^*$ in less than N steps. By Theorem 2.12 (ACT) the theory $\mathsf{ElDiag}(\mathcal{M})^*$ has a Δ_{k+1}-full model $\mathcal{M}'$. This ends the proof of Lemma 4.13. ⊣

Now, we proceed to the proof of Lemma 4.12, which will conclude the proof of Theorem 4.11.

PROOF OF LEMMA 4.12. We work in PA. Let $\mathcal{M}$ be any Δ_2-model of B. By Lemma 4.13, there exists a Δ_3-full recursively saturated model $\mathcal{N}$ of B. In our proof, we use a construction resembling the one given originally by Kripke in [15]. A very similar argument appeared before in [2] and [4]. By induction, we define a sequence of arithmetical formulae $\Gamma_c, c \in N$. That is, a sequence of elements $\Gamma_c \in N$ such that $\mathcal{N} \models \Gamma_c \in \mathsf{Form}^{\leq 1}_{\mathcal{L}_{\mathsf{PA}}}$. Let $\Gamma_0(x)$ be a definition of the atomic diagram of $\mathcal{N}$. More precisely, let

$$\Gamma_0(x) := \exists s, t \in \mathsf{ClTerm}_{\mathcal{L}_{\mathsf{PA}}} \ \ x = (s = t) \wedge s^\circ = t^\circ.$$

Having defined the formula Γ_n, we set $\Gamma_{n+1}(\phi)$ (which we also denote by $\phi \in \Gamma_{n+1}$) if and only if one of the following conditions is satisfied:

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

- $\bigvee_{j\leq n}\ \phi \in \Gamma_j$.
- $\exists t \in \mathsf{ClTerm}_{\mathcal{L}_\mathsf{B}}\ (\phi = T(t)) \wedge t^\circ \in \Gamma_n$.
- $\exists t \in \mathsf{ClTerm}_{\mathcal{L}_\mathsf{B}}\ (\phi = \neg T(t)) \wedge (\neg t^\circ) \in \Gamma_n$.
- $\exists \psi \in \mathsf{Sent}_{\mathcal{L}_T}\ (\phi = \neg\neg\psi) \wedge \psi \in \Gamma_n$.
- $\exists \psi, \eta \in \mathsf{Sent}_{\mathcal{L}_T}\ (\phi = (\psi \vee \eta)) \wedge (\psi \in \Gamma_n \vee \eta \in \Gamma_n)$.
- $\exists \psi, \eta \in \mathsf{Sent}_{\mathcal{L}_T}\ (\phi = \neg(\psi \vee \eta)) \wedge (\neg\psi \in \Gamma_n \wedge \neg\eta \in \Gamma_n)$.
- $\exists v \in \mathsf{Var}\ \psi \in \mathsf{Form}^{\leq 1}_{\mathcal{L}_T}\ (\phi = \exists v\ \psi) \wedge \exists x\ (\psi(\underline{x}) \in \Gamma_n)$.
- $\exists v \in \mathsf{Var}\ \psi \in \mathsf{Form}^{\leq 1}_{\mathcal{L}_T}\ (\phi = \neg\exists v\ \psi) \wedge \forall x((\neg\psi(\underline{x})) \in \Gamma_n)$.

Now, let T be the subset of the domain of N defined as the union $\bigcup_{i\in\mathbb{N}} \Gamma_i(\mathcal{N})$. In other words,

$$T(x) := \exists y\ \mathcal{N} \models \Gamma_y(x).$$

Consider the expanded model $(\mathcal{N}, T)$. Since the definition of $(\mathcal{N}, T)$ is Σ_1 in the complexity of $\mathcal{N}$, the complexity of the resulting model is Σ_3, hence in particular it is Δ_4. We would like to ensure that $(\mathcal{N}, T)$ is a model $\mathsf{KF}^-[\mathsf{B}]$. The model $(\mathcal{N}, T)$ satisfies B, since $\mathcal{N}$ does, so it is enough to check that $(\mathcal{N}, T)$ satisfies truth-theoretic axioms KF1–KF12.

This is obvious for KF1 and KF2. The axioms KF3 and KF4 are omitted if the base theory is PA, since in our formulation PA has no relation symbols. Let us check the claim for KF6. Suppose that $(\mathcal{N}, T) \models T(\phi \vee \psi)$. Since $(\mathcal{N}, T) \models T(\phi \vee \psi)$, there exists i such that

$$\mathcal{N} \models \Gamma_i(\phi \vee \psi).$$

Then by definition of Γ_i, either $\mathcal{N} \models \phi \in \Gamma_{i-1}$ (and, consequently, $T(\phi)$ holds) or $\mathcal{N} \models \psi \in \Gamma_{i-1}$ (and then $T(\psi)$ holds). Conversely, if $(\mathcal{N}, T) \models T(\phi)$ or $(\mathcal{N}, T) \models T(\psi)$, then for some i, $\mathcal{N} \models \phi \in \Gamma_i$ or $\mathcal{N} \models \psi \in \Gamma_i$. But then $\phi \vee \psi \in \Gamma_{i+1}$ and, consequently, $(\mathcal{N}, T) \models T(\phi \vee \psi)$. This guarantees that $(\mathcal{N}, T) \models$ KF6. The proofs for axioms KF5 and KF7 are similar, as are the proofs for axioms KF11 and KF12. Let us focus on axiom KF9.

Suppose that $(\mathcal{N}, T) \models T(\neg\exists v\ \psi)$. Then there exists i such that $\mathcal{N} \models (\neg\exists v\ \psi) \in \Gamma_i$. This implies that for all x, $\mathcal{N} \models \neg\psi(\underline{x}) \in \Gamma_{i-1}$. Therefore, for all $x \in N$, $(\mathcal{N}, T) \models T(\neg\psi(\underline{x}))$ holds.

Conversely, suppose that for all $x \in N$, $(\mathcal{N}, T) \models T(\neg\psi(\underline{x}))$. In other words, for every $x \in N$, there exists i such that $\mathcal{N} \models (\neg\psi(\underline{x})) \in \Gamma_i$. We claim that there exists k such that for all x, $\mathcal{N} \models (\neg\psi(\underline{x})) \in \Gamma_k$. Suppose otherwise. Then for every k, the following set of arithmetical formulae is realized in $\mathcal{N}$ by some x:

$$\neg\psi(\underline{x}) \notin \Gamma_0 \wedge \cdots \wedge \neg\psi(\underline{x}) \notin \Gamma_k.$$

Therefore, by recursive saturation, there exists an $a \in N$ such that for every k,

$$\neg\psi(\underline{a}) \notin \Gamma_k,$$

contrary to the assumption. This implies that there exists $k \in N$ such that $\mathcal{N} \models \neg\psi(\underline{x}) \in \Gamma_k$ for every $x \in N$, and therefore $\mathcal{N} \models (\neg\exists x\psi(x)) \in \Gamma_{k+1}$. We conclude that $(\mathcal{N}, T)$ satisfies KF9. The case of the axiom KF8 is straightforward.

In order to prove that KF10 holds, we check by induction on n (in PA) that this axiom is satisfied by formulae in Γ_n. The conclusion follows immediately, thus completing the proof of the lemma. ⊣

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

REMARK 4.14. Motivated by a question posed by the referee, let us consider the following two axioms which may be added to KF^- and have been studied in the literature:

$$\forall \phi \in \mathsf{Sent}_{\mathcal{L}_T} \ \neg(T(\phi) \wedge T(\neg\phi)). \tag{Cons}$$

$$\forall \phi \in \mathsf{Sent}_{\mathcal{L}_T} \ T(\phi) \vee T(\neg\phi). \tag{Comp}$$

"Cons" stands for "consistent" and "Comp" stands for "complete."

The argument presented above actually guarantees that both theories $\mathsf{KF}^- + \mathsf{Cons}$ and $\mathsf{KF}^- + \mathsf{Comp}$ can be feasibly reduced to PA. Indeed, we would like to show Lemma 4.12 for $\mathsf{KF}^-[\mathsf{B}] + \mathsf{Cons}$ and $\mathsf{KF}^-[\mathsf{B}] + \mathsf{Comp}$ in place of KF^-.

