Chapter 17
Interpretations over Heyting's arithmetic

Albert Visser

July 1995

Abstract

In this paper we experiment with a rather general notion of "interpretation in constructive arithmetical theories". We prove a number of elementary properties of the notion introduced. We prove a number of negative results for interpretations that commute with disjunction. These negative results diverge markedly from what is known in the classical case. We briefly consider interpetations in formula classes of bounded complexity. In an appendix we show how to do (the interpretation version of) the Henkin construction for Intuitionistic Predicate Logic inside Peano Arithmetic. This construction cannot be given in Heyting's Arithmetic.

1 Introduction

Relative interpretability has been well studied for classical arithmetical theories. What happens when we try to transfer the insights obtained classically to constructive theories? It turns out that the situation in the constructive case is markedly different. One of the pillars of the study of interpretations in the classical case, the formalized Henkin construction, cannot be constructivized. On the other hand there is a wealth of translations – like Kleene realizability, the Friedman translation, the double negation translation – that we are tempted to grant the honorific *interpretation*. The role of these interpretations is far more central to the study of theories like Heyting Arithmetic, than relative interpretations are to the study of Peano Arithmetic. No systematic study has been made of these interpretations. What is worse: no general definition has emerged of what an 'interpretation in a constructive theory' is.

It is our impression that it is too early to attempt a definition of interpretation for the constructive case. Still, it is already possible to say a lot without such a definition. Our approach is to study certain global properties that are plausible for interpretations.

The properties we consider have, mainly, to do with *the way the external numerical variables reappear inside the interpretation*. To understand what this means, let's turn back to the classical case for a moment. Consider a relative interpretation $\mathcal{K}$ of a theory U in a theory T. One heuristically fruitful way

to view $\mathcal{K}$ is to look at it as a definable mapping of models of T to models of U. For every model $\mathfrak{M}$ of T, the interpretation provides an *internal* model $\mathfrak{N}$ for U. If T extends Peano Arithmetic and if U contains some number theory, then the numbers of $\mathfrak{M}$, the *external* numbers, can be mapped, definably and uniformly, on the numbers of $\mathfrak{N}$, the internal numbers. Certain properties the external numbers have are reflected internally. E.g., if $x \neq y$ externally then the internal counterparts of x and y will be internally different, etcetera. In the constructive case our metaphor does not work: an interpretation is not always a mapping of models. Still it is useful to think of an interpretation *as if* it were such a mapping. We will set up our notion of interpretation in such a way that e.g. $(P(x))^{\mathcal{K}}$ means: the counterpart of external x has property P inside $\mathcal{K}$. Under this convention, the external-internal property, we mentioned, can be formulated as: $T \vdash x \neq y \rightarrow (x \neq y)^{\mathcal{K}}$. The external-internal properties combine in certain ways with other properties, like commutation with certain logical connectives.

The properties we consider are such that most known interpretations satisfy them. However, even if the present approach is 'top-down', we did not aim at full generality. For one thing, we did not consider other languages than those of ordinary predicate logic. For another thing, the properties of the internal appearance of external numerical variables studied here, are plausible only for theories having full induction, since they are motivated by the intuition that the external numbers are 'an initial segment' of the internal ones.

A spectacular difference between the constructive case and the classical case with relative interpretations, is as follows. If over Peano Arithmetic an interpretation is provably standard, then it is a faithful interpretation of Peano itself. The interpretation maps models to models isomorphic to themselves. On the other hand, the most salient examples over interpretations over Heyting Arithmetic – Kleene realizability, the double negation translation, the provability translation – keep the numbers standard, but engender new priciples – in our examples, respectively: ECT_0, Excluded Third, the Completeness principle. Moreover, if an interpretation commutes with disjunction and falsum, in a theory like HA then – as we will show – it cannot provide non-standard elements. Thus in classical arithmetic interesting interpretations are always non-standard, in constructive arithmetic interesting interpretations are quite often standard.

The contents of the paper are as follows. We provide the necessary preliminaries in section 2. Section 3 introduces our notion of 'interpretation' and provides some examples. In section 4 we verify in some detail the basic properties of interpretations. In section 5, we prove a number of limitative results in Rosser style. Specifically we show that if interpretations commute with disjunction and falsum, then they cannot be non-standard (in suitable constructive theories). Analogously, we show that interpretations that commute with disjunction and falsum are Π_2-conservative. In section 6 we have a look at interpretations into a bounded formula class. We provide an example of a bounded interpretation and show that in HA there are no primitive recursive bounded interpretations that commute with implication, disjunction and falsum. In appendix A, we show that the Henkin construction for Kripke models of constructive predicate logic

can be performed in Peano Arithmetic.

2 Preliminaries

2.1 Theories

The language of arithmetic, $\mathfrak{L}_{\mathsf{Ar}}$, contains, apart from $=$, the following symbols: $0, S, +, .$. The theory Ar is the $\mathfrak{L}_{\mathsf{Ar}}$-theory given by constructive logic plus the following principles:

Ar1 $Sx = Sy \to x = y$,

Ar2 $x + 0 = x$, $x + Sy = S(x + y)$,

Ar3 $x.0 = 0$, $x.Sy = x.y + x$,

Ar4 $x + Sy \neq x$

Ar5 $x = 0 \vee \exists y\, x = Sy$.

We write: $x < y :\leftrightarrow \exists u\, x + Su = y$. Evidently Ar verifies: $\neg x < x$, $y < Sy$, $x < y \to x < Sy$, $x < Sy \leftrightarrow (x < y \vee x = y)$.

The theory i-$I\Delta_0$ is the constructive counterpart of $I\Delta_0$, i.e., it is the theory in $\mathfrak{L}_{\mathsf{Ar}}$ containing Ar, plus induction for Δ_0-formulas. Exp is the axiom (correctly) expressing that exponentiation is total. (The usual results on the (verifiable) representability of the graph of exponentiation by a Δ_0-formula are easily seen to go through in i-$I\Delta_0$.) i-$I\Delta_0 + \mathsf{Exp}$ is finitely axiomatizable. We will work with its finitely axiomatized version. Another axiom that we will meet a few times is the axiom Ω_1, which says that the function $\omega_1(x) := 2^{\log(x)^2}$ is total.

For all languages $\mathfrak{U}$ for predicate logic, that we consider in this paper, we will assume that there is a relative interpretation (in the classical sense) $\mathfrak{Ar}_{\mathfrak{U}}$ of $\mathfrak{L}_{\mathsf{Ar}}$ into $\mathfrak{U}$. We consider this interpretation to be given with the language, so strictly speaking when we talk about $\mathfrak{U}$, we mean the pair $\langle \mathfrak{U}, \mathfrak{Ar}_{\mathfrak{U}} \rangle$, rather than $\mathfrak{U}$ proper. In practice, we will simply treat $\mathfrak{U}$ as a language, which has $\mathfrak{L}_{\mathsf{Ar}}$ as (designated) sublanguage. We will only consider theories that verify Ar on the designated numbers. The variables that we exhibit as $x, y, z, \ldots$ in $\mathfrak{U}$ will always range over the designated numbers.

2.2 Provability

We use underlining for our external numeral function. (In almost all cases, we will suppress the underlining, except at places where there is a real possibility of confusion.) *num* is the arithmetization of the numeral function. We will use $\Box_T$ for the formalization of provability in T. Suppose A is a formula with variables $x_1, \ldots, x_n$. We write $\#A$ and $\#t$, for: the Gödelnumber of A, respectively, t. We define $(\#A)\{(\#x_1) := x_1, \ldots, (\#x_n) := x_n\}$ or, briefly, $(\#A)\{(\#\mathbf{x}) := \mathbf{x}\}$ as: the Gödelnumber of the result of substituting the numerals of the $\mathbf{x}$'s for the

variables in A. (We use *binary numerals*, since their use is more appropriate in the context of weak theories.) Similarly, we can define $(\#A)\{(\#\mathbf{x}) := \mathbf{t}\}$, where $\mathbf{t}$ is any sequence of terms of the language. $\Box_T A(\mathbf{x})$ means $\mathsf{Prov}_T((\underline{\#A})\{(\underline{\#\mathbf{x}}) := \mathbf{x}\})$, where Prov_T is the arithmetization of the provability predicate of T and where $\mathbf{x}$ contains all free variables of A. (We need to arithmetize the function $(\#A)\{(\#\mathbf{x}) := \mathbf{x}\}$ to make sense of this.)

We illustrate the above by an example. We suppose Gödelnumbers are assigned as follows: $(\mapsto 11$, $) \mapsto 12$, $= \mapsto 15$, $S \mapsto 8$, $0 \mapsto 3$, $+ \mapsto 19$. Let $*$ be an arithmetization of the syntactical operation of concatenation. We have e.g.:

$$\mathsf{HA} \vdash num(\underline{3}) = \underline{8} * \underline{8} * \underline{8} * \underline{3}.$$

And:

$$\Box_T(x = x) \text{ means: } \mathsf{Prov}_T(\underline{11} * num(x) * \underline{15} * num(x) * \underline{12}).$$

Our notational convention evidently introduces a scope ambiguity. What is $\Box_T((x + y) = z)$ going to mean?

a) $\mathsf{Prov}_T(\underline{11} * num(x + y) * \underline{15} * num(z) * \underline{12})$ (wide scope) or:

b) $\mathsf{Prov}_T(\underline{11} * \underline{11} * num(x) * \underline{19} * num(y) * \underline{12} * \underline{15} * num(z) * \underline{12})$ (small scope)

For definiteness we stipulate that we always use the small scope reading. Fortunately by standard metamathematical results, we know that as long as the terms we employ stand for T-provably total recursive functions the different readings are provably equivalent. So (a) and (b) are provably equivalent. In this paper we will only employ terms for primitive recursive functions, so the ambiguity is mostly harmless. Note that in the context of the use of the Rosser ordering (see subsection 2.3 below) strictly speaking the choice still makes a difference – but the difference will always be inessential for the results we prove.

We write T_n for the theory axiomatized by the finitely many axioms of $i\text{-}I\Delta_0 + \mathsf{Exp}$, plus the axioms of T, which are smaller than n in the standard Gödelnumbering. We write $\vdash_{T,n} A$, or $T \vdash_n A$, for provability of A in T_n and $\mathsf{Prov}_{T,n}$ for the formalization of $\vdash_{T,n}$. We consider $\mathsf{Prov}_{T,n}$ as a form of *restricted provability* in T. The following well-known fact is quite important:

Fact 2.1 Suppose T is an RE extension of HA in $\mathfrak{L}_{\mathsf{Ar}}$. Then T is essentially reflexive (verifiably in HA). I.e. we have, for all n and for all formulas A with free variables $\mathbf{x}$, $T \vdash \forall \mathbf{x}\,(\Box_{T,n} A \to A)$. And, using $\mathsf{UC}(B)$ for *the universal closure of B, for all variables except x*, even:

$$\mathsf{HA} \vdash \forall x \forall A \in \mathsf{FOR}\ \ \Box_T \mathsf{UC}(\Box_{T,x} A \to A).$$

∎

Proof

The proof is roughly as follows. Ordinary cut-elimination for constructive predicate logic (or normalization in case we have a natural deduction system) can

be formalized in HA. Reason in HA. Let a number x and a formula A be given. Introduce a measure of complexity on arithmetical formulas that counts both the depth of quantifiers and of implications. Find y such that both the axioms of T_x and A have complexity $< y$. We can construct a truthpredicate True_y for formulas of complexity $< y$. We have: $\Box_T \mathsf{UC}(\mathsf{True}_y A \to A)$. Reason inside $\Box_T$. Suppose we have $\Box_{T,x} A$. By cut-elimination we can find a T_x-proof p of A in which only formulas of complexity $< y$ occur. We now prove by induction on the subproofs of p, that all subconclusions of p are True_y. So A is True_y. Hence we find A. ❑

In section 5 we will consider RE extensions of HA in $\mathfrak{L}_{\mathsf{Ar}}$ that are closed under the De Jongh rule.

DJ For *formulas* $A, B : T \vdash A \vee B \Rightarrow$ for some n $T \vdash A \vee \Box_{T,n} B$.

(Or equivalently:

For *formulas* $A, B :$ $T \vdash A \vee B \Rightarrow$ for some n $T \vdash \Box_{T,n} A \vee \Box_{T,n} B$.)

We will have a brief look at the question which theories can be seen to have this property. Note that if a theory T closed under the De Jongh Rule satisfies Σ-reflection, then it satisfies the Disjunction Property. We define a translation due to Dick de Jongh. Let T be an RE-extension of HA and let n be a natural number. Define a translation $[T]_n(.)$ as follows:

- $[T]_n P := P$ for P atomic,
- $[T]_n(.)$ commutes with $\wedge, \vee, \exists$,
- $[T]_n(A \to B) := ([T]_n A \to [T]_n B) \wedge \Box_{T,n}(A \to B)$,
- $[T]_n \forall y\, A(y) := \forall y\, [T]_n A(y) \wedge \Box_{T,n} \forall y\, A(y)$.

