The Closed Fragment of the Interpretability Logic of PRA with a Constant for $I\Sigma_1$.

Joost J. Joosten

February 24, 2003

Abstract

In this paper we characterize the closed fragment of the enriched provability logic of PRA and call the logic characterizing it **PGL**. These logics are enriched in the sense that they contain a constant symbol `S` which denotes the arithmetical sentence axiomatizing $I\Sigma_1$. We also determine the closed fragment of the interpretability logic of PRA with a constant for $I\Sigma_1$ which we baptize **PIL**. We show that $I\Sigma_1$ proves the consistency of PRA on a cut. By restricting the possible substitutions in Solovay's theorem we obtain a rough upperbound for the full interpretability logic of PRA.

1 Introduction

In this section we provide a plan of this paper, and a motivation of our study. We also fix some notation. The reader is suggested to skip Subsections 1.3 and 1.4 and only consult them if necessary in the rest of the paper.

1.1 Plan of this paper

The paper consists, apart from the introduction, of four sections. In the beginning of each section we give a brief summary of what is done in that section.

Section 2 fully characterizes the closed fragment of the provability logic of PRA with a constant for $I\Sigma_1$, with and without reflection. A modal semantics is provided and questions concerning decidability are dealt with.

In Section 3 these results are generalized to the setting of interpretability. We give two proofs of the arithmetical soundness of our logic **PIL**.

In Section 4 we make some remarks on the full interpretability logic (so, not just the closed fragment) of PRA and related systems. By a technique of restricting the possible substitutions in Solovay's theorem we obtain two "new" logics. One of them yields a rough upperbound for the full interpretability logic of PRA.

The final section, Section 5, is an appendix and gives a proof of the fact that $I\Sigma_1$ proves the consistency of PRA on a cut. This fact was already proved in Section 3 but the second proof employs entirely different methods.

1.2 Why do we think our study is interesting?

Interpretations in the form we will consider them have been around for quite a while in common mathematical practice. A good example is the interpretation of non-euclidean geometry in euclidean geometry. As a meta-mathematical tool interpretations were first introduced by Tarski in full generality in [TMR53] where they were used to show relative consistency and undecidability of theories.

The notion of interpretability we will study is essentially the same as in [TMR53]. Thus, an interpretation $\mathcal{K}$ of a theory T in a theory S —we write $\mathcal{K} : S \rhd T$— is nothing more but a translation of formulas of T to formulas of S such that the translation of any theorem of T is provable in S. In case such a translation exists we say that S interprets T or that T is interpretable in S and write $S \rhd T$. As in [TMR53] we are interested in relative interpretability. This means that in S we have a domain function $\delta(x)$ to which all our quantifiers are restricted/relativized. A precise and formal definition of relative interpretability can be found in, for example, [dJJ98] or [Vis97].

In these references and especially in [Vis91] the formalization of interpretability is studied. This gives rise to interpretability logics with a binary modal operator $\rhd$ for formalized interpretability. Just as in the case of provability logics we have that a modal sentence $A \rhd B$ is a valid principle for a theory T if for any arithmetical realization $*$ holds $T \vdash (T \cup \{A^*\}) \rhd (T \cup \{B^*\})$. Often $T + A^*$ will be written instead of $T \cup \{A^*\}$. Sometimes we will write $A^* \rhd_T B^*$ for $(T + A^*) \rhd (T + B^*)$. We will denote both the modal operator and the formalized notion of interpretability by the same symbol $\rhd$ but this will hardly lead to any confusion.

As the definition of interpretability invokes that of provability it does not come as a surprise that interpretability and provability logics are closely related. As a matter of fact, provability logics are literally included in the interpretability logics.

The interpretability logic for essentially reflexive theories has been determined independently by Berarducci ([Ber90]), and Shavrukov ([Sha88])

and is called **ILM**. Also the situation is known for finitely axiomatized theories in which case the logic is called **ILP** as was studied in [Vis90a].

No interpretability logic is known for a theory that is neither essentially reflexive nor finitely axiomatizable. PRA is such a theory. Thus we find it interesting to investigate the interpretability logic of this theory. More insight in the interpretability logic of PRA, from now on **IL**(PRA), can also shed some light on the question what interpretability principles hold in any reasonable theory as studied in Joosten and Visser, [JV00].