We claim that the expanded model $(\mathcal{N}, T)$ defined above actually satisfies the axiom Cons. Working in PA, we check by induction on y that no set Γ_y contains both a sentence and its negation. In other words:

$$\forall y \forall \phi \in \mathsf{Sent}_{\mathcal{L}_T} \ \neg\left(\phi \in \Gamma_y \wedge (\neg\phi) \in \Gamma_y\right).$$

We directly check by cases, depending on the syntactic shape of ϕ, that if both ϕ and $\neg\phi$ are in Γ_y, then there exists a $y' < y$ and a sentence ψ such that both ψ and $\neg\psi$ are in $\Gamma_{y'}$.

Now, suppose that $(\mathcal{N}, T)$ is a model of $\mathsf{KF}^-[\mathsf{B}] + \mathsf{Cons}$. Then we construct a model $(\mathcal{N}, T')$ of $\mathsf{KF}^-[\mathsf{B}] + \mathsf{Comp}$ simply by defining $T'(\phi)$ as $(\phi \in \mathsf{Sent}_{\mathcal{L}_T}) \wedge \neg T(\neg\phi)$.

4.3. Feasible reduction of $\mathsf{FS}^-[\mathsf{PA}]$ to PA. In this section we strengthen the conservativity proof from Section 3.3.3 by establishing the following result:

THEOREM 4.15. FS^- *is feasibly reducible to* PA.

The key step in our construction is to feasibly reduce the theory of ω-many truth predicates, $\mathsf{RT}^-_{<\omega}$, defined in Section 3.3.3, to PA. This is achieved in the following lemma.

LEMMA 4.16. $\mathsf{RT}^-_{<\omega}$ *is feasibly reducible to* PA.

PROOF. We shall prove that the assumptions of Corollary 2.39 hold for $\mathcal{T} = \mathsf{RT}^-_{<\omega}$ and $q(n) = n$. In fact, we shall show slightly more: working in PA, for an *arbitrary* coded fragment $\mathsf{B} \supseteq I\Delta_0 + \mathsf{Exp}$ of PA and a Δ_2-full model $\mathcal{M} \models \mathsf{B}$, we shall build a Δ_4-model of $\mathsf{RT}^-_{<\omega}[\mathsf{B}]$ (note that now we are talking about ω internally). To this end, working in PA, fix B as above.

The aim is to formalize the conservativity proof from Section 3.3.3 in PA. In order to do this we shall build a chain of uniformly definable Δ_3-full models $(\mathcal{M}_n)_{n\in\mathbb{N}}$ such that $\mathcal{M} \preceq_{\mathcal{L}_{\mathsf{PA}}} \mathcal{M}_0$ and for each $n \in \mathbb{N}$ the following holds:

1. $\mathcal{M}_n$ is a full Δ_3-model of $\mathsf{RT}^-_{<n+1}[\mathsf{B}]$; and
2. $\mathcal{M}_k \preceq_{\mathcal{L}_{<k+1}} \mathcal{M}_n$ for each $k < n$.

Clearly the limit model will be a model of RT^-_{ω} (even a full one—this follows by elementarity). To define the respective chain we shall implement the argument from Section 4.1: the chain $\mathcal{M}_0, \ldots, \mathcal{M}_k$ will be described by a Δ_2-theory $\mathcal{T}_k$ formulated in the language $\mathcal{L}_{\mathcal{T}_k}$ whose nonlogical symbols consist of:

1. symbols of $\mathcal{L}_{\mathsf{B}}$;
2. unary predicates: $\mathrm{M}_0, \ldots, \mathrm{M}_k$; and

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

3. unary predicates: $T_0, \ldots, T_k$.[24]

As in the proof for $\mathsf{CT}^-[\mathsf{B}]$ in Section 4.1, the axioms of $\mathcal{T}_k$ can be divided into three groups:

1. $\mathcal{M}$ is an elementary submodel of $\mathcal{M}_0$. Formally this is expressed as an infinite set of axioms: $\{\phi^{\mathrm{M}_0} \mid \phi \in \mathsf{ElDiag}(\mathcal{M})\}$.
2. $(\mathcal{M}_i)_{i \leq k}$ forms an elementary chain of submodels. More precisely: for each i, $\mathcal{M}_i$ is an $\mathcal{L}_{<i+1}$-elementary submodel of $\mathcal{M}_{i+1}$. Formally this is expressed analogously to the condition (R2) from the proof for CT^-.
3. For every $i \leq k$, $\mathcal{M}_i$ is a model of $\mathsf{RT}^-_{<i+1}$. This is expressed by formally relativizing the axioms of $\mathsf{RT}^-_{<i+1}$ to M_i.

Now, by induction on n we show that $\forall n \mathsf{Con}(\mathcal{T}_n)$. We proceed as in the sketch of the conservativity proof given in Section 3.3.3. For $n = 0$ we simply use the proof from Section 4.1 to build an elementary extension of $\mathcal{M}$ satisfying $\mathsf{CT}^-[\mathsf{B}]$. For the induction step, note that, using the same reasoning as we did in Section 4.1 to verify $\mathsf{Con}(\mathcal{T}_{k+1})$, it is enough to build a model for $\mathsf{RT}^-_{<k+1}$ which would be a full model for $\mathcal{L}_{<k}$ but will possibly leave some sentences with T_k undefined.

As in the conservativity proof for $\mathsf{RT}^-_{<\omega}$ we use the fact that $\mathsf{RT}^-_{<k+1}$ is deductively equivalent to the theory $\mathcal{IT}$ below[25]:

$$\mathsf{CT}^-[\mathsf{RT}^-_{<k}] + \forall \phi \in \mathsf{Sent}_{\mathcal{L}_{<k-1}} \ \big(T_{k-1}(\phi) \equiv T(\phi)\big). \tag{$\mathcal{IT}$}$$

From $\mathsf{Con}(\mathcal{T}_k)$ we obtain a Δ_2-full model $\mathcal{K}$ of $\mathsf{RT}^-_{<k}$. We build an extension satisfying $\mathcal{IT}$ in ω many steps via the union of chain argument. The following is the analogue of Lemma 4.8 in our situation:

Lemma 4.17 (Arithmetized Enayat–Visser Lemma+). *The universal generalization of the formula $\theta_1 \to \theta_2$ is provable in* PA *for every $l \in \mathbb{N}$, where θ_1 expresses*:

"B *is a good theory, $(\mathcal{M}, S, P)$ is a Δ_l-full model for $\mathcal{L}_{<k} \cup \{S\} \cup \{P\}$ such that $\mathcal{M}$ is a model of* B*; and S is a P-restricted satisfaction class for $\mathcal{L}_{<k}$."*

and θ_2 expresses:

"There exist a Δ_{l+1}-full model $\mathcal{N}$ and a Δ_{l+1}-set $S^* \subseteq N^2$ such that $\mathcal{M} \preceq \mathcal{N}$; S^* is an M-restricted satisfaction class for $\mathcal{L}_{<k}$ (we add a predicate for the universe of $\mathcal{M}$ to the language); $S \subseteq S^*$; and for every $\phi \in \mathsf{Form}_{<k-1}(\mathcal{M})$, $(\mathcal{N}, S^*, M) \models T_{k-1}(\phi) \to \forall \alpha S^*(\phi, \alpha)$."

Sketch Of The Proof. We indicate how to modify the proof Lemma 4.8. Firstly, we add the following sentences to the definition of the Enayat–Visser theory of $(\mathcal{M}, S, P)$:

$$\{\forall \alpha \ \ U_\phi(\alpha) \mid \mathcal{M} \models T_{k-1}(\phi)\}. \tag{$*$}$$

Now we work with a finite fragment F of the Enayat–Visser theory. The next step, which requires a modification, is the definition of rank^b for a coded set of sentences b. According to the previous definition, a formula ϕ was assigned rank^b zero if and only if either ϕ was atomic or some immediate subformula of ϕ was outside b. Now

[24] Here we are continuing to use the convention of our article of using the same symbol to represent a truth predicate and an interpretation of it, so officially these predicates should be written as $\mathrm{T}_0, \ldots \mathrm{T}_k$.