Let's first make a few quick observations, that make life easier:

i) $\mathsf{HA} \vdash [T]_n A \to \Box_{T,n} A, T \vdash [T]_n A \to A$,

ii) $\mathsf{HA} \vdash [T]_n((A \to B) \wedge (A' \to B')) \leftrightarrow$
$([T]_n A \to [T]_n B) \wedge ([T]_n A' \to [T]_n B') \wedge \Box_{T,n}((A \to B) \wedge (A' \to B'))$.
Similarly for conjunctions of more than two implications.

iii) $\mathsf{HA} \vdash [T]_n \forall y \forall z\, A(y,z) \leftrightarrow (\forall y \forall z\, [T]_n A(y,z) \wedge \Box_{T,n} \forall y \forall z\, A(y,z))$. Similary for larger blocks of universal quantifiers.

iv) $\mathsf{HA} \vdash [T]_n \forall y\, (A(y) \to B(y)) \leftrightarrow$
$\forall y\, ([T]_n A(y) \to [T]_n B(y)) \wedge \Box_{T,n} \forall y\, (A(y) \to B(y))$.

v) $\mathsf{HA} \vdash [T]_n \forall y < z\, A(y) \leftrightarrow \forall y < z\, [T]_n A(y)$.

vi) For $S \in \Sigma : \mathsf{HA} \vdash S \leftrightarrow [T]_n S$

In (vi) Σ is the set of Σ-formulas. We assume that these are the results of prefixing a block of existential quantifiers to a Δ_0-formula. Δ_0-formulas, in their turn, are generated from atoms and negated atoms using conjunction, disjunction and bounded quantification. (v) is immediate from the well known fact that:

$$\mathsf{HA} \vdash \forall y < z\, \Box_{T,n} A(y) \rightarrow \Box_{T,n} \forall y < z\, A(y).$$

(vi) is immediate from (v). Let's write $[T]_n\Gamma := \{[T]_n D \mid D_\Gamma\}$. We have:

vii) $\Gamma \vdash_{\mathsf{HA},n} A \Rightarrow [T]_n\Gamma \vdash_{\mathsf{HA}} [T]_n A$ (verifiably in HA).

Proof

Of (vii): The proof is by induction on the proof witnessing $\Gamma \vdash_{\mathsf{HA},n} A$. We treat two cases.

Suppose A is an induction axiom, say for $B(x)$, of HA_n. Clearly $[T]_n A$ is HA-provably equivalent to:

$$\begin{array}{llll} [\;\; \{ & [T]_n B(0) & \wedge & \forall x\, ([T]_n B(x) \rightarrow [T]_n B(x+1)) \\ & & \wedge & \Box_{T,n} \forall x\, (B(x) \rightarrow B(x+1)) \qquad \} \\ & & \rightarrow & (\forall x\, [T]_n B(x) \wedge \Box_{T,n} \forall x\, B(x)) \qquad\quad] \\ & & \wedge & \Box_{T,n} A \qquad\qquad\qquad\qquad\qquad\quad . \end{array}$$

We have, $\mathsf{HA} \vdash \Box_{T,n} A$. So it follows that:

$$\mathsf{HA} \vdash (\Box_{T,n} B(0) \wedge \Box_{T,n} \forall x\, (B(x) \rightarrow B(x+1))) \rightarrow \Box_{T,n} \forall x\, B(x).$$

Moreover (as an instance of induction for $[T]_n B(x)$):

$$\mathsf{HA} \vdash ([T]_n B(0) \wedge \forall x\, ([T]_n B(x) \rightarrow [T]_n B(x+1))) \rightarrow \forall x\, [T]_n B(x).$$

Combining these we find the promised: $\mathsf{HA} \vdash [T]_n A$.

Suppose $A = (D \rightarrow E)$ and the last step in the proof was by:

$$\Gamma, D \vdash_{\mathsf{HA},n} E \Rightarrow \Gamma \vdash_{\mathsf{HA},n} D \rightarrow E.$$

From $\Gamma, D \vdash_{\mathsf{HA},n} E$, we have, by the Induction Hypothesis,

$$[T]_n\Gamma, [T]_n D \vdash_{\mathsf{HA},n} [T]_n E$$

and hence: $[T]_n\Gamma \vdash_{\mathsf{HA},n} [T]_n D \rightarrow [T]_n E$. Moreover, for some finite $\Gamma_0 \subseteq \Gamma$, we have: $\Gamma_0, D \vdash_{\mathsf{HA},n} E$. Let B be the conjunction of the elements of Γ_0. We find:

$$[T]_n\Gamma \vdash_{\mathsf{HA},n} \Box_{T,n} B \text{ and } \vdash_{\mathsf{HA},n} \Box_{T,n}(B \rightarrow (D \rightarrow E)).$$

Hence: $[T]_n\Gamma \vdash_{\mathsf{HA},n} \Box_{T,n}(D \rightarrow E)$. We may conclude:

$$[T]_n\Gamma \vdash_{\mathsf{HA},n} ([T]_n D \rightarrow [T]_n E) \wedge \Box_{T,n}(D \rightarrow E)$$

❑

Let $\mathfrak{C}$ be the smallest class such that:

a) $\Sigma \subseteq \mathfrak{C}$,

b) for any $B \in \mathfrak{L}_{\mathsf{Ar}}$, and for any $C \in \mathfrak{C}$, $(B \to C) \in \mathfrak{C}$,

c) $\mathfrak{C}$ is closed under conjunction and universal quantification.

Note that the formulas of $\mathfrak{C}$ can always be brought in the form of a universal quantification of a conjunction of implications with a Σ-formula in the consequent.

viii) Suppose $A \in \mathfrak{C}$, then $\vdash_T \Box_{T,n} A \to [T]_n A$.

Proof

Of (viii): The proof is by induction on the definition of $\mathfrak{C}$. We treat case (b). Suppose $A = (B \to C)$. Reason in T. Suppose $\Box_{T,n} A$. We have to show $[T]_n A$. Clearly, it is sufficient to show $[T]_n B \to [T]_n C$. Suppose $[T]_n B$. Then $\Box_{T,n} B$. Combining this with $\Box_{T,n} A$, we get $\Box_{T,n} C$, and, hence, by the Induction Hypothesis: $[T]_n C$. ❑

Let T be an RE extension of HA, axiomatized (over HA) by $\mathfrak{C}$-formulas. We have:

ix) $\Gamma \vdash_{T,n} A \Rightarrow [T]_n \Gamma \vdash_T [T]_n A$.

In case the relevant properties of T are verifiable in HA, (ix) is also verifiable in HA.

Proof

Of (ix): Note that if C is a non-HA axiom of T which is smaller than n, then $T \vdash \Box_{T,n} C$, and, hence, $T \vdash [T]_n C$. ❑

Theorem 2.2 *Any $\mathfrak{C}$-axiomatized RE extension of HA is closed under the De Jongh Rule.*

Proof

Suppose $T \vdash B \vee C$, then, for some $n, T \vdash_n B \vee C$. Hence, by (ix), $T \vdash [T]_n(B \vee C)$. Ergo $T \vdash [T]_n B \vee [T]_n C$, and thus, by (i), $T \vdash B \vee \Box_{T,n} C$. ❑

Of course, there are many other extensions of HA, not covered by theorem 2.2 – like HA plus the uniform reflection principle for HA – that are closed under the De Jongh Rule.

We close this section by giving an application of the De Jongh Rule. We remind the reader of two forms of Markov's Rule for T (see Troelstra [73]):

MR $T \vdash \forall x\,(A(x) \vee \neg A(x))$ and $T \vdash \neg\neg\exists x\,A(x) \Rightarrow T \vdash \exists x\,A(x)$

$\mathsf{MR}_{\mathsf{PR}}$ $T \vdash \neg\neg S \Rightarrow T \vdash S$, for S a Σ-formula.

The second form is Primitive Recursive Markov's Rule. Closure under MP immediately implies closure under $\mathsf{MP}_{\mathsf{PR}}$.

Theorem 2.3 *Suppose T is an extension of* HA *closed under the De Jongh Rule and* $\mathsf{MP}_{\mathsf{PR}}$. *Then T is closed under* MP.

Proof

Suppose T satisfies the conditions of the theorem and $T \vdash \forall x\,(A(x) \vee \neg A(x))$. By the De Jongh Rule, we have for some n: $T \vdash \forall x\,(\Box_{T,n}A(x) \vee \neg A(x))$. If $T \vdash \neg\neg\exists x\,A(x)$, it follows that: $T \vdash \neg\neg\exists x\,\Box_{T,n}A(x)$. Hence, by $\mathsf{MP}_{\mathsf{PR}}$: $T \vdash \exists x\,\Box_{T,n}A(x)$. We may conclude: $T \vdash \exists x\,A(x)$. ❑

2.3 The Rosser ordering

The definitions of witness comparisons between sentences are as follows:

- $\exists x\,Ax \leq \exists y\,By :\Leftrightarrow \exists x\,(Ax \wedge \forall y < x\,\neg By)$
- $\exists x\,Ax < \exists y\,By :\Leftrightarrow \exists x\,(Ax \wedge \forall y \leq x\,\neg By)$.

In this paper we will only consider witness comparison between Σ-formulas with precisely one (outer) unbounded existential quantifier. For such Σ-formulas the obvious properties of the ordering, familiar from the classical case, also hold constructively. E.g., $i\text{-}I\Delta_0 \vdash S \rightarrow (S \leq S' \vee S' < S)$.

We write $(\exists x\,Ax \leq \exists y\,By)^{\perp}$ for $\exists y\,By < \exists x\,Ax$, and $(\exists x\,Ax < \exists y\,By)^{\perp}$ for $\exists y\,By \leq \exists x\,Ax$. For later use and for illustration, we reproduce an typical argument involving these notions.

Let T be an extension of $I\Delta_0 + \mathsf{Exp}$ and let R be the ordinary Σ Rossersentence for T. R satisfies: $T \vdash R \leftrightarrow \Box_T\neg R \leq \Box_T R$. We show: $T \vdash \neg R^{\perp} \Rightarrow T \vdash \perp$. Suppose $T \vdash \neg R^{\perp}$. Since $T \vdash \Box_T R \rightarrow (R \vee R^{\perp})$, it follows that $T \vdash \Box_T R \rightarrow R$. By the Löb's Rule, we find: $T \vdash R$. Hence, by Rosser's Theorem, $T \vdash \perp$.

3 What is an interpretation?

3.1 The definition

Let $\mathfrak{U}$ and $\mathfrak{V}$ be languages. An $(\mathfrak{U},\mathfrak{V}\text{-})$*translation* $\mathcal{M}$ is a mapping from the formulas of $\mathfrak{U}$ with only arithmetical variables free to the formulas of $\mathfrak{V}$ with only arithmetical variables free. We demand:

A1) $FV(A^{\mathcal{M}}) \subseteq FV(A)$.

We write: $A^{\mathcal{M}}$ or $A[\mathcal{M}]$ for: $\mathcal{M}(A)$.

An translation is *bounded* if its range is inside a restricted complexity class of formulas. The hallmark of such a class is the existence of a truth predicate with reasonable properties for it. A translation is *primitive recursive (recursive, ...)* if it is primitive recursive (recursive, ...) as a function.

An $\mathfrak{U}, \mathfrak{V}$-translation $\mathcal{M}$ is *an interpretation in* T. if it has the following properties for all $\mathfrak{U}$-formulas A and B:

A2) $T \vdash x = 0 \to (x = 0)^{\mathcal{M}}$.

A3) $T \vdash y = Sx \to (y = Sx)^{\mathcal{M}}$.

A4) $\mathsf{Ar} \vdash_{\mathfrak{U}} A \Rightarrow T \vdash A^{\mathcal{M}}$.

A5) $T \vdash (A^{\mathcal{M}} \wedge (A \to B)^{\mathcal{M}}) \to B^{\mathcal{M}}$.

The subscript $\mathfrak{U}$ in A4 indicates that we mean consequences of the axioms Ar in the language $\mathfrak{U}$. Note that interpretations can be completely wild. $\mathcal{M}$ need not even be recursive.

The free variables occurring in $A^{\mathcal{M}}$ are free variables of T. They should be viewed as standing for elements of the world of T being injected in $\mathcal{M}$. What this means will get clearer in the examples of subsection 3.2. The conditions A1-4 reflect the fact that we want the embedding of the numbers of T into the numbers of $\mathcal{M}$ to behave properly.

A parenthetical remark: our notion of interpretation is perhaps still too 'HA-centric', since we definitionally assume that *all* the numbers of T can be embedded. Generality could be gained by only asking embeddability of the numbers in some definable cut.

Let $\mathcal{M}$ be a $\mathfrak{U}, \mathfrak{V}$-interpretation in V. Let U be a theory in $\mathfrak{U}$. We write:

- $\mathcal{M} : V \rhd U :\Leftrightarrow$ for all sentences A in $\mathfrak{U}$, $U \vdash A \Rightarrow V \vdash A^{\mathcal{M}}$.
- $\mathcal{M} : V \rhd_{\mathsf{faith}} U :\Leftrightarrow$ for all sentences A in $\mathfrak{U}$, $U \vdash A \Leftrightarrow V \vdash A^{\mathcal{M}}$.