In this paper we constrain ourselves to the closed fragment of **IL**(PRA), that is, modal formulas without propositional variables. It is shown in an article by Hájek and Svejdar, [Hv91], that the closed fragment of any interpretability logic extending[1] **ILF** has the same characterization as the closed fragment of **GL**. It is easily seen that **IL**(PRA) indeed does extend **ILF**.

We have chosen to add an extra constant to our closed fragment that denotes the sentence axiomatizing $I\Sigma_1$. By writing $I\Sigma_1$ we will refer both to the finitely axiomatizable theory and to the finite axiom axiomatizing it. We can thus study what these theories have to say about each others provability and interpretability behaviour.

In this respect our enterprise is rather akin to a certain part of Beklemishev's paper [Bek96] on the classification of bimodal logics. As an example of his results he gives the provability logic (not just the closed fragment) of PRA with a constant for $I\Sigma_1$. The closed fragment of this logic is just the logic **PGL** we present in Section 2. We have chosen to give explicit proofs for the correctness and completeness of **PGL** again, so that we can easily extend them to the situation where interpretability is added to the vocabulary in Section 3.

This paper also is reminiscent of Visser's paper on exponentiation [Vis92]. In that paper the closed fragment of the interpretability logic of the arithmetical theory $\mathbf{\Omega}$ is presented with an additional constant exp in the language denoting the Π_2-formula stating the totality of the exponential function. (The theory $\mathbf{\Omega}$ is $I\Delta_0 + \Omega_1$. We refer the reader to consult [HP93] for definitions of the ω_n functions, definable cuts and other basic notions.)

A fundamental difference between Visser's [Vis92] and our paper is that although $I\Sigma_1$ is a proper extension of PRA, no new recursive functions are proved to be total, as $I\Sigma_1$ is a Π_2-conservative extension of PRA. In this sense the gap between PRA and $I\Sigma_1$ is smaller than the gap between $\mathbf{\Omega}$ and $\mathbf{\Omega} + \mathsf{exp}$. This difference is also manifested in the corresponding logics already when we just constrain ourselves to

[1] The logics and other notions mentioned in the introduction will be defined later on in the paper. Alternatively a reference is provided.

provability. For example we have that

$$\mathrm{PRA} + \mathsf{Con}(\mathrm{PRA}) \vdash \mathsf{Con}(\mathrm{I}\Sigma_1),$$

whereas

$$\mathbf{\Omega} + \mathsf{Con}(\mathbf{\Omega}) \nvdash \mathsf{Con}(\mathbf{\Omega} + \mathsf{exp}).$$

Actually even $\mathbf{\Omega} + \mathsf{exp} + \mathsf{Con}(\mathbf{\Omega})$ does not even prove $\mathsf{Con}(\mathbf{\Omega} + \mathsf{exp})$. It does hold however that $\mathbf{\Omega} + \mathsf{Con}(\mathsf{Con}(\mathbf{\Omega})) \vdash \mathsf{Con}(\mathbf{\Omega} + \mathsf{exp})$ and there are more similarities. We have that $\mathsf{Con}(\mathrm{PRA})$ is not provable in $\mathrm{I}\Sigma_1$. Similarly $\mathsf{Con}(\Omega)$ is not provable in $\mathbf{\Omega}+\mathsf{exp}$. In turn $\mathrm{I}\Sigma_1$ is not provable in PRA together with any iteration of consistency statements and the same holds for exp and $\mathbf{\Omega}$.[2]