[25] "$\mathcal{I}$" abbreviates "Induction" as this theory is used in the induction step of our construction.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

we will treat as formulae of rank^b zero all formulae from $\mathsf{Form}_{\mathcal{L}_{<k-1}}(\mathcal{M})$ as well. For such formulae ϕ we have an obvious candidate for the definition of $\theta_\phi(x)$ (i.e., the formula defining the extension for $U_\phi(x)$ in $\mathcal{M}$). We define:

$$\theta_\phi(\alpha) := T_{k-1}(\phi[\alpha]).$$

Note that T_{k-1} satisfies generalized regularity, so it is sufficient to verify the truth of ϕ on numerals naming the values of α. The definition of $\mathsf{rank}^b \geq x$ is now as follows: there exists a sequence y such that the following holds:

1. $\mathsf{len}(y) = x + 1$ and $(y)_x = \{\phi\}$.
2. For all $i < x + 1$ $(y)_i \subseteq b$.
3. For all $i < x$ for all θ, $\theta \in (y)_{i+1}$ iff $\theta \in \mathsf{Form}_{\mathcal{L}_{<k}} \setminus \mathsf{Form}_{\mathcal{L}_{<k-1}}$ and for all ψ such that $\mathcal{M} \models \psi \lhd \theta$, $\psi \in (y)_i$.[26]

The definitions of $\mathsf{rank}^b = x$ and $\widehat{b}$ (for an arbitrary b) are analogous to the ones from the original lemma. The last step which requires a modification is the definition of the formula $\zeta(x)$. Below, as in the proof for CT^-, c is the set of formulae ϕ such that U_ϕ occurs in F. We define $\zeta(x)$ to be the formula expressing:

"There exists a unique family of $\mathcal{L}_{<k} \cup \{S\}$-formulae $\{\theta_\phi\}_{\mathsf{rank}^{\widehat{c}}(\phi) \leq x}$ indexed with formulae of $\mathsf{rank}^{\widehat{c}} \leq x$ such that:

1. For every ϕ, if $\mathsf{rank}^{\widehat{c}}(\phi) = 0$, then:
   (a) if $\mathcal{M} \models \exists t_1, \ldots, t_a \in \mathsf{Term}_{\mathcal{L}_\mathsf{B}} \phi = R(t_0, \ldots, t_a)$ for a relation symbol $R \in \mathcal{L}_\mathsf{B}$, then $\theta_\phi(\alpha) = R(t_0^\alpha, \ldots, t_a^\alpha)$; and
   (b) if $\mathcal{M} \models \exists t \in \mathsf{Term}_{\mathcal{L}_\mathsf{B}}(\phi = T_{k-1}(t))$, then $\theta_\phi(\alpha) = T_{k-1}(t^\alpha)$; and
   (c) if $\phi \in \mathsf{Form}_{\mathcal{L}_{<k-1}}(\mathcal{M})$, then $\theta_\phi(\alpha) := T_{k-1}(\phi[\alpha])$; and
   (d) if ϕ is from P, then $\theta_\phi(\alpha) = S(\phi, \alpha)$; and
   (e) if for some $\psi \in P$, $\phi \approx^{\mathcal{M}} \psi$, then U_ϕ is defined from U_ψ using $(U_{\widehat{\phi}} \to U_\phi)$ and $(U_\phi \to U_{\widehat{\phi}})$;
   (f) otherwise $\theta_\phi(x) = (x \neq x)$.
2. $(\mathcal{M}, S, P) \models F \restriction x[\theta_\phi / U_\phi]_{\mathsf{rank}^{\widehat{c}}(\phi) \leq x}$.

Note that conditions (c) through (e) are the same as in the original definition.

The rest of the proof is as previously. ⊣

Once we can prove $\forall n \mathsf{Con}(\mathcal{T}_n)$, the construction of the chain $(\mathcal{M}_n)_{n \in \omega}$ and its union is precisely the same as in Section 4.1. ⊣

Now we want to finish the proof of Theorem 4.15. We have just shown that $\mathsf{RT}^-_{<\omega}[\mathsf{PA}]$ satisfies the assumptions of Corollary 2.39 (including the "moreover" part).

By Observation 2.42, it follows that $\mathsf{RT}^-_{<\omega}[\mathsf{PA}]$ is PA-provably feasibly strongly reflexive, i.e., there exists a P-time computable function $f(s_0, s_1)$ such that for all $n, k \in \mathbb{N}$, $f(\mathsf{tal}(n), \mathsf{tal}(k))$ is a PA-proof of the sentence

$$\forall \phi \in \mathsf{Sent}_{\mathcal{L}_\mathsf{PA}} \big(\mathsf{dp}(\phi) \leq \underline{k} \wedge \mathsf{Pr}_{\mathsf{RT}^-_{<\omega} \restriction \underline{n}}(\phi) \to \mathsf{Tr}_k(\phi)\big). \qquad (\mathrm{REF}_k(\mathsf{RT}^-_{<\omega} \restriction n))$$

Note that there exists a P-time computable function $g(s)$ such that for any n, $g(\mathsf{tal}(n))$ is a PA-proof of the sentence

$$\forall \phi \in \mathsf{Sent}_{\mathcal{L}_\mathsf{PA}} \left(\mathsf{dp}(\phi) \leq \underline{k} \wedge \mathsf{Pr}_{\mathsf{FS}^-_{\underline{n}}}(\phi) \to \mathsf{Pr}_{\mathsf{RT}^-_{<\underline{2n}+1}}(\phi)\right). \qquad (\mathrm{HR}_{n,k})$$

[26] Note that if $\phi \in \mathsf{Form}_{\mathcal{L}_{<k-1}}$, this condition implies that y has length 1.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

The above is in fact an easy consequence of the proof of Halbach's reduction (HR) of FS^- to $\mathsf{RT}^-_{<\omega}$ from Lemma 3.9.

Now, we would like to verify that there is a P-time computable function $e(s,t)$ such that for every n, $e(\mathsf{tal}(n),\mathsf{tal}(k))$ is a PA-proof of the sentence

$$\forall \phi \in \mathsf{Sent}_{\mathcal{L}_{\mathsf{PA}}} \left(\mathsf{dp}(\phi) \leq \underline{k} \wedge \mathsf{Pr}_{\mathsf{FS}^-_{\underline{n}}}(\phi) \rightarrow \mathsf{Tr}_k(\phi)\right). \qquad (\mathrm{REF}_k(\mathsf{FS}_n))$$

However, we should exercise caution, since in $\mathrm{REF}_k(\mathsf{RT}^-_{<\omega} \restriction n)$ and $\mathsf{HR}_{n,k}$ we are dealing with two different kinds of restrictions of $\mathsf{RT}^-_{<\omega}$: $\mathsf{RT}^-_{<\omega} \restriction n$ consists of axioms of $\mathsf{RT}^-_{<\omega}$ of length at most n and $\mathsf{RT}^-_{<2n+1}$ consists of compositional axioms for the first $2n+1$ iterations of the truth predicate. Since the length of T_n is roughly $\log(n)$, the length of each axiom of $\mathsf{RT}^-_{<2n+1}$ is bounded above by $(2n+1)(3\log(2n)+C) \leq n^3 + C'$, where C, C' are independent from n (this number bounds the length of RT6, which is a conjunction of $2n+1$ axioms of length $3\log(2n)+C$). This bound[27] can be readily calculated using the arithmetical definition of $\mathsf{RT}^-_{<2n+1}$, and as the proof is uniform in n, there exists a P-time computable function $c(s)$ such that for every n, $c(\underline{n})$ is a PA-proof of the following sentence:

$$\forall x \ \left(\mathsf{RT}^-_{<2\underline{n}+1}(x) \rightarrow \mathsf{len}(x) < \underline{n} \cdot \underline{n} \cdot \underline{n} + \underline{C'}\right).$$

Now, the proof of $(\mathrm{REF}_k(\mathsf{FS}_n))$ can be given as follows: given n,k compute $g(\mathsf{tal}(n),\mathsf{tal}(k))$ (i.e., a proof of $(\mathsf{HR}_{n,k})$) and let $l := n^3 + C'$. Next, compute $f(\mathsf{tal}(l),\mathsf{tal}(k))$ (i.e., a proof of $\mathrm{REF}_k(\mathsf{RT}^-_{<\omega} \restriction l)$) and by concatenating the two proofs (adding a constant number of logical transformations), obtain a proof of $(\mathrm{REF}_k(\mathsf{FS}_n))$, which is our value of $e(\mathsf{tal}(n),\mathsf{tal}(k))$.