In the first case we say that $\mathcal{M}$ is *an interpretation of* U in V. In the last case, we say that $\mathcal{M}$ is a *faithful* interpretation of U in V.

We proceed to define the notion of local interpretation. There are many possible definitions. More experimentation is certainly necessary to see which of the various possibilities is best. Here we present just one of the possible choices. Let $\mathcal{R}$ be a function from the natural numbers to $\mathfrak{U}, \mathfrak{V}$-interpretations in V. Let Δ range over finite subsets of sentences of $\mathfrak{U}$. Define:

- $\mathcal{R} : V \rhd_{\mathsf{loc}} U :\Leftrightarrow \forall \Delta \exists i \forall j \geq i \forall A \in \Delta\, (U \vdash A \Rightarrow V \vdash A^{\mathcal{R}(j)})$.
- $\mathcal{R} : V \rhd_{\mathsf{loc,faith}} U :\Leftrightarrow \forall \Delta \exists i \forall j \geq i \forall A \in \Delta\, (U \vdash A \Leftrightarrow V \vdash A^{\mathcal{R}(j)})$.

We will consider interpretations with further properties $P, Q, \ldots$ We will speak of P-interpretations, P, Q-interpretations etc. Examples of such properties are:

B1) $T \vdash (A \vee B)^{\mathcal{M}} \to A^{\mathcal{M}} \vee B^{\mathcal{M}}$.

B2) $T \vdash \bot^{\mathcal{M}} \rightarrow \bot$.

B3) $T \vdash \forall x\,(x < c)^{\mathcal{M}}$, for some numerical $\mathfrak{L}$-constant c.

B4) $T \vdash A^{\mathcal{M}} \rightarrow (A^{\mathcal{M}})^{\mathcal{M}}$.

B5(Γ) $T \vdash C \rightarrow C^{\mathcal{M}}$, for C a Γ-formula.

For B5 to make sense we need that Γ is both in $\mathfrak{U}$ and $\mathfrak{V}$. We can relax this condition a bit by only asking that Γ is mapped via standard relative interpretations on the formulas of both $\mathfrak{U}$ and $\mathfrak{V}$. We will take this relaxed view in case Γ is a set of arithmetical formulas.

A B1-interpretation will be called *disjunctive*, a B2-interpretation *consistent*, a B1,B2-interpretation *prime*, B3-interpretation *provably non-standard*, a B4-interpretation *inductive*. A B5(Γ)-interpretation is called Γ-*complete*. If Γ is all of $\mathfrak{V}$, we will simply say that the interpretation is complete.

Let's first note an elementary fact.

Fact 3.1 Suppose $\mathcal{M}$ is a interpretation in T.

i) $\mathcal{M}$ provably commutes with conjunction, i.e.,
$T \vdash (A \wedge B)^{\mathcal{M}} \leftrightarrow (A^{\mathcal{M}} \wedge B^{\mathcal{M}})$.

ii) $T \vdash (A^{\mathcal{M}} \vee B^{\mathcal{M}}) \rightarrow (A^{\mathcal{M}} \wedge B^{\mathcal{M}})$. Hence, if $\mathcal{M}$ is prime, then $\mathcal{M}$ provably commutes with disjunction.

iii) $T \vdash (\forall x\, A)^{\mathcal{M}} \rightarrow \forall x\,(A^{\mathcal{M}})$.

iv) $T \vdash \exists x\,(A^{\mathcal{M}}) \rightarrow (\exists x\, A)^{\mathcal{M}}$.

❑

Proof

(i) Reason in T. Suppose $(A \wedge B)^{\mathcal{M}}$. By A4: $((A \wedge B) \rightarrow A)^{\mathcal{M}}$. Hence by A5: $A^{\mathcal{M}}$. Similarly we find $B^{\mathcal{M}}$. Hence $(A^{\mathcal{M}} \wedge B^{\mathcal{M}})$. Conversely suppose $(A^{\mathcal{M}} \wedge B^{\mathcal{M}})$. By A4: $(A \rightarrow (B \rightarrow (A \wedge B)))^{\mathcal{M}}$. Hence by two applications of A5: $(A \wedge B)^{\mathcal{M}}$. (ii), (iii), and (iv) are trivial. ❑

3.2 Examples

There are plenty of interpretations in T. We just provide a selection of the possibilities.

3.2.1 Identity

The identity function $\mathcal{ID}$ from $\mathfrak{L}_T$ to $\mathfrak{L}_T$ is a faithful interpretation of T in T. $\mathcal{ID}$ is unbounded, prime, inductive and complete.

3.2.2 $(A \to (.))$

Let A be any $\mathfrak{L}_T$-sentence. The function mapping $\mathfrak{L}_T$-formulas B to $(A \to B)$ is a faithful interpretation of $T + A$ in T. If A is not refutable in T, the interpretation is unbounded, inductive and complete. Generally the interpretation is not consistent, nor disjunctive. It commutes with universal quantification and implication.

3.2.3 $(A_i \to (.))_{i \in \omega}$

Let U be an $\mathfrak{L}_T$-theory extending T. Fix an enumeration of the axioms of U (over T). Let A_i be the conjunction of the first i axioms of U. Then $\mathcal{R}$ with $\mathcal{R}(i)(B) := (A_i \to B)$ is a faithful local interpretation of U in T.

3.2.4 Adding a function symbol

Suppose the free variables of A are $\mathbf{x}, y$ and

$$T \vdash \forall \mathbf{x} \exists y\, A(\mathbf{x}, y) \text{ and } T \vdash \forall \mathbf{x}, y, z((A(\mathbf{x}, y) \wedge A(\mathbf{x}, z)) \to y = z).$$

Let $\mathfrak{L}$ extend the language of T with a new function symbol f with arity $|\mathbf{x}|$. The usual translation of $\mathfrak{L}$ into $\mathfrak{L}_T$ is an interpretation in our sense. The interpretation is unbounded, prime, inductive and complete. Moreover the interpretation commutes with all connectives.

3.2.5 Adding a generic P

Let P be a formula with only x free. Suppose $T \vdash \exists x\, P(x)$. Let $\mathfrak{L}$ extend the language of T with a new constant c. Define for A in $\mathfrak{L}$:

$$A^{\mathcal{M}} := \forall z\, (P[x := z] \to A[c := z]),$$

where z is the first variable not occurring in P, A.

It is easily seen that $\mathcal{M}$ is an interpretation. In case, c does not occur in A, we have: $T \vdash A \leftrightarrow A^{\mathcal{M}}$. So $\mathcal{M}$ is complete and consistent. It commutes with universal quantification and with implications with antecedent without c. It is easily seen that $\mathcal{M}$ cannot be disjunctive. Let e.g. $P(x)$ be $x = x$ and let T be HA. We have:

$$\mathsf{HA} \vdash (\text{`}c \text{ is even'} \vee \text{`}c \text{ is odd'})^{\mathcal{M}}.$$

But not:

$$\mathsf{HA} \vdash (\text{`}c \text{ is even'})^{\mathcal{M}} \vee (\text{`}c \text{ is odd'})^{\mathcal{M}}.$$

3.2.6 Adding a generic non-standard P

Let P be a formula with only x free. Suppose $T \vdash \forall y \exists x \neq y\, P(x)$. Let $\mathfrak{L}$ extend the language of T with a new constant c. Define for $A \in \mathfrak{L}$:

$$A^{\mathcal{M}} := \exists y \forall z \geq y\, (P[x := z] \to A[c := z]),$$

where z is the first variable not occurring in P, A.

It is easily seen that $\mathcal{M}$ is an interpretation. In case, c does not occur in A, we have: $T \vdash A \leftrightarrow A^{\mathcal{M}}$. So $\mathcal{M}$ is complete and consistent. We have: $T \vdash \forall x\,(x < c)^{\mathcal{M}}$. Just like in example 3.2.5 our interpretation is not disjunctive. In section 5 we will see that this feature is necessary in HA.

3.2.7 $\neg\neg$

The function mapping $\mathfrak{L}_T$-formulas A to $\neg\neg A$ is an interpretation in T. It is unbounded, consistent, inductive and complete. It commutes with implication. It interprets T plus propositional excluded third in T. It is not disjunctive.

3.2.8 The double negation translation $\mathfrak{dnt}$

Gödel's double negation translation $\mathfrak{dnt}$ can be obtained by replacing in a formula A every subformula that is a disjunction or an existential quantification by its double negation. The interpretation is unbounded, consistent, inductive. It is not complete. It commutes with implication and universal quantification. We have $\mathfrak{dnt} : \mathsf{HA} \rhd_{\mathsf{faith}} \mathsf{PA}$. $\mathfrak{dnt}$ is not disjunctive. It is easy to see that – under reasonable conditions – the non-disjunctivity is unavoidable for interpretations of PA in HA. Suppose e.g. that $\mathcal{M}$ is a primitive recursive, disjunctive interpretation of PA in HA. We show that $\mathcal{M}$ interprets $\mathsf{HA}+\bot$. Let T be given by: $T \vdash A \Leftrightarrow \mathsf{HA} \vdash A^{\mathcal{M}}$. By the properties of interpretations, T is an RE theory of predicate logic, extending PA. Let R be the ordinary Rosser sentence for T. We have:

$$\begin{aligned} T \vdash R \vee \neg R &\Rightarrow \mathsf{HA} \vdash (R \vee \neg R)^{\mathcal{M}} \\ &\Rightarrow \mathsf{HA} \vdash R^{\mathcal{M}} \vee (\neg R)^{\mathcal{M}} \\ &\Rightarrow \mathsf{HA} \vdash R^{\mathcal{M}} \text{ or } \mathsf{HA} \vdash (\neg R)^{\mathcal{M}} \\ &\Rightarrow T \vdash R \text{ or } T \vdash \neg R \\ &\Rightarrow T \vdash \bot\,. \end{aligned}$$

See the remark below theorem 5.1 for another proof of this fact.

3.2.9 Non-standard interpretations via the Henkin construction in PA

As is well known, since the work of Orey and Feferman, we can mimick the Henkin model construction in $\mathsf{PA} + \mathsf{Con}(V)$ to construct an interpretation of V. In the appendix we will see that the result can be widened to include the analogue of the Henkin construction of Kripke Models for theories in Intuitionistic Predicate Logic. This result holds for PA, not for HA. We can induce the result in HA by composing it with $\mathfrak{dnt}$. However, the resulting interpretation will not be disjunctive.

3.2.10 Kleene realizability $\mathfrak{r}$

Let $\mathbf{r}$ be ordinary Kleene realizability over HA (See Troelstra[72a] or Troelstra & van Dalen[88a]). We put: $A^{\mathfrak{r}} := \exists x\, x\mathbf{r}A$. Then $\mathfrak{r}$ is a faithful interpretation of $\mathsf{HA} + \mathsf{ECT}_0$ in HA. $\mathfrak{r}$ is prime, inductive.

3.2.11 q-realizability $\mathfrak{q}$

Let $\mathbf{q}$ be ordinary q-realizability over HA. We put: $A^{\mathfrak{q}} := \exists x\, x\mathbf{q}A$ (See Troelstra[72a] or Troelstra & van Dalen[88a]). q-realizability has two versions. The version we will use has the same clauses as ordinary realizalility, except in the cases of $\rightarrow$ and $\forall$, where a conjunction with the original formula is added to the clause of realizability. E.g., the clause for $\rightarrow$ is:

$$x\mathbf{q}(B \rightarrow C) := \forall y\, (y\mathbf{q}B \rightarrow \exists z\, (\{x\}y \simeq z \wedge z\mathbf{q}C)) \wedge (B \rightarrow C)$$

We have: $\mathfrak{q}$ is a faithful interpretation of HA in HA. $\mathfrak{q}$ is prime and inductive.

3.2.12 The Friedman translation

The Friedman translation was introduced in Friedman[77]. See also Troelstra & van Dalen[88a], 136-139. Let B be any $\mathfrak{L}_{\mathsf{Ar}}$-sentence. The $(.)^B : \mathfrak{L}_{\mathsf{Ar}} \rightarrow \mathfrak{L}_{\mathsf{Ar}}$ is defined as follows. $(A)^B$ is the result of replacing each atomic formula P in A by $P \vee B$. The Friedman translation is an interpretation of HA in HA, but it is only consistent in T if $T \vdash \neg B$. The Friedman translation commutes with all connectives (except $\bot$), but is not generally inductive. (It is inductive, if B is in Σ.) The Friedman translation is extremely useful for proving derived rules.

3.2.13 Provability

Consider the mapping of $\mathfrak{L}_{\mathsf{Ar}}$-formulas A to $\Box_T A$. This gives us an interpretation of T in T, which is – assuming T to be consistent – not disjunctive and not consistent.