The interpretability logics have similarities and differences too. For example we have that $\mathrm{PRA} \rhd \mathrm{PRA} + \neg\mathrm{I}\Sigma_1$ and $\mathbf{\Omega} \rhd \mathbf{\Omega} + \neg\mathsf{exp}$. Also $\mathrm{PRA} + \mathsf{Con}(\mathrm{PRA}) \rhd \mathrm{I}\Sigma_1$ and $\mathbf{\Omega} + \mathsf{Con}(\mathbf{\Omega}) \rhd \mathbf{\Omega} + \mathsf{exp}$. On the other hand $\mathrm{I}\Sigma_1 \not\rhd \mathrm{PRA} + \mathsf{Con}(\mathrm{PRA})$ whereas $\mathbf{\Omega} + \mathsf{exp} \rhd \mathbf{\Omega} + \mathsf{Con}(\mathbf{\Omega})$. However we do have that $\mathrm{I}\Sigma_1 \rhd \mathbf{\Omega} + \mathsf{Con}(\mathrm{PRA})$. We have that $\mathrm{I}\Sigma_1 \not\rhd \mathrm{PRA} + \mathsf{Con}(\mathrm{PRA})$ but PRA itself cannot see this. PRA can only see that $\mathrm{I}\Sigma_1 \rhd \mathrm{PRA} + \mathsf{Con}(\mathrm{PRA}) \rightarrow \neg\mathsf{Con}(\mathrm{PRA})$.

I would like to thank Lev Beklemishev, Volodya Shavrukov and Albert Visser for their multitude of good suggestions and discussions. Also a word of gratitude is due to Vincent van Oostrom, Dimitri Hendriks, Sander Bruggink, Dick de Jongh, Evan Goris, Rosalie Iemhoff, Nick Bezhanishvili, Panikovsky, Clemens, Herr Kupke, Yoav Seginer, Maarten Janssen, and probably many others too.

1.3 What is Primitive Recursive Arithmetic?

The base theory in this enterprise is PRA which is a system of arithmetic that goes by by many different formulations. We will briefly mention these formulations here and then stick to one of them. In a rudimentary form PRA was first introduced by Skolem in 1923 ([Sko67]). The emergence of PRA is best understood in the light of Hilbert's programme and finitism (see [Tai81]) or instrumentalism as Ignjatovic calls it in [Ign90].

Since Π_1-sentences or open formulas played a prominent role in Hilbert's programme, the first versions of PRA were formulated in a

[2] It is known that $\mathrm{I}\Sigma_1 \equiv \mathrm{RFN}_{\Pi_3}(\mathrm{EA})$ and by Fact 2.3 this schema is not contained in any Σ_3-extension of EA. Consistency statements are all Π_1-sentences. For the case of $\mathbf{\Omega}$ and exp reason as follows. Take any non-standard model of true arithmetic together with the set $\{2^c > \omega_2^k(c) \mid k \in \omega\}$. Take the smallest set containing c being closed under the ω_2 function. Consider the initial segment generated by this set. This initial segment is a model of $\mathbf{\Omega}$ and of all true Π_1 sentences but clearly not closed under exp.

quasi-equational setting without quantifiers but with a symbol for every primitive recursive function. (See for example Goodstein [Goo57], or Schwartz [Sch87a], [Sch87b].)

Other formulations are in the full language of predicate logic and also contain a function symbol for every primitive recursive function. The amount of induction can either be for Δ_0-formulas or for open formulas. Both choices yield the same set of theorems. This definition of PRA has, for example, been used in [Smo77].[3]

In this paper we will associate to each arithmetical theory T in a uniform way a proof predicate $\Box_T$ as is done by Feferman in [Fef60]. Thus, we will also have the obvious properties of this predicate like $\Box_{T+\varphi}\psi \leftrightarrow \Box_T(\varphi \to \psi)$ available in any theory of some reasonable minimal strength. We will also extensively make use of reflection principles.

For a theory T and a class of formulas Γ we define the uniform reflection principle for Γ over T to be a set of formulas in the following way: $\mathrm{RFN}_\Gamma(T) := \{\forall x\, (\Box_T \gamma(\dot{x}) \to \gamma(x)) \mid \gamma \in \Gamma\}$. This set of formulas is often equivalent to a single formula also denoted by $\mathrm{RFN}_\Gamma(T)$. For ordinals $\alpha \leq \omega$ we define $(T)^\Gamma_0 := T$, $(T)^\Gamma_{\alpha+1} := (T)^\Gamma_\alpha + \mathrm{RFN}_\Gamma((T)^\Gamma_\alpha)$ and $(T)^\Gamma_\omega := \cup_{\beta<\omega}(T)^\Gamma_\beta$. This can be extended to transfinite ordinals, provided an elementary system of ordinal notation is given. If Γ is just the class of Π_n formulas we write $(T)^n_\alpha$ instead of $(T)^{\Pi_n}_\alpha$.