Finally, let us observe that the relation $R(k,n,m) \subseteq \mathbb{N}^3$ defined via:

"k is an FS^-_n-proof of m"

is P-time computable, so by Corollary 2.24 it is feasibly numerable in $\mathsf{I}\Delta_0 + \mathsf{Exp}$. This gives us a P-time function $h(s_0,s_1,s_2)$ such that for every FS^-_n-proof π (with code k) of a sentence ϕ, $h(\underline{k},\underline{n},\underline{\phi})$ is a PA-proof of $\mathsf{Proof}_{\mathsf{FS}^-_{\underline{n}}}(\underline{k},\underline{\phi})$.

Our desired reduction can now be defined as follows: given an FS^--proof π with code k of a sentence ϕ compute n and m such that there are exactly n applications of NEC and CONEC in π and ϕ is of depth m. Observe that both $|\mathsf{tal}(n)|$ and $|\mathsf{tal}(m)|$ are at most $|\underline{k}|$. Using h find a proof of $\mathsf{Pr}_{\mathsf{FS}^-_{\underline{n}}}(\underline{\phi})$. More precisely this proof will simply start with $h(\underline{k},\underline{n},\underline{\phi})$ followed by $\mathsf{Pr}_{\mathsf{FS}^-_{\underline{n}}}(\underline{\phi})$ (that can be inferred by existential generalization). Compute $e(\mathsf{tal}(n),\mathsf{tal}(k))$, i.e., a proof of $(\mathrm{REF}_k(\mathsf{FS}_n))$. Apply modus ponens to conclude $\mathsf{Tr}_k(\underline{\phi})$. Finally, apply Theorem 2.30 to compute the proof of

$$\mathsf{Tr}_k(\underline{\phi}) \equiv \phi.$$

The concatenation of the above proofs yields a PA-proof of ϕ. ⊣

4.4. Feasible interpretability of truth theories. In Section 2.5 we gave a terse proof of Theorem 2.36; that proof did not directly link the notions of feasible reducibility with feasible interpretability, which is how we originally conceived of—and arrived at—our main results. Since interpretations, especially of the feasible variety, are

[27] In fact, this bound is even explicitly given in the arithmetical definition of $\mathsf{RT}^-_{<2n+1}$.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

of foundational and philosophical interest in connection with axiomatic theories of truth, we now explain the interpretability-theoretic perspective of our work by establishing the following result that shows that each of the truth theories $\mathsf{CT}^-[\mathsf{PA}]$, $\mathsf{KF}^-[\mathsf{PA}]$, and $\mathsf{FS}^-[\mathsf{PA}]$ can be interpreted in PA via a feasibly neat family of interpretations (the notion of a feasibly neat family of interpretations was introduced in Definition 2.44).

THEOREM 4.18 (Feasible interpretability of truth theories). *Let $\mathcal{T}$ be any of the truth theories $\mathsf{CT}^-[\mathsf{PA}]$, $\mathsf{KF}^-[\mathsf{PA}]$, or $\mathsf{FS}^-[\mathsf{PA}]$. Then there exists a feasibly neat family $\{I_n\}_{n\in\mathbb{N}} : \mathcal{T} \to \mathsf{PA}$ of interpretations.*

Note that, by Proposition 2.46, the existence of a feasibly neat family of interpretations guarantees feasible reducibility. The proof of Theorem 4.18 can be readily read-off the second proof of Theorem 2.36, which we present in this section; this second proof employs a feasible version of the Arithmetized Completeness Theorem to demonstrate that the assumptions of Theorem 2.36 about a theory $\mathcal{T}$ imply the existence of a feasibly neat family of interpretations of $\mathcal{T}$ in PA. This will make it clear that Theorem 4.18 holds since we have already verified in Sections 4.1, 4.2, and 4.3 that the assumptions of Theorem 2.36 are met when $\mathcal{T}$ is any of the truth theories $\mathsf{CT}^-[\mathsf{PA}]$, $\mathsf{KF}^-[\mathsf{PA}]$, or $\mathsf{FS}^-[\mathsf{PA}]$.

We begin with presenting a feasible version of the Arithmetized Completeness Theorem.

- In Lemma 4.19 below an *n-theory* (respectively: *n-model*, *n-full model*, *n-language*) is a theory (respectively: model, full model, language) definable by a formula of depth n, where "definable" should be understood to be in the sense of the feasible Sat_n predicates discussed in Section 2.4.

LEMMA 4.19 (Feasible Arithmetized Completeness Theorem, FACT). *There is a polynomial $p(n)$ such that $\{\psi_n\}_{n\in\mathbb{N}}$ is PA-feasibly provable, where ψ_n expresses*:

"Every consistent n-theory $\mathcal{T}$ in an n-language $\mathcal{L}$ has a $p(n)$-full model."

Moreover, there exists a P-time computable function $f(s_0, s_1)$ such that for any $l, k \in \mathbb{N}$, $f(\mathsf{tal}(l), \mathsf{tal}(k))$ is a PA-proof of the sentence expressing

"If $\mathcal{M}$ is a full l-model of a k-theory $\mathcal{T}$, then $\mathcal{T}$ is consistent."

As a corollary to the proof of the above lemma we can obtain the following proposition.

PROPOSITION 4.20 (Uniformity of FACT). *Suppose that $\phi(x, \bar{y})$ is an $\mathcal{L}_{\mathsf{PA}}$-formula (where $\bar{y}$ denotes the parameters) such that*:

$$\mathsf{PA} \vdash \text{“}\underline{\phi(x, \bar{y})} \text{ defines a theory.”}$$

Then there exists a formula $\phi'(x, \bar{y})$ such that:

$$\mathsf{PA} \vdash \text{“If } \underline{\phi} \text{ is consistent, then } \underline{\phi'} \text{ defines a model for } \underline{\phi}\text{.”} \qquad (\mathsf{ACT}_\phi)$$

Moreover, there exists a P-time computable function $f(s)$, such that $f(\phi)$ is a PA-proof of ACT_ϕ.

PROOF OF LEMMA 4.19. See the Appendix. ⊣

In addition to FACT, we also have the Feasible Compactness Theorem (the proof of which is rather obvious, as we deal here with consistency in the syntactical sense).

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

LEMMA 4.21 (Arithmetized Compactness Theorem). *There exists a P-time computable function $f(s)$ such that for every $n \in \mathbb{N}$, $f(\mathsf{tal}(n))$ is a* PA*-proof of the sentence expressing*:

"An n-theory $\mathcal{T}$ is consistent if and only if each bounded fragment of $\mathcal{T}$ is consistent."

We are now ready to present the second proof of Theorem 2.36. We include the statement here for the benefit of the reader.