3.2.14 The provability translation $\mathfrak{p}_T$

The provability translation $\mathfrak{p}_T$ was studied extensively in Visser [82]. It is a simplification of Beeson's fp-realizability (see Beeson [75]). Viewed in a different way it is just a variant of Gödel's translation of intuitionistic logic into $S4$, where we substitute provability for necessity and where we do not necessarily start with classical logic. The translation commutes with all connectives except $\rightarrow$ and $\forall$. Here the clauses are:

- $(A \rightarrow B)^{\mathfrak{p}} := (A^{\mathfrak{p}} \rightarrow B^{\mathfrak{p}}) \wedge \Box_T(A^{\mathfrak{p}} \rightarrow B^{\mathfrak{p}})$
- $(\forall x\, A)^{\mathfrak{p}} := \forall x\, A^{\mathfrak{p}} \wedge \Box_T \forall x\, A^{\mathfrak{p}}$

If we take $T := \mathsf{HA}$, we find, e.g., that $\mathfrak{p}_{\mathsf{HA}} : \mathsf{HA} \rhd_{\mathsf{faith}} \mathsf{HA}^*$, where HA^* is the unique RE theory satisfying the following equation *verifiably in* HA:

- $\mathsf{HA}^* = \mathsf{HA} + \{A \to \Box_{\mathsf{HA}^*} A | A \in \mathfrak{L}_{\mathsf{Ar}}\}$.

(For a proof of the uniqueness of the solution of the equation, see Visser [82].) The provability translation is prime and inductive.

3.2.15 The De Jongh translation ∂j

The De Jongh translation can be considered as a local interpretation. For example for the case of HA it is a faithful local interpretation of HA in HA.

3.2.16 Feferman provability

The Feferman predicate for an $\mathfrak{L}_{\mathsf{Ar}}$-extension T of HA is the predicate $\triangle_T A$ given by: $\triangle_T A := \exists x\,(\Box_{T,x} A \wedge \Diamond_{T,x} \top)$. (Here as usual, $\Diamond_{T,x}$ is $\neg\Box_{T,x}\neg$.) The mapping A to $\triangle_T A$ is an interpretation of T in T (due to the fact that T is reflexive). It is consistent, but not disjunctive. In section 6 we will sketch the construction of a modified Feferman predicate for HA that is prime.

4 Basic facts about interpretations

In this section we verify the basic facts of interpretations. These facts are not the most exciting things in the world, but necessary to see that our interpretations behave decently. Our notion of interpretation is mainly 'about' how external variables behave internally. The facts of this section substantiate that the behaviour is as expected. In this section $\mathcal{M}$ is an interpretation in T.

Lemma 4.1 $T \vdash x = y \to (x = y)^{\mathcal{M}}$. ❑

Proof

We use Ar5. Suppose $x = y$. In case $x = 0$, we have $y = 0$. By A2 $(x = 0)^{\mathcal{M}}$ and $(y = 0)^{\mathcal{M}}$. So by A4, A5: $(x = y)^{\mathcal{M}}$. In case $x = Su$, we have $y = Su$. So by A3: $(x = Su)^{\mathcal{M}}$ and $(y = Su)^{\mathcal{M}}$. Hence by A4, A5: $(x = y)^{\mathcal{M}}$. ❑

With lemma 4.1 in hand we are ready to settle a hairy detail: interpretations behave decently w.r.t. substitutions of variables. Even if the fact proved here is a complete and utter triviality, still, experience teaches, it is very easy to make mistakes in this area.

Let σ be a function from a *finite* set V of variables to variables. Consider a formula A. In case for every v in V $v\sigma$ is substitutable for v in A, $A\sigma$ is the result of substituting, for all $v \in V$, $v\sigma$ for all free occurrences of v in A. In case not every $v\sigma$ is substitutable for v, $A\sigma$ is obtained as follows. We first go to an α-variant A' of A (i.e. we replace some bound variables by appropriate other ones) such that all the $v\sigma$'s are substitutable in A' and then substitute the $v\sigma$.

Lemma 4.2 We have: $T \vdash A^{\mathcal{M}}\sigma \leftrightarrow A\sigma^{\mathcal{M}}$. ❑

Proof

By A1 we may assume that $V \subseteq FV(A)$. We first assume that $V \cap \mathsf{range}(\sigma) = \emptyset$. By lemma 4.1 we have:

$$T \vdash \bigwedge\{v = v\sigma \,|\, v \in V\} \rightarrow (\bigwedge\{v = v\sigma | v \in V\})^{\mathcal{M}}.$$

Ergo by A4, A5:

$$T \vdash \bigwedge\{v = v\sigma | v \in V\} \rightarrow (A \leftrightarrow A\sigma)^{\mathcal{M}}.$$

And, hence, by A4, A5:

$$T \vdash \bigwedge\{v = v\sigma | v \in V\} \rightarrow (A^{\mathcal{M}} \leftrightarrow A\sigma^{\mathcal{M}}).$$

Since V is disjoint from $\mathsf{range}(\sigma)$, V will be disjoint from $FV(A\sigma)$ and, hence, V will also be disjoint from $FV(A\sigma^{\mathcal{M}})$. By predicate logic, using the fact that $\mathsf{range}(\sigma)$ is disjoint from V, we find: $T \vdash A^{\mathcal{M}}\sigma \leftrightarrow A\sigma^{\mathcal{M}}$.

Now let σ be arbitrary. Let μ be a bijection between V and some set of variables U disjoint from $FV(A) \cup \mathsf{range}(\sigma)$. Since $\mathsf{range}(\mu)$ is disjoint from V, we have:

$$T \vdash A^{\mathcal{M}}\mu \leftrightarrow A\mu^{\mathcal{M}}.$$

Consider $\nu = \mu^{-1}\sigma$ (i.e., first μ^{-1}, then σ). Clearly it follows that:

$$T \vdash A^{\mathcal{M}}\mu\nu \leftrightarrow A\mu^{\mathcal{M}}\nu.$$

Note that $dom(\nu) = U$ and $\mathsf{range}(\nu) = \mathsf{range}(\sigma)$. Since $\mathsf{range}(\nu)(= \mathsf{range}(\sigma))$ is disjoint from U we have:

$$T \vdash A\mu^{\mathcal{M}}\nu \leftrightarrow A\mu\nu^{\mathcal{M}}.$$

Hence: $T \vdash A^{\mathcal{M}}\mu\nu \leftrightarrow A\mu\nu^{\mathcal{M}}$. Since U is disjoint from $FV(A)$ and hence from $FV(A^{\mathcal{M}})$, we find: $A\mu\nu = A(\mu\nu) = A\sigma$ and $A^{\mathcal{M}}\mu\nu = A^{\mathcal{M}}(\mu\nu) = A^{\mathcal{M}}\sigma$.[1] Ergo:

$$T \vdash A^{\mathcal{M}}\sigma \leftrightarrow A\sigma^{\mathcal{M}}.$$

❑

In lemma 4.3, we will strengthen lemma 4.2 a bit.

Lemma 4.3 i) $T \vdash (y = 0)^{\mathcal{M}}[y := 0]$

ii) $T \vdash (y = Sx)^{\mathcal{M}}[y := Sx]$

Let σ be a function from a set V of variables to terms of one of the following forms: $0, x, Sy$. Consider a formula A. $A\sigma$ is defined in the obvious way.

iii) $T \vdash A^{\mathcal{M}}\sigma \leftrightarrow A\sigma^{\mathcal{M}}$.

❑

[1] Note the pittfall here. If, e.g., A is $(x = y)$ and $\mu := [y := x]$ and σ is the empty susbstitution, id, then: $\nu = \mu^{-1}\sigma = [x := y]id = [x := y]$ and $\mu\nu = id = \sigma$. But $(x = y)\mu\nu = (x = x)\nu = (y = y) \neq (x = y) = (x = y)\sigma$. The reason for this phenomenon is that μ^{-1} is only the inverse of μ as a function on its domain, not the inverse of its extension μ^* to all variables given by: $\mu^*(x) = \mu(x)$ if $x \in V$, $\mu^*(x) = x$ otherwise.

Proof

i) By A2, $T \vdash y = 0 \to (y = 0)^{\mathcal{M}}$. Specializing, we find:

$$T \vdash 0 = 0 \to (y = 0)^{\mathcal{M}}[y := 0].$$

Hence, $T \vdash (y = 0)^{\mathcal{M}}[y := 0]$. The proof of (ii) is similar.

iii) The proof follows the same general lines as the proof of lemma 4.2. We just treat the special case that $\sigma = [y := 0]$. We have:

a)	$\mathsf{Ar} \vdash (y = 0 \to (A \leftrightarrow A[y := 0]))$	
b)	$T \vdash (y = 0 \to (A \leftrightarrow A[y := 0]))^{\mathcal{M}}$	$a, A4$
c)	$T \vdash (y = 0)^{\mathcal{M}} \to (A^{\mathcal{M}} \leftrightarrow A[y := 0]^{\mathcal{M}})$	$b, A5$
d)	$T \vdash (y = 0)^{\mathcal{M}}[y := 0] \to (A^{\mathcal{M}}[y := 0] \leftrightarrow A[y := 0]^{\mathcal{M}}[y := 0])$	c
e)	$T \vdash A^{\mathcal{M}}[y := 0] \leftrightarrow A[y := 0]^{\mathcal{M}}$	$d, (i), A1$

❑

With lemmas 4.2 and 4.3 done we can devote our attention to the more interesting 'preservation' theorems. For these theorems we often need a modicum of induction. Let C be a class of numerical constants of $\mathfrak{L}$. $\Delta_0(\mathcal{M}, C)$ is the smallest class such that:

- $(s = t) \in \Delta_0(\mathcal{M}, C)$,
- $B \in \Delta_0(C) \Rightarrow B^{\mathcal{M}} \in \Delta_0(\mathcal{M}, C)$,
- $\Delta_0(\mathcal{M}, C)$ is closed under the propositional connectives and bounded quantification.

T is $\mathcal{M}, C$-*adequate* if it satisfies $\Delta_0(\mathcal{M}, C)$-induction. T is $\mathcal{M}$-*adequate* if it is $\mathcal{M}, \emptyset$-adequate. Note that HA is $\mathcal{M}, C$-adequate for all interpretations $\mathcal{M}$ in HA.

Lemma 4.4 Suppose T is $\mathcal{M}$-adequate. We have:

i) $T \vdash x + y = z \to (x + y = z)^{\mathcal{M}}$.

ii) $T \vdash x.y = z \to (x.y = z)^{\mathcal{M}}$.

iii) $T \vdash x \neq y \to (x \neq y)^{\mathcal{M}}$.

❑

Proof

Reason in T.

i) We show by induction on y that: $\forall z\,(x + y = z \to (x + y = z)^{\mathcal{M}})$. This induction can be viewed as a $\Delta_0(\mathcal{M})$ induction, since we can, trivially, bound z

by $x+Sy$. We show first: $(\forall z\,(x+y=z \to (x+y=z)^{\mathcal{M}}))[y:=0]$. By lemma 4.3 this is equivalent to: $\forall z\,(x+0=z \to (x+0=z)^{\mathcal{M}})$. Suppose $x+0=z$, then $x=z$, hence $(x=z)^{\mathcal{M}}$ and so by A4,A5: $(x+0=z)^{\mathcal{M}}$. Next we show: $(\forall z\,(x+y=z \to (x+y=z)^{\mathcal{M}}))[y:=Su]$. By lemma 4.3 this is equivalent to: $(\forall z\,(x+Su=z \to (x+Su=z)^{\mathcal{M}})$. Suppose $x+Su=z$. It follows that for some $v: z=Sv$ and $x+u=v$. By the induction hypothesis (in combination with lemma 4.2), we find: $(x+u=v)^{\mathcal{M}}$. Hence, by A4,A5: $(x+Su=Sv)^{\mathcal{M}}$. Moreover by A3: $(z=Sv)^{\mathcal{M}}$. Ergo, by A4,A5: $(x+Su=z)^{\mathcal{M}}$. The proof of (ii) is similar.

iii) Suppose $x \neq y$. Since T contains i-$I\Delta_0$, it follows that $x<y$ or $y<x$. (Here $x<y$ is defined as: $\exists u\, x+Su=y$). We prove by induction on y that: for all $z<y$ $(z\neq y)^{\mathcal{M}}$. If $y=0$, this is trivial. Suppose $y=Su$. Consider $z<Su$. By i-$I\Delta_0$, $z<u$ or $z=u$. In case $z=0$, we are easily done. Suppose $z=Sv$. We have $v<u$. It follows, by the Induction Hypothesis that $(z\neq y)^{\mathcal{M}}[z:=v, y:=u]$, i.o.w. $(v\neq u)^{\mathcal{M}}$. By A4,A5, we get: $(Sv\neq Su)^{\mathcal{M}}$. Note that by A3 we have: $(y=Su)^{\mathcal{M}}$ and $(z=Sv)^{\mathcal{M}}$. Hence, by A4,A5: $(z\neq y)^{\mathcal{M}}$. ❑

Lemma 4.5 Suppose T is $\mathcal{M}$-adequate. We have for $\mathfrak{L}_{\mathsf{Ar}}$-terms t and u:

i) $T \vdash t=u \to (t=u)^{\mathcal{M}}$.

ii) $T \vdash t\neq u \to (t\neq u)^{\mathcal{M}}$.

iii) $T \vdash (A^{\mathcal{M}} \wedge B^{\mathcal{M}}) \to (A\wedge B)^{\mathcal{M}}$.

iv) $T \vdash (A^{\mathcal{M}} \vee B^{\mathcal{M}}) \to (A\vee B)^{\mathcal{M}}$.

v) $T \vdash \exists x\, A^{\mathcal{M}} \to (\exists x\, A)^{\mathcal{M}}$.

vi) $T \vdash \forall x<y\, A^{\mathcal{M}} \to (\forall x<y\, A)^{\mathcal{M}}$ for $A\in\Delta_0$.