For some purposes it is not convenient that these definitions of PRA are in a language properly extending the language of PA. One can thus also take PRA to be $\mathrm{EA} + \Sigma_1\text{-IR}$ which is formulated in the language of PA and is obtained by adding to EA the induction rule for Σ_1 formulas. Thus, the Σ_1 induction rule allows you to conclude $\forall x\, \sigma(x)$ from $\sigma(0)$ and $\forall x\, (\sigma(x) \to \sigma(x{+}1))$. The theories $\mathrm{EA}+\Sigma_n\text{-IR}$ are defined likewise and we denote them by $\mathrm{I}\Sigma^R_n$. The theory EA is just $\mathrm{I}\Delta_0 + \mathsf{exp}$.

In [Bek97] it is shown that $\mathrm{I}\Sigma^R_n$ can be axiomatized by reflection principles in the following sense, $\mathrm{I}\Sigma^R_n \equiv (\mathrm{EA})^n_\omega$. All the above definitions of PRA give rise to the same theory and these equivalences are all provable in PRA itself. In our approach we will take $(\mathrm{EA})^2_\omega$ to be the definition of PRA. It turns out that this is a very convenient formulation for us. It is also nice that this is an axiomatic formulation in the language of PA.

Moreover we will fix an enumeration of the axioms of PRA. It is known that EA is finitely axiomatizable. Since we have partial truth definitions and we are talking global reflection we have that $\{\forall x\, (\Box_{\mathrm{EA}} \pi(\dot{x}) \to \pi(x)) \mid \pi \in \Pi_2\}$ can be expressed by a single sentence

[3] Confusingly enough Smoryński later defines in [Smo85] a version of PRA which is equivalent to $\mathrm{I}\Sigma_1$.

$\mathrm{RFN}_{\Pi_2}(\mathrm{EA})$. Likewise we see that $(\mathrm{EA})^2_\alpha$ can be expressed by a single sentence for any $\alpha < \omega$. In our enumeration of PRA, the i-th axiom will be $(\mathrm{EA})^2_i$.

By taking this definition of PRA we get almost for free that every extension of PRA with a Σ_2 sentence σ is reflexive. For, reason in $\mathrm{PRA} + \sigma$ and suppose $\Box_{\mathrm{PRA}\restriction n+\sigma}\bot$. Then $\Box_{\mathrm{PRA}\restriction n}\neg\sigma$, and as $\neg\sigma$ is Π_2 we get $\neg\sigma$ by Π_2-reflection. But this contradicts σ whence $\neg\Box_{\mathrm{PRA}\restriction n+\sigma}\bot$.

1.4 Notational Conventions and Basic Notions

Many of the interesting properties of interpretability are only provable in the presence of the Σ_1-collection principle $\mathsf{B}\Sigma_1$. Our base theory PRA lacks $\mathsf{B}\Sigma_1$ and thus, for example, $A \rhd B \to (\Diamond A \to \Diamond B)$ is not provable in PRA by the standard argument. We will thus rather talk of *smooth interpretability* as introduced in [Vis91]. This notion of interpretability can be seen as the notion where the needed collection has been built in by defining it accordingly. When we speak of interpretablility we will in this paper always mean the smooth version. In presence of $\mathsf{B}\Sigma_1$ the two versions of interpretability coincide.

The interpretability logics we will use in our study often have different names depending on how they are defined. A short word on our convention on the systematic nomenclature is thus in order. A logic usually comes with three names:

- The arithmetically motivated name. For some arithmetical theory T we will denote by $\mathbf{IL}(\mathrm{T})$ its interpretability logic, which is the collection of modal sentences that are provable in T under any arithmetical realization. In other words $\mathbf{IL}(\mathrm{T}) := \{A \mid \forall * \ \mathrm{T} \vdash A^*\}$ where as usual the $*$ ranges over realizations. Likewise we write $\mathbf{PL}(T)$ for the provability logic of some theory T.