THEOREM 4.22 (Theorem 2.36 redux). *Let $\mathcal{T}$ be an NP-time computable theory extending* PA*, and suppose that there is a polynomial $q(n)$ such that the family $\{\psi_n\}_{n\in\mathbb{N}}$ is polynomially* PA*-provable, where*:

$$\psi_n := \forall \phi \in \mathsf{Sent}_{\mathcal{L}_{\mathsf{PA}}} \left(\left(\mathsf{dp}(\phi) \leq \underline{n} \wedge \mathsf{Pr}_{\mathcal{T}\restriction \underline{n}}(\phi) \right) \rightarrow \mathsf{Pr}_{\mathsf{PA}\restriction \underline{q(n)}}(\phi) \right).$$

Then PA *polynomially simulates $\mathcal{T}$. Moreover, if $\mathcal{T}$ is P-time computable and $\{\psi_n\}_{n\in\mathbb{N}}$ is feasibly* PA*-provable, then $\mathcal{T}$ is feasibly reducible to* PA.

SECOND PROOF OF THEOREM 2.36, SKETCH. We will construct a feasibly neat family of interpretations $\{I_n\}_{n\in\mathbb{N}} : \mathcal{T} \rightarrow \mathsf{PA}$. Let us define the theory Φ_n:

$$x \in \Phi_n := (\mathsf{Tr}_{n+1}(x) \vee x \in \mathcal{T}) \wedge \exists y \left(\mathsf{len}(x) \leq y \wedge \mathsf{Con}^{\mathsf{Tr}_{n+1}}_{\mathcal{T}\restriction y} \right), \tag{2}$$

where $\mathsf{Con}^{\mathsf{Tr}_n}_{\mathcal{T}\restriction y}$ says that there is no proof of contradiction using as axioms sentences in $\mathcal{T} \restriction y$, or true sentences of depth n (see Remark 2.25 for an explanation).

Observe that the length of (the formula defining) Φ_n is polynomial in n and its shape depends uniformly on n. Then for every n, $\mathsf{PA} \vdash \mathsf{Con}_{\Phi_n}$ and the proof is uniform in n, so in fact there exists a polynomial $p_1(n)$ such that for every n,

$$\mathsf{PA} \vdash^{p_1(n)} \mathsf{Con}_{\Phi_n}.$$

(for the precise argument see [10], Theorem 2.37). By FACT we know that there exists a formula $\mathcal{M}_{\Phi_n}$ and a polynomial $p_2(n)$ such that:

$$\mathsf{PA} + \mathsf{Con}_{\Phi_n} \vdash^{p_2(n)} \mathcal{M}_{\Phi_n} \text{ is a full model for } \Phi_n. \tag{$**$}$$

Now I_n is defined as a relativization to $\mathcal{M}_{\Phi_n}$ i.e., a function defined on formulae of the language of $\mathcal{T}$ which preserves boolean operations such that for every relational symbol R and all terms $s_1, \ldots, s_n$ we have:

$$(R(s_1, \ldots, s_n))^{I_n} = \ulcorner R(s_1, \ldots, s_n) \urcorner \in \mathcal{M}_{\Phi_n},$$

(recall that full models are coded as elementary diagrams) and for every existential formula $\exists x \phi$,

$$(\exists x \phi)^{I_n} = \exists x \in M_{\Phi_n} \ \phi^{I_n}.$$

We next verify that the family of interpretations $\{I_n\}_{n\in\mathbb{N}}$ is polynomially neat. To check condition (a) of polynomial neatness we use contraposition. Work in PA. Assume $\neg\phi$, where ϕ is of length at most n. We will derive $(\neg\phi)^{I_n}$ via a proof whose length is bounded above by a polynomial in n. Surely, $\neg\phi$ is of length at most $n+1$. Let $r_2(n)$ be as in Theorem 2.29. Then, by provable Tarski biconditionals, with a proof of length at most $r_2(n)$ we conclude

$$\mathsf{Tr}_{n+1}(\neg\phi).$$

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

We show that this implies $\Phi_n(\neg\phi)$. By our assumption and Lemma 2.37 we obtain a polynomial $p_3(n,k)$ such that:

$$\mathsf{PA} \vdash^{p_3(n,n)} \mathsf{Con}^{\mathsf{Tr}_{n+1}}_{\mathcal{T}\restriction\underline{n+1}},$$

which directly implies that $\neg\phi$ belongs to Φ_n. Consequently PA proves:

$$\mathcal{M}_{\Phi_n} \models \neg\phi,$$

with a proof of length $O(p_1(n) + p_2(n) + q(n) + p_3(n,n))$. Now using at most $|\phi|$ many steps involving formulae of length polynomial in n we obtain

$$(\neg\phi)^{I_n}.$$

To show that $\{I_n\}_{n\in\mathbb{N}}$ satisfies condition (b) of polynomial neatness it suffices to show that for some polynomial $p(n,k)$ and each $k \in \mathbb{N}$,

$$\mathsf{PA} \vdash^{p(n,k)} \psi^{I_k},$$

for every sentence ψ in $\mathcal{T} \restriction n$. We use polynomial numerability of $\mathcal{T} \restriction n$ to guarantee that there is a polynomial p_4 such that for all $n \in \mathbb{N}$ and all $\psi \in \mathcal{T} \restriction n$,

$$\mathsf{PA} \vdash^{p_4(n)} \psi \in \mathcal{T} \restriction \underline{n}.$$

Now, as previously, we have

$$\mathsf{PA} \vdash^{p_3(n,k)} \mathsf{Con}^{\mathsf{Tr}_k}_{\mathcal{T}\restriction\underline{n}}.$$

Hence, adding a few more steps, we also have

$$\mathsf{PA} \vdash^{p_5(n,k)} \psi \in \Phi_k.$$

Then, as previously, we check that ψ^{I_k} is satisfied.

The "moreover" part holds since if $\mathcal{T}$ is P-time computable, then we can feasibly find a PA-proof witnessing that

$$\mathsf{PA} \vdash \phi \in \mathcal{T} \restriction \underline{n}.$$

The rest of the steps are fully analogous. ⊣

§5. Open questions. The proofs of our main results in Section 4 suggest that the answers to the following questions are both in the positive; we pose them here since definitive positive answers to them require a number of technical verifications that are yet to be carried out.

QUESTION A. Is the conservativity of $\mathsf{CT}^-[\mathsf{PA}]$, $\mathsf{KF}^-[\mathsf{PA}]$, and $\mathsf{FS}^-[\mathsf{PA}]$ over PA provable in Buss's system S^1_2? Note that the proofs given in this article make it clear that the answer to this question is positive if Buss's system S^1_2 is strengthened to $\mathsf{I}\Delta_0 + \mathsf{Exp}$.

QUESTION B. Suppose B is a sequential theory that is inductive; i.e., the scheme of induction over the natural numbers of B is provable in B. Are $\mathsf{CT}^-[\mathsf{B}]$, $\mathsf{KF}^-[\mathsf{B}]$, and $\mathsf{FS}^-[\mathsf{B}]$ feasibly reducible to B?

§6. Appendix. In order to minimally distract the reader from the main flow of the argument, we decided to relegate some of the technical verifications to this Appendix.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

6.1. Feasible reflexivity. In Section 2.5, we proved Theorem 2.36 which is the technical core of our article, and which provides us with a uniform way of obtaining polynomial simulations and feasible reductions. We presented two proofs of that theorem. The first one used the following lemma, which was originally presented as Lemma 2.37:

LEMMA 6.1. *There is a P-time computable function* $f(s_0, s_1)$ *such that for every* $n, k \in \mathbb{N}$, $f(\mathsf{tal}(n), \mathsf{tal}(k))$ *is a* PA*-proof of*

$$\forall \phi \in \mathsf{Sent}_{\mathcal{L}_{\mathsf{PA}}}\big(\big(\mathsf{dp}(\phi) \leq \underline{k} \wedge \mathsf{Pr}_{\mathsf{PA}\restriction \underline{n}}(\phi)\big) \to \mathsf{Tr}_k(\phi)\big).$$

The above result states that we can feasibly find PA-proofs of the uniform reflection for bounded fragments of PA. In the proof, we assumed that the statement holds for the axioms in these fragments, and that arithmetical satisfaction predicates enjoy certain regularity properties. This is in order to avoid taking universal closures of axioms. Below, we formulate these results in a precise manner and prove them with the help of the following definitions:

- If α is a valuation and v is in the domain of α, then by $\alpha[v \mapsto x]$ we mean a valuation α' that is the same as α except for the variable v, whose value is x.
- If y is a formula, then $\mathsf{Ind}(y, v)$ denotes the instantiation of the induction scheme (with parameters) with formula y w.r.t. v, i.e., the following formula

$$\big(y[0/v] \wedge \forall v\, \big(y \to y[S(v)/v]\big) \to \forall v\, y\big)\,.$$

Let us observe that, living inside PA, we know that every object can be named by a closed term. Proposition 6.2 says that for every formula ϕ being satisfied by a sequence y is equivalent to the truth of the sentence $\phi[y]$.[28] The proof is routine; it uses induction in PA on the complexity of formulae, and the feasibly PA-provable Tarski conditions for Sat_n predicates.