❑

Proof

Let x be a variable not in t. It is sufficient to show

$*$ $T \vdash t=x \to (t=x)^{\mathcal{M}}$,

since $(*)$ implies:

$$\begin{array}{rcl} T\vdash t=u & \to & \exists x\,(t=x\wedge u=x) \\ & \to & \exists x\,((t=x)^{\mathcal{M}}\wedge(u=x)^{\mathcal{M}}) \\ & \to & (t=u)^{\mathcal{M}}. \end{array}$$

The proof of $(*)$ is by meta-induction on t. Suppose e.g. $t = t' + t''$. We have, for fresh variables x' and x'',

$$\begin{array}{lll} T \vdash t = x & \rightarrow & \exists x', x''\,(t' = x' \wedge t'' = x'' \wedge x' + x'' = x) \\ & \rightarrow & \exists x', x''\,((t' = x')^{\mathcal{M}} \wedge (t'' = x'')^{\mathcal{M}} \wedge (x' + x'' = x)^{\mathcal{M}}) \\ & \rightarrow & (t = x)^{\mathcal{M}}. \end{array}$$

(ii)-(v) are left to the reader.

vi) Reason in T. We prove by induction on $y : \forall x < y\, A^{\mathcal{M}} \rightarrow (\forall x < y\, A)^{\mathcal{M}}$. In case $y = 0$, this is easy. Suppose $y = Su$ and $\forall x < Su\ A^{\mathcal{M}}$. It follows that $\forall x < u\, A^{\mathcal{M}}$ and hence by the Induction Hypothesis: $(\forall x < y\, A)^{\mathcal{M}}[y := u]$, i.e., $(\forall x < u\, A)^{\mathcal{M}}$. Also we have: $A^{\mathcal{M}}[y := u]$ and hence $(A[y := u])^{\mathcal{M}}$. We have $(y = Su)^{\mathcal{M}}$ and, hence, $(\forall x\,(x < y \leftrightarrow (x < u \vee x = u))^{\mathcal{M}}$. We may conclude: $(\forall x < y\, A)^{\mathcal{M}}$. ❑

Remember that the Δ_0-formulas are built up from atoms and negated atoms using conjunction, disjunction and bounded quantification. This rather restricted definition could make a difference inside $(.)^{\mathcal{M}}$, since Ar does not verify the equivalence with more general forms.

Theorem 4.6 *Suppose T is $\mathcal{M}$-adequate. Then $\mathcal{M}$ is Σ-complete in T.*

Proof

By a simple meta-induction on Σ-formulas, using lemma 4.5. ❑

Remark 4.7 Suppose T contains Σ-induction and $\mathcal{M}(A) := \Box_U A$, for U containing Ar. Then theorem 4.6 implies T-provable Σ-completeness for U-provability. We can see that for such applications our result is not optimal, since e.g. $T = i\text{-}I\Delta_0 + \mathsf{Exp}$ already proves Σ-completeness for U-provability. We leave the exploration of possible refinements for later work. ❑

Theorem 4.8 (Substitution of terms) *Let σ map variables of a finite set of variables V to $\mathfrak{L}_{\mathsf{Ar}}$-terms. Let T be an $\mathcal{M}$-adequate theory. We have: $T \vdash (A\sigma)^{\mathcal{M}} \leftrightarrow A^{\mathcal{M}}\sigma$.*

Proof

We treat the case that $\sigma = [x := t]$ and t is substitutable for x in A. the general argument is similar. Suppose first that t does not contain x. Since $T \vdash x = t \rightarrow (x = t)^{\mathcal{M}}$ it follows that:

$$T \vdash x = t \rightarrow (A[x := t] \leftrightarrow A)^{\mathcal{M}}.$$

Ergo, $T \vdash (A[x := t])^{\mathcal{M}} \leftrightarrow A^{\mathcal{M}}$ and so

$$T \vdash (A[x := t])^{\mathcal{M}}[x := t] \leftrightarrow A^{\mathcal{M}}[x := t].$$

We find: $T \vdash (A[x := t])^{\mathcal{M}} \leftrightarrow A^{\mathcal{M}}[x := t]$.

In case x does occur in t, let u be a fresh variable not in A or t. We get:

$$T \vdash (A[x := t[x := u]])^{\mathcal{M}} \leftrightarrow A^{\mathcal{M}}[x := t[x := u]].$$

And hence: $T \vdash (A[x := t[x := u]])^{\mathcal{M}}[u := x] \leftrightarrow A^{\mathcal{M}}[x := t[x := u]][u := x]$. Lemma 4.2 gives us the desired result. ❑

5 Limitative results concerning prime interpretations

In this section, we direct our attention mainly to HA and its extensions. We will show that if an interpretation is prime/disjunctive, then, under reasonable further assumptions, it has various other properties, like Π_2-conservativity. The results of this section are very much like improvisations on one single theme.

The immediate ancestors of these results are (i) Friedman's proof that the Σ Disjunction Property plus Consistency imply Σ-reflection (see Friedman [75] and remark 5.11 below), (ii) the work of McCarthy on the non-existence of non-standard models of arithmetic (see McCarthy [88]).

We start with the simplest result of 'non-existence of non-standard, prime interpretations'.

Theorem 5.1 *Suppose:*

i) $\mathcal{M}$ *is a prime interpretation of* $i\text{-}I\Delta_0$ *in* T,

ii) $\mathcal{M}$ *is provably non-standard in* T,

iii) $\mathcal{M}$ *is* Σ*-complete in* T

iv) T *has the disjunction property.*

Then T *is inconsistent.*

Proof

Suppose $T \vdash \forall x\, (x < c)^{\mathcal{M}}$, for the numerical $\mathfrak{L}$-constant c. Let $C := \exists x\, x = c$. Note that for any $S \in \Sigma$, we have: $T \vdash S \to (S < C)^{\mathcal{M}}$. Let R be the ordinary (Σ) Rosser sentence for T. We find:

$$i\text{-}I\Delta_0(c) \vdash (R < C) \ \vee \ (C \leq R).$$

So $T \vdash (R < C \vee C \leq R)^{\mathcal{M}}$, and hence: $T \vdash (R < C)^{\mathcal{M}} \vee (C \leq R)^{\mathcal{M}}$. Since $T \vdash R \to (R < C)^{\mathcal{M}}$, we find: $T \vdash (C \leq R)^{\mathcal{M}} \to \neg R$. Moreover $T \vdash R^{\perp} \to (R^{\perp})^{\mathcal{M}}$, and hence $T \vdash R^{\perp} \to (\neg R)^{\mathcal{M}}$. So $T \vdash (R < C)^{\mathcal{M}} \to \neg R^{\perp}$. We may conlude: $T \vdash \neg R^{\perp} \vee \neg R$. By the disjunction property: $T \vdash \neg R^{\perp}$ or $T \vdash \neg R$. Hence by Rosser's Theorem: $T \vdash \perp$. ❑

Note that we could as well have taken c to be a term of $\mathfrak{L}$. An immediate consequence of theorem 5.1 is a proof of the fact that there is no prime interpretation of PA in HA. Suppose there was one, say $\mathcal{M}$. There is a provably non-standard, prime interpretation $\mathcal{N}$ in PA. So, as is easily verified, $\mathcal{N} \circ \mathcal{M}$ (first $\mathcal{N}$, then $\mathcal{M}$) is prime and provably non-standard. A contradiction with theorem 5.1.

Note that in the present formulation of theorem 5.1 we used that the Rosser sentence is written in the strict Σ_1-form of $\mathfrak{L}_{\mathsf{Ar}}$. Nothing, however, beyond the verifiability of the properties of the Rosser ordering is used inside $\mathcal{M}$.

A disadvantage of theorem 5.1, is that e.g. HA cannot verify its own disjunction property, so we cannot verify one of the assumptions of the theorem in HA for HA. One way to get around this, is to use closure under the De Jongh Rule instead of the Disjunction Property. Another way is to use closure under Church's Rule. We first treat the approach using De Jongh closure. Note that the use of restricted provability only makes sense in the presence of full induction.

Theorem 5.2 *Suppose:*

i) $\mathcal{M}$ *is a prime interpretation of* i-$I\Delta_0$ *in* T,

ii) $\mathcal{M}$ *is provably non-standard in* T,

iii) T *satisfies full induction,*

iv) T *is closed under the De Jongh Rule.*

Then, T *is inconsistent.*

Proof

Suppose $T \vdash \forall x\,(x < c)^{\mathcal{M}}$, for the numerical $\mathfrak{L}$-constant c. Using the Gödel Fixed Point Lemma find $R(b)$ in Σ (with free variable b), such that:

$$i\text{-}I\Delta_0 \vdash R(b) \leftrightarrow \Box_{T,b}\neg R(b) \leq \Box_{T,b}R(b).$$

Let $R^{\perp}(b) := \Box_{T,b}R(b) < \Box_{T,b}\neg R(b)$. Reasoning as in the proof of theorem 5.1, we find: $T \vdash \neg R(b) \vee \neg R^{\perp}(b)$. So by the De Jongh rule, for some m:

$$T \vdash \Box_{T,m}\neg R(b) \vee \Box_{T,m}\neg R^{\perp}(b).$$

So by substituting m for b, we find:

$$T \vdash \Box_{T,m}\neg R(m) \vee \Box_{T,m}\neg R^{\perp}(m).$$

Since $R(m)$ is a Rosser-sentence for T_m, we find, by the formalization of Rosser's Theorem: $T \vdash \Box_{T,m} \perp$ and, hence, by provable reflection, $T \vdash \perp$. ❑

Lemma 5.3 Let c be a numerical constant of $\mathfrak{L}$. Suppose T is $\mathcal{M}$, $\{c\}$-adequate and disjunctive. Then: $T \vdash \forall x\,((c \leq x)^{\mathcal{M}} \rightarrow \exists y \leq x\,(c = y)^{\mathcal{M}})$. ❑

Proof

By a simple induction on x in T. ❑

T is closed under the *Strong De Jongh Rule* if, for each A, $T + \neg A$ is closed under the De Jongh Rule. As we have seen HA is closed under the Strong De Jongh Rule.

Corollary 5.4 Let c be a numerical constant of $\mathfrak{L}$. Suppose:

i) $\mathcal{M}$ is a prime interpretation of i-$I\Delta_0$ in T,

ii) T satisfies full induction,

iii) T is closed.under the Strong De Jongh Rule.

Then (a): $T \vdash \neg\neg\exists x\,(x = c)^{\mathcal{M}}$. If we also have:

iv) T is closed under Primitive Recursive Markov's Rule $\mathsf{MR}_{\mathsf{PR}}$.

Then: (b) $T \vdash \exists x\,(x = c)^{\mathcal{M}}$. ❑

Proof

(a) It is easy to see that the conditions of theorem 5.2 are satisfied by $T + \neg\neg\forall x\,(x < c)^{\mathcal{M}}$. It follows that $T \vdash \neg\forall x\,(x < c)^{\mathcal{M}}$. We have:

$$T \vdash (x < c \vee c \leq x)^{\mathcal{M}}, \text{ so } T \vdash (x < c)^{\mathcal{M}} \vee (c \leq x)^{\mathcal{M}}.$$

We may conclude: $T \vdash \neg\neg\exists x\,(c \leq x)^{\mathcal{M}}$. By lemma 5.3, we obtain the desired result.

b) Since $\mathcal{M}$ is prime, we have: $T \vdash (x = c)^{\mathcal{M}} \vee \neg(x = c)^{\mathcal{M}}$. By (a): $T \vdash \neg\neg\exists x\,(x = c)^{\mathcal{M}}$. By theorem 2.3, T is closed under MR. Hence, $T \vdash \exists x\,(x = c)^{\mathcal{M}}$. ❑

We turn to the more Kleene style treatment. Let $T(e, x, p)$ be Kleene's T-predicate. Here e is the index of a partial recursive function, x is the (sequence of) input(s) and p is the computation. Let res be the elementary result extracting function. We write:

- $\{e\}x \simeq y$ for: $\exists p\,(T(e, x, p) \wedge \mathsf{res}(p) = y)$,
- $\{e\}x \not\simeq y$ for: $\exists p\,(T(e, x, p) \wedge \mathsf{res}(p) \not\simeq y)$,
- $\{e\}x \simeq_c y$ for: $\exists p \leq c\,(T(e, x, p) \wedge \mathsf{res}(p) = y)$,
- $\{e\}x \not\simeq_c y$ for: $\exists p \leq c\,(T(e, x, p) \wedge \mathsf{res}(p) \not\simeq y)$.