  Sometimes it will be useful to restrict the possible realizations. If Γ is some set of arithmetical sentences we write $* \in \mathsf{Sub}(\Gamma)$ to indicate that all propositional variables will be mapped to some sentence in Γ by $*$. The Γ *interpretability logic* of T is the collection of modal sentences that are provable in T under any arithmetical realization "in Γ". In other words $\mathbf{IL}_\Gamma(\mathrm{T}) := \{A \mid \forall * \in \mathsf{Sub}(\Gamma)\ \mathrm{T} \vdash A^*\}$.

  Let $\sigma_1, \ldots, \sigma_n$ be some designated arithmetical sentences. The interpretability logic that arises by adding constants $\mathtt{s_1}, \ldots, \mathtt{s_n}$ to our modal language and mapping these constants under any realization to their corresponding arithmetical sentence, that is, $(\mathtt{s_i})^* = \sigma_i$, is denoted by $\mathbf{IL}(\mathrm{T})[\sigma_1, \ldots, \sigma_n]$. The closed fragment of any of the above logics is indicated with a superscript cl. The logic **PIL** presented in Section 3 is thus the same as $\mathbf{IL}^{\mathsf{cl}}(\mathrm{PRA})[\mathrm{I}\Sigma_1]$.

- The modally motivated name. When we consider a modal interpretability logic we will always assume that it contains the basic interpretability logic **IL**. A logic is then fully determined if we just indicate which modal axioms (axiom schemes) are added to **IL**. This is indicated by just postfixing the names of the additional axioms to "**IL**". Again the closed fragment of any such logic is indicated with a superscript cl. The logic **PIL** presented in Section 3 is thus the same as $\mathbf{IL}^{\mathsf{cl}}\mathsf{S}_1\mathsf{S}_2\mathsf{S}_3\mathsf{S}_4\mathsf{W}$.
- The convenient name. Rather than using the complete and systematic name of a logic we often prefer to introduce a new and shorter name for it.

When writing modal formulas we will omit superfluous brackets. These omissions do not bring the unique readability of formulas to danger due to our binding conventions. The strongest binding connectives are negation and the modalities $\Box$ and $\Diamond$. The connectives $\vee$ and $\wedge$ bind less strong but still stronger than the $\rhd$ modality which in its turn binds stronger than $\rightarrow$. We will also omit outer brackets. Thus, $A \rhd B \rightarrow A \wedge \Box\neg C \rhd B \wedge \Box\neg C$ is short for $((A \rhd B) \rightarrow ((A \wedge \Box(\neg C)) \rhd (B \wedge \Box(\neg C))))$. Often we will use $A \rhd B \rhd C$ as short for $(A \rhd B) \wedge (B \rhd C)$ and we do the same for implication.

2 The Closed Fragment of the Provability Logic of PRA with a Constant for $\mathrm{I}\Sigma_1$.

In this section we will introduce a modal logic that generates the closed fragment of the provability logic of PRA with a constant for $\mathrm{I}\Sigma_1$. We call this logic **PGL** (the **P** in **PGL** stands for PRA) and discuss its most important properties.

In Subsection 2.1 the logic is introduced and the main theorem, the arithmetical completeness, is formulated. In 2.2 the arithmetical soundness of the logic is proved. The arithmetical completeness is dealt with in 2.3. Also the logic **PGLS** is treated there. This is the logic **PGL** plus reflection. The **S** stands for Solovay. In the last subsection, 2.4, a modal semantics is provided for **PGL** and the decidability of the logic is discussed.

2.1 The Logic PGL.

We will introduce here the logic **PGL**. Inductively we define F, the formulas of **PGL**.

$$F := \quad \bot \mid \top \mid \mathsf{S} \mid F \wedge F \mid F \vee F \mid F \rightarrow F \mid \neg F \mid \Box F.$$

The symbol S is a constant in our language just as $\bot$ is a constant. There are no propositional variables. As always we will use $\Diamond A$ as an abbreviation for $\neg\Box\neg A$. We define $\Box^0\bot := \bot$ and $\Box^{n+1}\bot := \Box(\Box^n\bot)$. We also define $\Box^\gamma\bot$ to be $\top$ for limit ordinals γ.