PROPOSITION 6.2. *Each of the families* $\{\phi_n\}_{n\in\mathbb{N}}$ *and* $\{\phi'_n\}_{n\in\mathbb{N}}$ *of formulae* (*asserting regularity properties for* Sat_n *predicates*) *is feasibly* PA*-provable, where*:

1. $\phi_n := \Big(\big(\mathsf{dp}(y) \leq \underline{n} \wedge \alpha' = \alpha[v \mapsto S(\alpha(v))]\big) \to \mathsf{Sat}_n(y, \alpha') \equiv \mathsf{Sat}_n(y[S(v)/v], \alpha)\Big)$.
2. $\phi'_n := \Big(\big(\mathsf{dp}(y) \leq \underline{n} \wedge \alpha' = \alpha[v \mapsto z]\big) \to \mathsf{Sat}_n(y, \alpha') \equiv \mathsf{Sat}_n(y[\underline{z}/v], \alpha)\Big)$.

PROPOSITION 6.3 (Essentially Pudlák, [17]). *Each of the families* $\{\psi_n\}_{n\in\mathbb{N}}$ *and* $\{\psi'_n\}_{n\in\mathbb{N}}$ *of formulae is feasibly* PA*-provable, where*:

1. $\psi_n := \Big(\mathsf{dp}(x) \leq \underline{n} \wedge \text{“}x \text{ is a logical axiom”} \wedge \alpha \in \mathsf{Asn}(x) \to \mathsf{Sat}_n(x, \alpha)\Big)$.
2. $\psi'_n := \Big(\mathsf{dp}(y) \leq \underline{n} \wedge \text{“}y \text{ is of the form } x \to z\text{”} \wedge \alpha \in \mathsf{Asn}(y) \wedge \mathsf{Sat}_n(y, \alpha) \wedge \mathsf{Sat}_n(x, \alpha) \to \mathsf{Sat}_n(z, \alpha)\Big)$.

Now we prove that the family of sentences of the form “all PA-axioms of induction of depth n are true” is feasibly PA-provable. For the sake of simplicity we assume that PA is axiomatized by the induction scheme with free variables treated as parameters.

[28] For the notation $\phi[y]$, see Definition 2.4.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

PROPOSITION 6.4. *The family* $\{\theta_n\}_{n\in\mathbb{N}}$ *of formulae is feasibly* PA*-provable, where*:

$$\theta_n := \Big(\forall y \in \mathsf{Form}^1_{\mathcal{L}_{\mathsf{PA}}} \forall v \in \mathsf{Var}\,\big(v \in \mathsf{FV}(y) \wedge \mathsf{dp}(\mathsf{Ind}(y,v)) \leq \underline{n} \wedge \alpha \in \mathsf{Asn}(y) \longrightarrow \mathsf{Sat}_n(\mathsf{Ind}(y),\alpha)\big)\Big).$$

PROOF OF PROPOSITION 6.4. For the purposes of this proof, we say that y is *small* if $\mathsf{dp}(\mathsf{Ind}(y,v)) \leq \underline{n}$. Let $\phi_1(y,v,\alpha)$ be the formula that expresses:

"y is a small formula such that v is a free variable of y, and α is an assignment for y."

Moreover, let $\phi_2(y,v,\alpha,x)$ abbreviate

$$\exists \alpha'\big(\alpha' = \alpha[v \mapsto x] \wedge \mathsf{Sat}_n(y,\alpha')\big),$$

and let $\phi(x,v,y,\alpha) = \phi_1(y,v,\alpha) \wedge \phi_2(y,v,\alpha,x)$. The idea is that α encodes a sequence of parameters used in the induction and x is the varying value assigned to the variable v while proving $\forall v y$ by induction. We work in PA. We start with $\mathsf{Ind}(\phi(x,v,y,\alpha),x)$, which is an axiom whose length is bounded above by a polynomial in n (since the length of $\phi(x,v,y,\alpha)$ is bounded above by a polynomial in n). Using a few transformations (the number of which is independent of n) we obtain

$$\forall x,v,y,\alpha\ \big(\phi_1(y,v,\alpha) \to \mathsf{Ind}(\phi_2(y,v,\alpha,x),x)\big).$$

Let us look at $\mathsf{Ind}(\phi_2(y,v,\alpha,x),x)$. Observe that by Proposition 6.2 we have:

1. $\phi_2(y,v,\alpha,0)$ is equivalent to $\mathsf{Sat}_n(y[\underline{0}/v],\alpha)$ and
2. $\phi_2(y,v,\alpha,S(x))$ is equivalent to $\mathsf{Sat}_n(y[S(v)/v],\alpha)$.

Hence $\mathsf{Ind}(\phi_2(y,v,\alpha,x),x)$ implies

$$\mathsf{Sat}_n(y[\underline{0}/v],\alpha) \wedge \forall x\,\big(\mathsf{Sat}_n(y,\alpha[v\mapsto x]) \to \mathsf{Sat}_n(y[S(v)/v],\alpha[v\mapsto x])\big) \longrightarrow \longrightarrow \forall x\ \mathsf{Sat}_n(y,\alpha[v\mapsto x]).$$

Now by the compositional axioms for Sat_n the above is equivalent to

$$\mathsf{Sat}_n(\mathsf{Ind}(y,v)).$$ ⊣

6.2. Congruence Lemma. We sketch the proof of the following lemma from Section 4.1.

LEMMA 6.5 (Congruence lemma). *For all* ϕ, ϕ', ψ' *the following holds*:

$$\big(\phi \lhd \phi' \wedge \phi' \approx^{\mathcal{M}} \psi'\big) \Rightarrow \exists \psi\ \ \big(\psi \lhd \psi' \wedge \psi \approx^{\mathcal{M}} \phi\big). \tag{C}$$

Sketch of the proof. We prove the lemma by distinguishing cases depending on the main connective or quantifier in ϕ'. The only nontrivial step is the one for $\exists$. Assume $\phi' = \exists v\ \phi$. Then $\psi' = \exists v\ \psi$. Take $\widehat{\phi'}$ $(= \widehat{\psi'})$, which, by definition, is of the form $\exists v\ \eta$. In η replace all the occurrences of maximal terms in η (i.e., the ones that do not occur within a term) which only contain free variables (in η) with fresh variables, without using the same variable twice. Then rename the free variables of the resulting formula according to the procedure adopted in condition 4. of the definition of the term trivialization. In this way we obtain the term trivialization of both ψ and ϕ. ⊣

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

6.3. FACT. In Section 4.4, we presented an alternative proof of Theorem 2.36. This alternative proof used the fact that the arithmetized completeness theorem can be proved with a proof of length polynomial in the length of the formula defining $\mathcal{T}$. This was stated without proof as Lemma 4.19. In this subsection, we provide a proof of this Lemma.

LEMMA 6.6 (Feasible Arithmetized Completeness Theorem, FACT). *There is a polynomial $p(n)$ such that $\{\psi_n\}_{n\in\mathbb{N}}$ is* PA*-feasibly provable, where ψ_n expresses*:

"Every consistent n-theory $\mathcal{T}$ in an n-language $\mathcal{L}$ has a $p(n)$-full model."