Since formalization in i-$I\Delta_0$ is rather unwieldy we will work with interpretations that also satify the axiom Ω_1. The predicate $T(e,x,p)$ can be represented under the usual coding as a $\Delta_0(\omega_1)$-formula. By tricks well known from the classical context, we can verify $\Delta_0(\omega_1)$-induction in i-$I\Delta_0 + \Omega_1$. Thus we may comfortably work with the T-predicate inside i-$I\Delta_0 + \Omega_1$ and verify the usual elementary facts, like unicity of output.

Church's Rule for T is the following rule.

CR $T \vdash \forall x \exists y\, A(x,y) \Rightarrow$ for some e,
$T \vdash \forall x \exists p\, T(e,x,p) \wedge \forall x, p\, (T(e,x,p) \rightarrow A(x, res(p)))$

T is closed under the Strong Church Rule if, for every A, $T + \neg A$ is closed under Church's Rule. E.g. HA is closed under Strong Church's Rule.

Theorem 5.5 *Suppose:*

i) $\mathcal{M}$ *is a prime interpretation of* i-$I\Delta_0 + \Omega_1$ *in* T,

ii) $\mathcal{M}$ *is provably non-standard in* T,

iii) T *is adequate for* $\mathcal{M}$,

iv) T *is closed under Church's Rule.*

Then $T \vdash \bot$.

Proof

Suppose $T \vdash \forall x\, (x < c)^{\mathcal{M}}$, for the numerical $\mathfrak{L}$-constant c. We have:

$$T \vdash \forall e\, (\neg\{e\}e \simeq_c 0 \vee \neg\{e\}e \not\simeq_c 0)^{\mathcal{M}}.$$

Hence: $T \vdash (\neg\{e\}e \simeq_c 0)^{\mathcal{M}} \vee (\neg\{e\}e \not\simeq_c 0)^{\mathcal{M}}$. Since $\mathcal{M}$ is prime and c provably non-standard, we find: $T \vdash \neg\{e\}e \simeq 0 \vee \neg\{e\}e \not\simeq 0$. So by closure under Church's Rule, we can find an index m such that:

$$\begin{array}{lll} T \vdash \text{“}\{m\} \text{ is total”} & \wedge & \forall e\, ((\{m\}e \simeq 0 \rightarrow \neg\{e\}e \simeq 0) \ \wedge \\ & & (\{m\}e \not\simeq 0 \rightarrow \neg\{e\}e \not\simeq 0)). \end{array}$$

We find:

$$\begin{array}{lll} T \vdash \text{“}\{m\} \text{ is total”} & \wedge & ((\{m\}m \simeq 0 \rightarrow \neg\{m\}m \simeq 0) \ \wedge \\ & & (\{m\}m \not\simeq 0 \rightarrow \neg\{m\}m \not\simeq 0)). \end{array}$$

It is immediate that: $T \vdash \bot$. ❑

Theorem 5.6 *Let c be a numerical constant of $\mathfrak{L}$. Suppose:*

i) *$\mathcal{M}$ is a prime interpretation of i-$I\Delta_0 + \Omega_1$ in T,*

ii) *T is adequate for $\mathcal{M}, c$,*

iii) *T is closed under the Strong Church's Rule.*

Then (a) $T \vdash \neg\neg\exists x\,(c = x)^{\mathcal{M}}$. Moreover (b) if:

iv) *T is closed under primitive recursive Markov's rule $\mathsf{MR}_{\mathsf{PR}}$, then: $T \vdash \exists x\,(c = x)^{\mathcal{M}}$.*

Proof

We prove (a) from theorem 5.5, in the same way as we proved (a) of corollary 5.4 from theorem 5.2. We prove (b). Since T is prime we have: $T \vdash (x = c)^{\mathcal{M}} \vee \neg(x = c)^{\mathcal{M}}$. Hence by Church's Rule: for some e:

$$T \vdash \text{“}e \text{ is total”} \quad \wedge \quad \forall x\,((\{e\}x \simeq 0 \to (x = c)^{\mathcal{M}}) \quad \wedge \quad (\{e\}x \not\simeq 0 \to \neg(x = c)^{\mathcal{M}})).$$

Hence $T \vdash (x = c)^{\mathcal{M}} \leftrightarrow \{e\}x \simeq 0$. By (a): $T \vdash \neg\neg\exists x\,\{e\}x \simeq 0$. So by Markov's Rule: $T \vdash \exists x\,\{e\}x \simeq 0$ and hence: $T \vdash \exists x\,(x = c)^{\mathcal{M}}$. ❑

After these results involving the constant c, we turn to results, which are in essence about Σ-definable elements. Here we can often obtain somewhat sharper results. We will say that $\mathcal{M}$ is Γ-disjunctive in T if $T \vdash (A \vee B)^{\mathcal{M}} \to (A^{\mathcal{M}} \vee B^{\mathcal{M}})$ for all $A, B \in \Gamma$.

Theorem 5.7 *Let $A(x, y)$ be in $\Delta_0(\omega_1)$ Suppose:*

1. *$\mathcal{M}$ is a Σ-disjunctive interpretation of i-$I\Delta_0 + \Omega_1$ in T,*
2. *$T \vdash \forall x\,(\exists y\,A(x, y))^{\mathcal{M}}$,*
3. *T satisfies full induction,*
4. *T is closed under the De Jongh Rule.*

Then: $T \vdash \forall x\,(\exists y\,A(x, y) \vee \bot^{\mathcal{M}})$.

Proof

Let $B(a) := \exists y\,A(a, y)$. Before proceeding, we must make a stipulation on the use of variables nested inside $\mathcal{M}$ and then a box. We ask that, e.g., $\Box_{T,b}A(a, b)^{\mathcal{M}}$ means: $\mathsf{Prov}_{T,b}(\underline{\#A(a, b)}^{\mathcal{M}})\{\underline{\#a} := a, \underline{\#b} := b\}$. This stipulation has the advantage that it avoids the use of $\mathcal{M}$ as a function inside the arithmetized context.

We want to find an R such that:

$$i\text{-}I\Delta_0 + \Omega_1 \vdash R(a, b) \leftrightarrow ((\Box_{T,b}(\neg R(a, b))^{\mathcal{M}}) \vee B(a)) \leq \Box_{T,b}R(a, b)^{\mathcal{M}}.$$

Now if it were given that $(.)^{\mathcal{M}}$ is provably recursive in $i\text{-}I\Delta_0 + \Omega_1, R$ could be found by an immediate application of the Gödel Fixed Point Lemma. Since such an assumption is quite plausible in practice – all p-time computable functions are provably recursive in $i\text{-}I\Delta_0{+}\Omega_1$ – we could simply add it to our list of conditions. However, a minor modification of the Fixed Point Argument delivers a slightly modified version, that is as good in practice. The reader who is content with the demand of provable recursiveness may well skip the argument immediately below.

We show: there are formulas $R(a,b)$, $G(a,b)$ and $H(a,b)$, such that:

- $i\text{-}I\Delta_0 + \Omega_1 \vdash R(a,b) \leftrightarrow ((\Box_{T,b} G(a,b)) \vee B(a)) \leq \Box_{T,b} H(a,b)$,
- $T \vdash G(a,b) \leftrightarrow (\neg R(a,b))^{\mathcal{M}}$ and $T \vdash H(a,b) \leftrightarrow R(a,b)^{\mathcal{M}}$.

(G and H have no other variables than those displayed.) We sketch the argument. $C[x := t]$ is the usual unformalized substitution function. Define:

$$\begin{array}{lll} A(x,y,a,b) & := & (\mathsf{Prov}_{T,b}(x\{\underline{\#x} := x, \underline{\#y} := y, \underline{\#a} := a, \underline{\#b} := b\}) \vee B(a)) \leq \\ & & \mathsf{Prov}_{T,b}(y\underline{\#x} := x, \underline{\#y} := y, \underline{\#a} := a, \underline{\#b} := b\}). \end{array}$$

Now define:

$$C(x,y,a,b) := (\neg A(x,y,a,b))^{\mathcal{M}}, \quad D(x,y,a,b) := (A(x,y,a,b))^{\mathcal{M}}.$$

Note that C and D are fixed standard formulas. Let their Gödelnumbers be $\#C$ and $\#D$. Take:

$$\begin{array}{lll} R(a,b) & := & A(\underline{\#C}, \underline{\#D}, a, b), \\ G(a,b) & := & (\neg A(x,y,a,b))^{\mathcal{M}}[x := \underline{\#C}, y := \underline{\#D}], \\ H(a,b) & := & A(x,y,a,b)^{\mathcal{M}}[x := \underline{\#C}, y := \underline{\#D}]. \end{array}$$

By theorem 4.8 we have immediately:

$$T \vdash G(a,b) \leftrightarrow (\neg R(a,b))^{\mathcal{M}} \text{ and } T \vdash H(a,b) \leftrightarrow R(a,b)^{\mathcal{M}}.$$

Note that for a sufficiently large m we will also have:

- $i\text{-}I\Delta_0 + \Omega_1 \vdash b \geq \underline{m} \rightarrow \Box_{T,b}(G(a,b) \leftrightarrow (\neg R(a,b))^{\mathcal{M}})$
- $i\text{-}I\Delta_0 + \Omega_1 \vdash b \geq \underline{m} \rightarrow \Box_{T,b}(H(a,b) \leftrightarrow R(a,b)^{\mathcal{M}})$.

The following are equivalent in $i\text{-}I\Delta_0 + \Omega_1$:

1. $R(a,b)$
2. $(\mathsf{Prov}_{T,b}(\underline{\#C}\{\underline{\#x} := \underline{\#C}, \underline{\#y} := \underline{\#D}, \underline{\#a} := a, \underline{\#b} := b\}) \vee B(a)) \leq$
   $\mathsf{Prov}_{T,b}(\underline{\#D}\{\underline{\#x} := \underline{\#C}, \underline{\#y} := \underline{\#D}, \underline{\#a} := a, \underline{\#b} := b\})$
3. $(\mathsf{Prov}_{T,b}(\underline{\#G}\{\underline{\#a} := a, \underline{\#b} := b\}) \vee B(a)) \leq$
   $\mathsf{Prov}_{T,b}(\underline{\#H}\{\underline{\#a} := a, \underline{\#b} := b\})$

4. $((\Box_{T,b}G(a,b)) \vee B(a)) \leq \Box_{T,b}H(a,b)$.

In the following we simply confuse, e.g. $\Box_{T,b}G$ with $\Box_{T,b}(\neg R)^{\mathcal{M}}$. It is easy to see that this is harmless, assuming that the m in the argument below is large enough to yield the desired equivalences. We have in T:

a)	$(B(a))^{\mathcal{M}}$			2
b)	$(R(a,b) \vee R^{\perp}(a,b))^{\mathcal{M}}$			a
c)	$R(a,b)^{\mathcal{M}} \vee R^{\perp}(a,b)^{\mathcal{M}}$			1,b
d)	$\Box_{T,m}R(a,b)^{\mathcal{M}} \vee \Box_{T,m}R^{\perp}(a,b)^{\mathcal{M}}$			4,c, some m
e)	$\Box_{T,m}R(a,m)^{\mathcal{M}} \vee \Box_{T,m}R^{\perp}(a,m)^{\mathcal{M}}$			d
f)	$\Box_{T,m}R(a,m)^{\mathcal{M}}$	$\rightarrow$	$(R(a,m) \vee R^{\perp}(a,m)) \wedge R(a,m)^{\mathcal{M}}$	
		$\rightarrow$	$((\Box_{T,m}(\neg R(a,m))^{\mathcal{M}} \vee B(a)) \wedge R(a,m)^{\mathcal{M}}) \vee (R^{\perp}(a,m) \wedge R(a,m)^{\mathcal{M}})$	
		$\rightarrow$	$(((\neg R(a,m))^{\mathcal{M}} \vee B(a)) \wedge R(a,m)^{\mathcal{M}}) \vee (R^{\perp}(a,m)^{\mathcal{M}} \wedge R(a,m)^{\mathcal{M}})$	
		$\rightarrow$	$\perp^{\mathcal{M}} \vee B(a)$	
g)	$\Box_{T,m}R^{\perp}(a,m)^{\mathcal{M}}$	$\rightarrow$	$(R(a,m) \vee R^{\perp}(a,m)) \wedge R^{\perp}(a,m)^{\mathcal{M}}$	
		$\rightarrow$	$(R(a,m) \wedge R^{\perp}(a,m)^{\mathcal{M}}) \vee (\Box_{T,m}R(a,m)^{\mathcal{M}} \wedge R^{\perp}(a,m)^{\mathcal{M}})$	
		$\rightarrow$	$R(a,m)^{\mathcal{M}} \wedge R^{\perp}(a,m)^{\mathcal{M}}$	
		$\rightarrow$	$\perp^{\mathcal{M}}$	
h)	$\perp^{\mathcal{M}} \vee B(a)$			e,f,g
p)	$\forall x\,(\perp^{\mathcal{M}} \vee \exists y\,A(x,y))$			h

❑

Remark 5.8 It is immediate from 5.7, that if $\mathcal{M} : \mathsf{HA} \rhd T$ and $\mathcal{M}$ is prime, then HA is Π_2-conservative over T. Note, on the other hand, that PA is also Π_2-conservative over T, but it is not prime interpretable in HA. ❑

Theorem 5.9 *Let $A(x,y)$ be in Δ_0. Suppose:*

1. *$\mathcal{M}$ is a Σ-prime interpretation of i-$I\Delta_0$ in T,*
2. *$T \vdash \forall x\,(\exists y\,A(x,y))^{\mathcal{M}}$,*
3. *$\mathcal{M}$ is Σ-complete in T,*
4. *T is closed under Church's Rule.*

Then: $T \vdash \forall x \exists y\,A(x,y)$.