Throughout this paper we shall reserve $B, B_0, B_1, \ldots$ to denote boolean combinations of formulas of the form $\Box^n\bot$ with $n \in \omega + 1$. The axioms of **PGL** are all propositional tautologies in the language of F and all instantiations in F of the following schemes.

$$\begin{array}{ll} \mathsf{L}_1: & \Box(C \to D) \to (\Box C \to \Box D) \\ \mathsf{L}_2: & \Box A \to \Box\Box A \\ \mathsf{L}_3: & \Box(\Box A \to A) \to \Box A \\ \mathsf{S}_1: & \Box(\mathsf{S} \to B) \to \Box B \\ \mathsf{S}_2: & \Box(\neg\mathsf{S} \to B) \to \Box B \end{array}$$

So, by our notational convention both in S_1 and in S_2 the B is a boolean combination of formulas of the form $\Box^n\bot$ with $n \in \omega$. The rules of **PGL** are necessitation and modus ponens. Immediate consequences of S_1 and S_2 are that both $\Diamond(\mathsf{S} \wedge B)$ and $\Diamond(\neg\mathsf{S} \wedge B)$ are equivalent in **PGL** to $\Diamond B$. The logic **PGL** without S_1 and S_2 but with propositional variables is just the provability logic **GL**.

Every sentence in F can also be seen as an arithmetical statement as follows: we translate S to the canonical sentence $\mathrm{I}\Sigma_1$ (the single sentence axiomatizing the theory $\mathrm{I}\Sigma_1$), $\bot$ to, for example, 0=1 and $\top$ to 1=1. As usual we inductively extend this translation to what is sometimes called an arithmetical interpretation by taking for the translation of $\Box$ the proof predicate for PRA as fixed in Section 1.3.

If there is no chance of confusion we will use the same letter to indicate both a formal sentence of **PGL** and the arithmetical statement expressed by it. With this convention we can formulate the main theorem of this section.

Theorem 2.1 *For all sentences $A \in F$ we have*

$$\mathrm{PRA} \vdash A \Leftrightarrow \mathbf{PGL} \vdash A.$$

The implication "$\Leftarrow$" is proved in the next subsection and more specifically in Corollary 2.2 and Lemma 2.5. The other direction, the completeness, is dealt with in 2.6 where it is reduced to some more fundamental lemmas.

In [Bek96] it is shown that Theorem 2.1, in a slightly different formulation, actually holds not only for the closed fragment of PRA with a constant for $\mathrm{I}\Sigma_1$ but rather for the full language of the propositional modal logic under consideration. His schemes B_1-Cons' and B_1-NCons'

are just stronger versions of our schemes S_1 and S_2. The result presented here is thus an instance of [Bek96] and as expected the methods of proof are more elementary.

2.2 Arithmetical Soundness of PGL

This subsection is devoted to showing that everything that is proved by **PGL** is also a theorem of PRA (the translation, that is). It is known that the principles L_1 to L_3 are provable under any translation in all theories extending, say, $\mathrm{I}\Delta_0 + \Omega_1$.

So, certainly in our case we see that all instantiations of these principles are indeed provable in PRA. We also see that PRA is closed under the translation of the rules, so, the only case of interest is thus reduced to proving that all instances of S_1 and S_2 are indeed provable in PRA.

Axiom S_1 can be seen as a direct consequence of the formalization of Parsons' theorem ([Par70],[Par72]). As is pointed out for example in the first proof of [Joo02], the proof of Parsons' theorem essentially relies on Cut-elimination. The proof can thus be formalized as soon as the totality of the superexponential function is provable. Recall that $\mathrm{EA} \vdash \Box_{\mathrm{I}\Sigma_1} C \leftrightarrow \Box_{\mathrm{PRA}+\mathrm{I}\Sigma_1} C \leftrightarrow \Box_{\mathrm{PRA}}(\mathrm{I}\Sigma_1 \to C)$.