Moreover, there exists a P-time computable function $f(s_0, s_1)$ such that for any $l, k \in \mathbb{N}$, $f(\mathsf{tal}(l), \mathsf{tal}(k))$ is a PA*-proof of the sentence expressing*:

"If $\mathcal{M}$ is a full l-model of a k-theory $\mathcal{T}$, then $\mathcal{T}$ is consistent."

PROOF. We follow the "leftmost branch" strategy; the proof is routine but we present it for thoroughness. Fix a a formula $\theta(x)$ of depth n that defines a consistent $\mathcal{L}$-theory. Note that we need not care about the rise in Σ_k-complexity of the formula defining a model for $\theta(x)$ as long as the construction of the relevant formula can be feasibly computed in the input $\mathsf{tal}(n)$. Let $\mathsf{Form}^H_{\mathcal{L}}$ be the set of formulae of $\mathcal{L}$ enriched with Henkin constants (we denote the Henkin constant for a formula ϕ with c_ϕ, and assume that the function $\phi \mapsto c_\phi$ is Δ_1).

Working towards building a complete and consistent Henkin extension, let $\theta'(x)$ be the formula that defines the Henkin extension of the theory defined by $\theta(x)$, i.e.,

$$\theta'(x) := \Big(\theta(x) \vee \exists \phi, v \Big(v \in \mathsf{Var} \wedge \phi \in \mathsf{Form}^H_{\mathcal{L}} \wedge x = (\exists v \phi \to \phi[c_\phi/v])\Big)\Big).$$

Note that the depth of $\theta'(x)$ is polynomial in n, hence to talk about the set it defines we can use a truth predicate of length polynomial in n. We check that $\mathsf{Con}_{\theta'(x)}$ holds: each proof of $\exists x(x \neq x)$ from the axioms of $\theta'(x)$ can be transformed into a $\theta(x)$-proof of

$$\neg \bigwedge_{j\le a} (\exists v_{i_j} \phi_j \to \phi_j[c_{\phi_j}/v_{i_j}]).$$

Then we check that the provability of the above entails that $\theta(x)$ proves:

$$\bigvee_{j\le a} (\exists v_{i_j} \phi_j \wedge \forall v_{i_j} \neg \phi_{i_j}),$$

which contradicts the consistency of θ. Note that the above argument is uniform in θ.

Let $\sigma(i) = x$ be an enumeration of $\mathsf{Form}^H_{\mathcal{L}}$ (i.e., $\sigma(i) \in \mathsf{Form}^H_{\mathcal{L}}$ for each number i, and every formula $x \in \mathsf{Form}^H_{\mathcal{L}}$ is of the form $\sigma(i)$ for some number i). For any binary sequence τ of length y let $\mathsf{enum}(\sigma, \tau, y)$ be the theory defined by:

$$x \in \mathsf{enum}(\sigma, \tau, y) := \exists i < y \left(\big(\tau(i) = 0 \wedge x = \sigma(i)\big) \vee \big(\tau(i) = 1 \wedge x = \neg\sigma(i)\big)\right).$$

Note that $\mathsf{enum}(\sigma, \tau, y)$ describes the theory obtained by enumerating the first y formulae given by σ, and then adding a negation in front of the i-th element if $\tau(i) = 1$. Let $\theta^H(x)$ be the formula that asserts that there exists a unique pair (y, τ) that satisfies the following properties:

1. τ is a binary sequence of length y;
2. $\forall i < y \ \big(\tau(i) = 0 \iff \mathsf{Con}_{\theta' + \mathsf{enum}(\sigma,\tau,i) + \sigma(i)}\big)$; and

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

3. $\left(\mathsf{Con}_{\theta'+\mathsf{enum}(\sigma,\tau,y)+\sigma(y)} \wedge x = \sigma(y)\right) \vee \left(\neg\mathsf{Con}_{\theta'+\mathsf{enum}(\sigma,\tau,y)+\sigma(y)} \wedge x = \neg\sigma(y)\right)$.

Thus $\theta^H(x)$ describes the "leftmost consistent branch" of the binary tree each node of which represents a finite approximation to a completion of the theory defined by θ', as in the usual proof of the arithmetical completeness theorem. Once again the depth of $\theta^H(x)$ is bounded above by a polynomial function applied to n, and this polynomial does not depend on the initial choice of θ and $\mathcal{L}$. We show that θ^H is a complete and consistent theory with Henkin sentences (which, according to our definitions is the same as a full model). The whole proof was carried out uniformly in n and can be produced by a P-time computable function f.

To prove the "moreover" part, we show by induction on the lengths of proofs that any l-full model (which, recall, is the same as a complete consistent Henkinized theory) is closed under reasoning in first-order logic. This argument is carried out uniformly, except that we use different feasible satisfaction predicates depending on the complexity of the model. ⊣

6.4. A glossary of technical notions. This article contains a fairly large number of technical definitions; here we enclose a glossary of such terms for the benefit of the reader.

- $x \in \mathsf{Asn}(y)$ means that x is an assignment for a formula or a term y (or for a set of terms or formulae y), i.e., x is a function whose domain includes the free variables of y (or whose domain includes the free variables of each element of y); see Definition 2.3 and Convention 2.5.
- $x \in \mathsf{ClTerm}_{\mathcal{L}}$ means that x is a closed term of the language $\mathcal{L}$; see Definition 2.3 and Convention 2.5.
- $x \in \mathsf{ClTermSeq}_{\mathcal{L}}$ means that x is a sequence of closed terms of the language $\mathcal{L}$; see Definition 2.3 and Convention 2.5.
- $\mathsf{Con}_{\mathcal{T}}$ is an arithmetized consistency statement for $\mathcal{T}$; see Corollary 2.24.
- CT^- is the compositional theory of truth over PA; see Definition 3.1.
- $\mathsf{CT}^-[\mathsf{B}]$ is the compositional theory of truth over a theory B extending $\mathsf{I}\Delta_0+\mathsf{Exp}$; see Definition 3.1.
- $\mathsf{dp}(\phi)$ is the syntactic depth of a formula ϕ; see Definition 2.27.
- $\mathsf{ElDiag}(\mathcal{M})$ (elementary diagram of $\mathcal{M}$) is the same as a full model $\mathcal{M}$, this notation is used when $\mathcal{M}$ is viewed as a complete Henkinized theory, rather than a structure; see Definition 2.7.
- $\mathsf{Form}_{\mathcal{L}}(x)$ means that x is a formula of the language $\mathcal{L}$; see Definition 2.3 and Convention 2.5.
- $\mathsf{Form}^{\leq 1}_{\mathcal{L}}(x)$ means that x is a formula of the language $\mathcal{L}$ with at most one free variable; see Definition 2.3 and Convention 2.5.
- $\mathsf{Form}^{1}_{\mathcal{L}}(x)$ means that x is a formula of the language $\mathcal{L}$ with exactly one free variable; see Definition 2.3 and Convention 2.5.
- FS^- is the Friedman–Sheard theory of untyped truth over PA without extra induction; see Definition 3.3.
- $\mathsf{FS}^-[\mathsf{B}]$ is the Friedman–Sheard untyped theory of truth over a theory B extending $\mathsf{I}\Delta_0 + \mathsf{Exp}$ without extra induction; see Definition 3.3.
- $\mathsf{FV}(x, y)$ means that y is a free variable of x; see Definition 2.3 and Convention 2.5.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