Proof

Let $B(a) := \exists y\,A(a,y)$ and $E(b,a) := ((\{b\}(a) \simeq 0 \vee B(a)) \leq (\{b\}a \not\simeq 0))$. We have in T:

a)	$B(a)^{\mathcal{M}}$			
b)	$(E^{\perp}(b,a) \vee E(b,a))^{\mathcal{M}}$			a
c)	$E^{\perp}(b,a)^{\mathcal{M}} \vee E(b,a)^{\mathcal{M}}$			1,b
d)	$(\neg E(b,a))^{\mathcal{M}} \vee (\neg E^{\perp}(b,a))^{\mathcal{M}}$			c
e)	$\neg E(b,a) \vee \neg E^{\perp}(b,a)$			3,d
f)	"$\{m\}$ is total"	$\wedge$	$\forall b,a\,((\{m\}(b,a) \simeq 0 \rightarrow \neg E(b,a))$	
		$\wedge$	$(\{m\}(b,a) \not\simeq 0 \rightarrow \neg E^{\perp}(b,a)))$	4,e, some m
g)	$\forall a\,\{n\}a = \{m\}(n,a)$			Rec. Thm, some n
h)	"$\{n\}$ is total"	$\wedge$	$\forall a\,((\{n\}a \simeq 0 \rightarrow \neg E(n,a))$	
			$\wedge(\{n\}a \not\simeq 0 \rightarrow \neg E^{\perp}(n,a)))$	f,g
i)	"$\{n\}$ is total"	$\wedge$	$\forall a\,((\{n\}a \simeq 0 \rightarrow \{n\}a \not\simeq 0)$	
		$\wedge$	$(\{n\}a \not\simeq 0 \rightarrow (\{n\}a \simeq 0 \vee B(a))))$	h
j)	$B(a)$			i
k)	$\forall x \exists y\, A(x,y)$			j

❑

Theorem 5.10 *Let $A(x,y)$ be in Δ_0. Suppose:*

i) *$\mathcal{M}$ is a prime interpretation of i-$I\Delta_0$ in T,*

ii) *$\mathcal{M}$ is Σ-complete in T,*

iii) *$T \vdash \mathsf{CT}_0$.*

Then: $T \vdash \forall x\,(\exists y\, A(x,y))^{\mathcal{M}} \rightarrow \forall x \exists y\, A(x,y)$.

Proof

The proof is a minor variation of the proof of theorem 5.9. ❑

Remark 5.11 The proofs of the results above bear strong resemblance to Friedman's work on the Disjunction and the Existence Property (see Friedman [75]). Still Friedman's result is not precisely the same. We provide a result that is much closer to the original Friedman proof.

Suppose $(S(a))^{\mathcal{M}}$ is Σ for any $S(a)$ in Σ. We suppose that $\mathcal{M}$ interprets i-$I\Delta_0 + \mathsf{Exp}$ in T. Let B be a Σ-formula. Let True be the usual Σ-truthpredicate. We write $\mathsf{True}^{\mathcal{M}}(t)$ for $\mathsf{True}(x)^{\mathcal{M}}[x := t]$. By the Gödel Fixed Point Lemma, we find R such that:

$$i\text{-}I\Delta_0 \vdash R \leftrightarrow (\mathsf{True}^{\mathcal{M}}(R^{\perp}) \vee B) \leq \mathsf{True}^{\mathcal{M}}(R).$$

Suppose $\mathcal{M}$ is Σ-prime and Σ-complete in T. We have in T:

$$\begin{array}{rcl} B^{\mathcal{M}} & \rightarrow & (R \vee R^{\perp})^{\mathcal{M}} \\ & \rightarrow & R^{\mathcal{M}} \vee R^{\perp \mathcal{M}} \\ & \rightarrow & \mathsf{True}(R)^{\mathcal{M}} \vee \mathsf{True}(R^{\perp})^{\mathcal{M}} \\ & \rightarrow & \mathsf{True}^{\mathcal{M}}(R) \vee \mathsf{True}^{\mathcal{M}}(R^{\perp}) \\ & \rightarrow & R \vee R^{\perp}. \end{array}$$

And:

$$\begin{array}{rcl} R & \rightarrow & (\mathsf{True}^{\mathcal{M}}(R^{\perp}) \vee B) \wedge R^{\mathcal{M}} \\ & \rightarrow & (R^{\perp \mathcal{M}} \vee B) \wedge R^{\mathcal{M}} \\ & \rightarrow & \perp^{\mathcal{M}} \vee B. \end{array}$$

Moreover:

$$\begin{array}{rcl} R^{\perp} & \rightarrow & \mathsf{True}^{\mathcal{M}}(R) \wedge R^{\perp \mathcal{M}} \\ & \rightarrow & R^{\mathcal{M}} \wedge R^{\perp \mathcal{M}} \\ & \rightarrow & \perp^{\mathcal{M}}. \end{array}$$

Ergo: $T \vdash B^{\mathcal{M}} \rightarrow B \vee \perp^{\mathcal{M}}$. ∎

6 Notes on bounded interpretations

There is a strong feeling that there is only a very restricted possibility to have bounded interpretations of HA into itself. We provide an example of a bounded interpretation satisfying some constraints, followed by a negative result.

We sketch the construction of an interpretation of HA in HA which is primitive recursive, bounded and prime. Define complexity classes Γ_i for $\mathfrak{L}_{\mathsf{Ar}}$-formulas as follows.

- $\Gamma_0 := \Sigma$, where we, locally, take Σ as being closed under $\wedge, \vee$ and $\exists$
- $A \in \Gamma_{i+1}$ iff there is a computation sequence σ such that $A = (\sigma)_{\mathsf{length}(\sigma)-1}$, and, for all $j < \mathsf{length}(\sigma)$, $(\sigma)_j \in \Sigma$ or $((\sigma)_j \notin \Sigma$ and
  $(\sigma)_j = (B \wedge C)$ or $(\sigma)_j = (B \vee C) \Rightarrow \exists k, m < j\, ((\sigma)_k = B$ and $(\sigma)_m = C)$, and
  $(\sigma)_j = \exists x\, B \Rightarrow \exists k < j\, (\sigma)_k = B$, and
  $(\sigma)_j = (B \rightarrow C) \Rightarrow B, C \in \Gamma_i$, and
  $(\sigma)_j = \forall x\, B \Rightarrow B \in \Gamma_i$.

We define truthpredicates True_i for the Γ_i in such a way that each definition can be formalized in HA and such that its properties can be verified.

- True_0 is the usual Σ-truthpredicate.

- $\mathsf{True}_{i+1}(A)$ iff there is a computation sequence σ such that $A = (\sigma)_{\mathsf{length}(\sigma)-1}$, and, for all $j < \mathsf{length}(\sigma)$, $(\sigma)_j \in \Sigma$ and $\mathsf{True}_0((\sigma)_j)$ or $((\sigma)_j \notin \Sigma$ and
  $(\sigma)_j = (B \wedge C) \Rightarrow \exists k, m < j\,((\sigma)_k = B$ and $(\sigma)_m = C)$, and
  $(\sigma)_j = (B \vee C) \Rightarrow \exists k < j\,((\sigma)_k = B$ or $(\sigma)_k = C)$, and
  $(\sigma)_j = \exists x\, B \Rightarrow \exists k < j \exists n < \sigma\,(\sigma)_k = B[x := \underline{n}]$, and
  $(\sigma)_j = (B \to C) \Rightarrow (\mathsf{True}_i(B) \to \mathsf{True}_i(C))$, and
  $(\sigma)_j = \forall x\, B \Rightarrow \forall n\, \mathsf{True}_i(B[x := \underline{n}])$.

We now modify the notion of restricted provability of subsection 2.2 as follows. Let T be a theory in $\mathfrak{L}_{\mathsf{Ar}}$. We write T_n for the theory axiomatized by the finitely many axioms of i-$I\Delta_0 + \mathsf{Exp}$, plus the axioms of T, which are in Γ_n. We write $\mathsf{Prov}_{T,n}$ for the formalization of provability in T_n. We have the following analogue of theorem 2.1:

Fact 6.1 Suppose T is a *finite* $\mathfrak{C}$-axiomatized extension of HA in $\mathfrak{L}_{\mathsf{Ar}}$. (see subsection 2.2 for the definition of $\mathfrak{C}$.) We have, for all n, and for all formulas A with free variables $\mathbf{x}$,

$$T \vdash \forall \mathbf{x}\,(\Box_{T,n} A \to A).$$

And (using UC for: 'the universal closure of') even:

$$\mathsf{HA} \vdash \forall x \forall A \in \mathsf{FOR}\, \Box_T \mathsf{UC}(\Box_{T,x} A \to A).$$

❑

Proof

As usual we find a cut-free x-proof p of A and show by induction on subproofs that p's conclusion is true. The only problem is the fact that in our new circumstances T_i has infinitely many axioms. Thus we cannot handle the axioms case for case. We are saved by the fact that – due to our stipulation that T is a *finite extension* of HA – the only axioms that we have, are finitely many special axioms plus infinitely many induction axioms. The induction axioms have a *schematic* form. Thus (in extremely sloppy notation):

$$\mathsf{True}_i((A[x := \underline{0}] \wedge \forall x\,(A \to A[x := Sx])) \to \forall x\, A),$$

where A is a *variable*, is equivalent to:

$$(\mathsf{True}_{i-2}(A[x := \underline{0}]) \wedge \forall x\,(\mathsf{True}_{i-2}(A[x := \underline{x}]) \to$$

$$\mathsf{True}_{i-2}(A[x := \underline{Sx}]))) \to \forall x\, \mathsf{True}_{i-2}(A[x := \underline{x}]).$$

But this new fomula is itself an instance of induction with induction formula $B = \mathsf{True}_{i-2}(A[x := \underline{x}])$ and with A as free parameter. ❑

We define $[T]_n$ just as in subsection 2.2, using the new notion of restricted provability. (i)-(vi) of subsection 2.2 are verified without problems. Now note that if A is in Γ_i, then $[T]_n A$ is also in Γ_i. Thus we get the following strengthened version of (ix) of subsection 2.2:

ix$^+$) $\Gamma \vdash_{T,n} A \Rightarrow [T]_n\Gamma \vdash_{T,n} [T]_n A$ (verifiably in HA).

Consider the following Feferman Predicate for HA:

$$\Delta_{\mathsf{HA}} A :\Leftrightarrow \exists x\, (\Box_{\mathsf{HA},x} A \wedge \forall S \in \Sigma_1\, (\Box_{\mathsf{HA},x} S \rightarrow \mathsf{True}_0(S))).$$

Take $\mathcal{M}(A) := \Delta_{\mathsf{HA}} A$. We easily see – using fact 6.1 – that $\mathcal{M} : \mathsf{HA} \rhd \mathsf{HA}$. Clearly, $\mathcal{M}$ is primitive recursive, bounded and consistent.

Theorem 6.2 *$\mathcal{M}$ is disjunctive.*

Proof

Reason in HA. Suppose $\Delta_{\mathsf{HA}}(B \vee C)$. Then for some x : $\Box_{\mathsf{HA},x}(B \vee C)$ and $\forall S \in \Sigma_1\, (\Box_{\mathsf{HA},x} S \rightarrow \mathsf{True}_0(S))$. By the the closure of HA_x under the De Jongh translation $[\mathsf{HA}]_x$, we find: $\Box_{\mathsf{HA},x}(\Box_{\mathsf{HA},x} B \vee \Box_{\mathsf{HA},x} C)$. Hence, by Σ_1-reflection: $(\Box_{\mathsf{HA},x} B \vee \Box_{\mathsf{HA},x} C)$, and so: $(\Delta_{\mathsf{HA}} B \vee \Delta_{\mathsf{HA}} C)$. ❑

The formulas of the intuitionistic propositional calculus IPC modulo provable equivalence form the Rieger Nishimura Lattice (see Troelstra & van Dalen [88b], 707). $g_n(p)$ is a standard enumeration of these formulas.

Theorem 6.3 *[De Jongh] Let A be a sentence of $\mathfrak{L}_{\mathsf{Ar}}$, such that $\mathsf{HA} \nvdash \neg\neg A$ and $\mathsf{HA} \nvdash \neg\neg A \rightarrow A$. Then there is no $\mathfrak{L}_{\mathsf{Ar}}$-formula $B(x)$, such that for all n:*

$$\mathsf{HA} \vdash g_n(A) \leftrightarrow B(\underline{n}).$$

Proof

This theorem is an immediate consequence of theorems 2.1 and 5.1 of De Jongh [82]. ❑

A formula is *almost negative* if it is built up from Σ-formulas using all connectives except $\vee$ and $\exists$. Our definition is a minor variation of Definition 3.2.9 of Troelstra [73], 193, alternatively: Definition 4.4, 197 of Troelstra & van Dalen [88a].

Theorem 6.4 $\mathsf{HA} + \mathsf{ECT}_0$ *is conservative over* HA *w.r.t almost negative formulas.*

Proof

This result is an immediate consequence of Troelstra [73], Lemma 3.2.11, 193 and Corollary 3.2.19, 196. Alternatively see Troelstra & van Dalen [88a], Proposition 4.5, 197 and Theorem 4.10, 199. ❑

Theorem 6.5 *There is no provably recursive, bounded, prime interpretation* $\mathcal{M}$ *in* HA *which commutes with implication.*

Proof

Suppose $\mathcal{M}$ is provably recursive, bounded and prime. Since $\mathcal{M}$ is bounded, there is a truthpredicate $\mathsf{T}(x)$ for its range. Suppose also that $\mathcal{M}$ commutes with implication. Let S be a Σ-sentence such that $\mathsf{HA} \nvdash \neg\neg S$ and $\mathsf{HA} \nvdash \neg\neg S \to S$. A well known example of such a formula is $\Box_{\mathsf{HA}} \bot$. As is easily seen, there is a primitive recursive function f such that $f(n)$ is the Gödelnumber of $(g_n(S))^{\mathcal{M}}$. Take $B(x) := \mathsf{T}(f(x))$. We have: $\mathsf{HA} \vdash g_n(S^{\mathcal{M}}) \leftrightarrow (g_n(S))^{\mathcal{M}} \leftrightarrow B(\underline{n})$. To derive a contradiction it is, by theorem 6.3, sufficient to show that $\mathsf{HA} \nvdash \neg\neg S^{\mathcal{M}}$ and $\mathsf{HA} \nvdash \neg\neg S^{\mathcal{M}} \to S^{\mathcal{M}}$. Suppose e.g. $\mathsf{HA} \vdash \neg\neg S^{\mathcal{M}} \to S^{\mathcal{M}}$. It follows that $\mathsf{HA} + \mathsf{CT}_0 \vdash \neg\neg S^{\mathcal{M}} \to S^{\mathcal{M}}$. Clearly $\mathcal{M}$ is a prime interpretation in $\mathsf{HA} + \mathsf{CT}_0$. By theorem 4.6 and theorem 5.10: $\mathsf{HA} + \mathsf{CT}_0 \vdash S^{\mathcal{M}} \leftrightarrow S$. Ergo: $\mathsf{HA} + \mathsf{CT}_0 \vdash \neg\neg S \to S$. Since $\neg\neg S \to S$ is almost negative, by theorem 6.4, we find that $\mathsf{HA} \vdash \neg\neg S \to S$. Quod non. The proof for $\neg\neg S$ is similar. ❑

The present result is not all that satisfactory, since commuting with implication is not a frequent property among useful interpretations. So improvements of theorem 6.5 would be wellcome.

A The Henkin construction

Let T be an RE theory of Intuitionistic Predicate Logic and let A be a sentence in the language of T. We assume that the language of T is relational, except for the possible presence of constants. (If were is not, we could first convert it to the relational format, using the usual procedure.) We show that in $\mathsf{PA} +$ "$T \nvdash A$" we can formalize the Henkin construction of a Kripke model of T, which does not force A. This gives us an interpretation $\mathcal{K}$ of T into $\mathsf{PA} +$ "$T \nvdash A$", which commutes with atomic formulas, $\wedge$, $\vee$, $\exists$, behaves in a more complicated way in the cases of $\to$ and $\forall$. We have for sentences A, B:

$$T \vdash B \Rightarrow \mathsf{PA} + \text{“}T \nvdash A\text{”} \vdash B^{\mathcal{K}} \text{ and } \mathsf{PA} + \text{“}T \nvdash A\text{”} \vdash \neg(A^{\mathcal{K}}).$$

We work informally inside of $\mathsf{PA} +$ "$T \nvdash A$". For every $n > 0$ we define X_n as the set of numbers divided by one of the first n prime numbers. let $Y_{n+1} := X_{n+1} \setminus X_n$. Let $\mathfrak{L}$ be the language of T. We can arrange the coding of syntax in such a way that fresh constants can be coded by pairs, say, $\langle 723, i\rangle$. We write

c_i for $\langle 723, i\rangle$. For any set of numbers I, we write $\mathfrak{L}(I)$ for the result of adding the constants c_i for $i \in I$ to $\mathfrak{L}$. Let $\mathfrak{L}_n$ be $\mathfrak{L}(X_n)$.

The worlds of our Henkin model will be given by triples $\langle n, B, C\rangle$. Here:

- B is a Π_2-formula in one free variable, defining a set of $\mathfrak{L}_n$-sentences Δ, that extends T and is saturated, i.e. for $\mathfrak{L}_n$-sentences D, E and $\exists x\, F$:
  - $\Delta \vdash D \Rightarrow D \in \Delta$
  - $\Delta \nvdash \bot$
  - $\Delta \vdash D \vee E \Rightarrow \Delta \vdash D$ or $\Delta \vdash E$,
  - $\Delta \vdash \exists x\, F \Rightarrow \Delta \vdash F[x := c_i]$ for some $i \in X_n$.
- C is a Π_2-formula in one free variable, defining the set of $\mathfrak{L}_n$-sentences not in Δ.

We will write, informally, $\langle n, \Delta\rangle$ for our worlds, where Δ is the Δ_2 set of $\mathfrak{L}_n$-sentences presented by B and C. Since PA contains a truth-predicate for Π_2-formulas our definition makes sense inside PA. The domain associated to a world $\langle n, \Delta\rangle$ is simply the set of c_i for i in X_n modulo Δ-provable identity. We write for $P(c)$ atomic in $\mathfrak{L}_n$:

$$\langle n, \Delta\rangle \models P(c) :\Leftrightarrow P(c) \in \Delta.$$

Etcetera. The ordering relation of our model is as follows:

$$\langle n, \Delta\rangle \preceq \langle n', \Delta'\rangle \Leftrightarrow n \leq n' \text{ and } \Delta \subseteq \Delta'.$$

The forcing relation is extended to the full language present at a world in the usual way. We do not do this uniformly in PA, we just provide, for each E, a PA-formula $\langle n, \Delta\rangle \models E$, in which n and Δ are variables. We will show $\langle n, \Delta\rangle \models E \Leftrightarrow E \in \Delta$.

We want to show that if $T \nvdash A$, then there is a world $\langle n, \Delta\rangle$, such that $T \subseteq \Delta$ and $A \notin \Delta$. Moreover we want to show for each $B \in \mathfrak{L}_n : \langle n, \Delta\rangle \models B \Leftrightarrow B \in \Delta$. The following lemma is sufficient.

Lemma A.1 [PA] Suppose Γ is a Δ_2-theory in $\mathfrak{L}_n$. Let U_0 and V_0 be finite sets of $\mathfrak{L}_{n+1}$-sentences, such that $\Gamma, U_0 \nvdash V_0$. (The "V_0" after the provability sign gets the disjunctive reading.) Then we can explicitely give a world $\langle n+1, \Delta\rangle$, such that $\Gamma, U_0 \subseteq \Delta$ and $\Delta \nvdash V_0$. ❑

Proof

Let $d_1, d_2, \ldots$ be an effective enumeration of the constants c_j with index in Y_{n+1}, which are not in U_0, V_0. We effectively enumerate the sentences of $\mathfrak{L}_{n+1}$ as, say, $C_1, C_2, \ldots$ in such a way that:

- d_i occurs in $C_j \Rightarrow i \leq j$,

- $C_i = \exists x\, D \Rightarrow C_{i+1} = D[x := d_{i+1}]$,
- for any $E \in \mathfrak{L}_{n+1}$, we can primitive recursively find j such that $E = C_j$.

We first describe the construction of Δ and verify its properties and then worry about its arithmetical complexity. We define a sequence of pairs of finite sets of $\mathfrak{L}_{n+1}$-sentences U_n, V_n as follows.

- U_0, V_0 are given.
- If $\Gamma, U_i, C_{i+1} \nvdash V_i$, then $U_{i+1} := U_i \cup \{C_{i+1}\}$, $V_{i+1} := V_i$.
- If $\Gamma, U_i, C_{i+1} \vdash V_i$, then $U_{i+1} := U_i$, $V_{i+1} := V_i \cup \{C_{i+1}\}$.

We take Γ the union of the U_i's.

By an easy induction, using the properties of disjunction, we find that $\Gamma, U_i \nvdash V_i$. It follows that $\Delta \nvdash V_0$ and that $\Gamma, U_0 \subseteq \Delta$.

Δ is saturated. We treat the case of existential quantification. Suppose $\Delta \vdash \exists x\, D$. Say $\Gamma, U_j \vdash \exists x\, D$ and $C_{i+1} = \exists x\, D$. It follows that $\Gamma, U_i, C_{i+1} \nvdash V_i$ (otherwise we would have $\Gamma, U_{\max(i,j)} \vdash V_{\max(i,j)}$). It follows that $\Gamma, U_{i+2}, D[x := d_{i+2}] \nvdash V_{i+2}$, since d_{i+2} does not occur in $\Gamma, U_{i+2}, D, V_{i+2}$. Since $C_{i+2} = D[x := d_{i+2}]$, we find $D[x := d_{i+2}] \in U_{n+2}$.

Finally we check that Δ is Δ_2 in PA. First note that the 'decisions' we need in our construction are Δ_2, since Γ-provability is supposed to be Δ_2 and:

$$\Gamma, U_i, C_{i+1} \vdash V_i \Leftrightarrow \Gamma \vdash \mathfrak{Cl}_{i+1}(((\bigwedge U_i \wedge C_{i+1}) \rightarrow \bigvee V_i)).$$

Here $\mathfrak{Cl}_{i+1}$ is the result of first replacing the Y_{i+1}-indexed constants by fresh variables and then taking the universal closure of the resulting formula.

We can represent our recursive definition of the U's and V's by coding the sequence of yes and no decisions of the construction as a sequence of 0's and 1's. Say, at place $i+1$, we put 0 in the sequence if $\Gamma, U_i, C_{i+1} \nvdash V_i$, and 1 if $\Gamma, U_i, C_{i+1} \vdash V_i$. If we have such a sequence σ of length k we can read off the $U_i := U_i(\sigma)$ and $V_i := V_i(\sigma)$ for $i \leq k$ primitive recursively from σ. Given E, we can primitive recursively find $j := j(E)$ such that $E = C_j$. From j we can derive a (primitive recursive) estimate $k(E)$ of the 0,1-sequences of length j. We find:

$$\begin{aligned} E \in \Gamma \quad \Leftrightarrow \quad & \exists \sigma \leq F(E)\, [\mathsf{length}(\sigma) = j(E) \text{ and } (\sigma)_{j(E)} = 0 \wedge \\ & \forall i < j(E)\ P(U_i(\sigma), V_i(\sigma), U_{i+1}(\sigma), V_{i+1}(\sigma))]. \end{aligned}$$

Here P is the obvious Δ_2-condition (incorporating inclusion of the given finite U_0 and V_0). Hence Γ is Δ_2. ❑

We show that for all $B \in \mathfrak{L}_n$ and for all worlds $\langle n, \Delta\rangle$: $\langle n, \Delta\rangle \models B \Leftrightarrow B \in \Delta$, by (external) induction on the complexity of B. We just treat the case that $B = \forall x\, C$. Consider a world $\langle n, \Gamma\rangle$. From right to left is trivial. We prove from left to right by contraposition. Suppose $B \notin \Gamma$. So $\Gamma \nvdash \forall x\, C$. Let c be the first

constant with index in Y_{n+1}. Clearly $\Gamma \nvdash C[x := c]$. Apply the lemma with $U_0 := \emptyset$, $V_0 := \{C[x := c]\}$. We find a world $\langle n+1, \Delta\rangle \succeq \langle n, \Gamma\rangle$, such that $C[x := c] \notin \Delta$. By the $IH : \langle n, \Gamma\rangle \not\models C[x := c]$.

Finally, if $T \nvdash A$, we can find by the lemma a node $\langle 1, \Delta\rangle$ such that $T \subseteq \Delta$ and $A \notin \Delta$.

Suppose T is a theory containing Ar. We can associate with each number n in a primitive recursive way the $\mathfrak{L}_1$-constant e_n corresponding to the sentence "$\exists x\ x = \underline{n}$". We translate $B(x_1, \ldots, x_n)$ to: "$\#B[\#x_1 := e_{x_1}, \ldots, \#x_n := e_{x_n}] \in \Delta$". It is easy to see that this yields an interpretation in our sense. This interpretation evidently commutes with atomic formulas, conjunction, disjunction and existential quantification. Reasoning outside of PA, we see: $\mathsf{PA} \vdash \neg(A \in \Delta)$ and $\mathsf{PA} \vdash (\Box_T B \to B \in \Delta)$. Hence, if $T \vdash B$, then $\mathsf{PA} \vdash \Box_T B$ and so $\mathsf{PA} \vdash B \in \Delta$.

Acknowledgement

I thank Volodya Shavrukov for his careful reading of the penultimate version of this paper.