- $\mathsf{FVSeq}(x, y)$ means that y is a coded sequence whose elements are (some) free variables of x; see Definition 2.3 and Convention 2.5.
- KF^- is the Kripke–Feferman theory of untyped truth over PA without extra induction; see Definition 3.2.
- $\mathsf{KF}^-[\mathsf{B}]$ is the Kripke–Feferman theory of untyped truth over a theory B extending $\mathsf{I}\Delta_0 + \mathsf{Exp}$ without extra induction; see Definition 3.2.
- $\mathsf{len}(s)$ is the length of a *sequence* of strings s; and $|s|$ is the length of a string s; see Definition 2.3 and Convention 2.5.
- $\underline{n}$ is the numeral representing n using the binary expansion of n; see Definition 2.2.
- n-theory (respectively: n-model, n-full model) is a theory (respectively: model, full model) defined with a formula of depth n (via a feasible Sat_n predicate); see the bullet item before Lemma 4.19.
- $\mathsf{Pr}_{\mathcal{T}}(y)$ means that there exists a proof of y in the theory $\mathcal{T}$; see Corollary 2.24.
- $\mathsf{Proof}_{\mathcal{T}}(m, n)$ means that m is a proof of n from the theory $\mathcal{T}$; see Corollary 2.24.
- $\mathsf{Sent}_{\mathcal{L}}(x)$ means that x is a sentence of the language $\mathcal{L}$; see Definition 2.3 and Convention 2.5.
- $\mathsf{tal}(n)$ is the tally numeral representing n; see Definition 2.2.
- $\mathsf{Term}_{\mathcal{L}}(x)$ means that x is a term of the language $\mathcal{L}$; see Definition 2.3 and Convention 2.5.
- $\mathsf{TermSeq}_{\mathcal{L}}(x)$ means that x is a sequence of terms of the language $\mathcal{L}$; see Definition 2.3 and Convention 2.5.
- $\mathsf{Var}(x)$ means that x is a variable; see Definition 2.3.
- $\beta \sim_v \alpha$ means α and β are functions, v is a variable and $\alpha(w) = \beta(w)$ for all variables w, possibly except for v which also might belong only to the domain of β; see Definition 2.3.
- $\phi[\alpha]$ is a formula ϕ with the numeral $\underline{\alpha(v)}$ substituted for every occurrence of v for every free variable v of ϕ; see Definition 2.4.
- $\phi[s/v]$ denotes the formula ϕ with the term s substituted for the variable v; see Definition 2.4.
- $\| \phi \|_{\mathcal{T}}$ is the length of the shortest proof of ϕ in $\mathcal{T}$; see Definition 2.13.
- $\phi \lhd \psi$ means that ϕ is an immediate subformula of ψ; this is a key notion in the proof of Lemma 4.8.
- $\widehat{\phi}$ is the term trivialization of ϕ; see remarks preceding Lemma 4.9.
- $\mathcal{M} \preceq_{\mathcal{L}} \mathcal{N}$ means that $\mathcal{M}$ is an elementary submodel of $\mathcal{N}$ with respect to formulae from the language $\mathcal{L}$; see Definition 2.8 (together with the subsequent remarks), and Definition 2.9.
- t^{α} is the value of the term t in which every free variable v has been evaluated to $\alpha(v)$; see Definition 2.4.
- $\mathcal{T} \vdash^n \phi$ means that there is a proof in $\mathcal{T}$ of ϕ whose length is at most n; see Definition 2.13.
- $\mathcal{T} \restriction n$ for a theory $\mathcal{T}$ means the set of axioms of $\mathcal{T}$ of length at most n; see Definition 2.23.

https://doi.org/10.1017/jsl.2019.24 Published online by Cambridge University Press

- $x^\circ = y$ means that y is the value of the term x; see Definition 2.3 and Convention 2.5.

Acknowledgments. We wish to record our gratitude to Albert Visser for inspiring us to carry out this project through many enlightening conversations and emails. It was Albert's joint work with Ali Enayat on the conservativity of $\mathsf{CT}^-[\mathsf{PA}]$ over PA, and on the interpretability of $\mathsf{CT}^-[\mathsf{PA}]$ in PA that naturally led to our investigation of feasible reducibility of $\mathsf{CT}^-[\mathsf{PA}]$ and other truth theories to PA. We are also indebted to Matt Kaufmann for contributing a key idea that led to the particular arithmetization of the Enayat–Visser construction that is employed in the proof of Theorem 4.1, to Fedor Pakhomov for suggesting Theorem 2.36 as the main technical tool for proving our results in a transparent and efficient fashion, and to Leszek Kołodziejczyk, for steering us away from an incorrect way of stating some of our results that pertain to polynomial time computations. We would also like to thank Jeremy Avigad and Tin Lok Wong for their valuable comments on the original manuscript, and to the anonymous referee for a most helpful report. The work of Mateusz Łełyk and Bartosz Wcisło on this article was supported by The National Science Centre, Poland (NCN), 2017/27/B/HS1/01830.

REFERENCES

[1] P. Caldon and A. Ignjatović, *On mathematical instrumentalism*, this Journal, vol. 70 (2005), no. 13, pp. 778–794.

[2] A. Cantini, *Notes on formal theories of truth*. ***Zeitschrift für Mathematische Logik und Grundlagen der Mathematik***, vol. 35 (1989), no. 1, pp. 97–130.

[3] C. Cieśliński, ***The Epistemic Lightness of Truth: Deflationism and its Logic***, Cambridge University Press, Cambridge, 2018.

[4] C. Cieśliński, M. Łełyk, and B. Wcisło, *Models of* PT^- *with internal induction for total formulae*. ***The Review of Symbolic Logic***, vol. 10 (2017), no. 1, pp. 187–202.

[5] A. Enayat, Problem 1 in "A list of open problems". Submitted to the conference ***Model Theory and Proof Theory of Arithmetic*** (2012). https://www.impan.pl/∼kz/KR/slides/KR_OpenProblems.pdf.

[6] 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-Verlag, Berlin, 2015, pp. 321–335.

[7] S. Feferman, *Reflecting on incompleteness*, this Journal, vol. 56 (1991), no. 1, pp. 1–49.

[8] M. Fischer, *Truth and speed-up*. ***The Review of Symbolic Logic***, vol. 7 (2014), no. 2, pp. 319–340.

[9] H. Friedman and M. Sheard, *An axiomatic approach to self-referential truth*. ***Annals of Pure and Applied Logic***, vol. 33 (1987), pp. 1–21.

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

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

[12] R. Kaye, ***Models of Peano Arithmetic***, Clarendon Press, Oxford, 1991.

[13] R. Kossak and B. Wcisło, *Disjunctions with stopping condition*. ***Bulletin of Symbolic Logic***, accepted.

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

[15] S. Kripke, *Outline of a theory of truth*. ***The Journal of Philosophy***, vol. 72 (1975), no. 19, pp. 690–716.

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

[17] P. Pudlák, *On the length of proofs of finitistic consistency statements in first order theories*, ***Logic Colloquium 84*** (J. B. Paris, A. Wilkie, and G. Wilmers, editors), 1986 .

[18] ———, The lengths of proofs, ***Handbook of Proof Theory*** (S. R. Buss, editor), Elsevier, Amsterdam, 1998, pp. 547–642.

[19] S. G. Simpson, ***Subsystems of Second Order Arithmetic***, second ed., Perspectives in Logic, Cambridge University Press, Cambridge, 2009.

[20] Vann McGee, *How truthlike can a predicate be? A negative result*. ***Journal of Philosophical Logic***, vol. 14 (1985), no. 4, pp. 399–410.

[21] R. Verbrugge, ***Feasible interpretations***, Ph.D thesis, University of Amsterdam, 1993.

[22] 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. 455–480.

DEPARTMENT OF PHILOSOPHY, LINGUISTICS, AND THEORY OF SCIENCE
UNIVERSITY OF GOTHENBURG, BOX 200
405 30 GOTHENBURG, SWEDEN
E-mail: ali.enayat@gu.se

INSTITUTE OF PHILOSOPHY
UNIVERSITY OF WARSAW
UL. KRAKOWSKIE PRZEDMIEŚCIE 3
00-927 WARSZAWA, POLAND
E-mail: mlelyk@uw.edu.pl

INSTITUTE OF MATHEMATICS, POLISH ACADEMY OF SCIENCES
UL. ŚNIADECKICH 8
00-656 WARSZAWA, POLAND
E-mail: b.wcislo@impan.pl