Corollary 2.2 $\mathrm{PRA} \vdash \Box_{\mathrm{PRA}}(\mathrm{I}\Sigma_1 \to B) \to \Box_{\mathrm{PRA}} B$ *for* $B \in \Pi_2$ *and thus certainly whenever* B *is as in* S_1.

When we talk about a Σ_n-extension of T we shall mean some extension of T by Σ_n-sentences and similarly for Π_n. To show that all instantiations of S_2 are also provable in PRA we reason as follows.

Fact 2.3 *(Kreisel, Levy)* $T + \mathrm{RFN}_{\Pi_n}(T)$ *is not contained in any consistent* Σ_n*-extension of* T*, and dually we have that* $T + \mathrm{RFN}_{\Sigma_n}(T)$ *is not contained in any consistent* Π_n *(and even* Σ_{n+1}*) -extension of* T.

Proof of Fact 2.3. Let S be some collection of Σ_n-sentences such that $T + S$ extends $T + \mathrm{RFN}_{\Pi_n}(T)$. We also have

$$T + S \vdash \forall x\, (\Box_T(\mathsf{Tr}_{\Pi_n}(\dot{x})) \to \mathsf{Tr}_{\Pi_n}(x)).$$

By compactness we have for some particular Σ_n-sentence σ that
$T + \sigma \vdash \forall x\, (\Box_T(\mathsf{Tr}_{\Pi_n}(\dot{x})) \to \mathsf{Tr}_{\Pi_n}(x))$. Consequently,
$T + \sigma \vdash \forall x\, (\Box_T(\mathsf{Tr}_{\Pi_n}(\ulcorner \neg\sigma \urcorner)) \to \mathsf{Tr}_{\Pi_n}(\ulcorner \neg\sigma \urcorner))$ and thus
$T \vdash \sigma \to (\Box_T(\ulcorner \neg\sigma \urcorner) \to \neg\sigma)$. But we also have
$T \vdash \neg\sigma \to (\Box_T(\ulcorner \neg\sigma \urcorner) \to \neg\sigma)$ hence
$T \vdash \Box_T(\ulcorner \neg\sigma \urcorner) \to \neg\sigma$. Löb's rule gives us $T \vdash \neg\sigma$ in which case $T + S$ is inconsistent. For the dual fact it suffices to see that $T + \mathrm{RFN}_{\Sigma_n}(T) \equiv T + \mathrm{RFN}_{\Pi_{n+1}}(T)$.

QED

This fact is used to prove the following theorem. (See Beklemishev [Bek96] for a somewhat weaker statement.)

Theorem 2.4 *$T + \neg \mathrm{I}\Sigma_1$ is Π_3-conservative over T whenever T is a Σ_3-axiomatized theory containing* EA.

PROOF OF THEOREM 2.4. It is well-known that $\mathrm{I}\Sigma_n \vdash \mathrm{RFN}_{\Pi_{n+2}}(\mathrm{EA})$. (See [Lei83] or [HP93].) Let T be some Σ_3-axiomatized theory and suppose that for some $\pi \in \Pi_3$, $T + \neg \mathrm{I}\Sigma_1 \vdash \pi$. Then $T + \neg\pi \vdash \mathrm{I}\Sigma_1$ and thus $T + \neg\pi \vdash \mathrm{RFN}_{\Pi_3}(\mathrm{EA})$. By Fact 2.3 we get that $T + \neg\pi$ can not be consistent thus $T \vdash \pi$. QED

The above reasoning can be carried out in a formalized setting. We will do so in the proof of Lemma 2.5. Lemma 2.5 implies the arithmetical correctness of axiom scheme S2.

Lemma 2.5 $\mathrm{EA} \vdash \Box_{\mathrm{PRA}}(\neg \mathrm{I}\Sigma_1 \to B) \to \Box_{\mathrm{PRA}} B$ *whenever* $B \in \Pi_3$ *and thus certainly whenever B is as in* S_2.

PROOF OF LEMMA 2.5. The formalization of the statement $\mathrm{I}\Sigma_1 \vdash \mathrm{RFN}_{\Pi_3}(\mathrm{EA})$ is a true Σ_1-sentence and thus provable in EA or even Robinson's Q. As $\mathrm{EA} \vdash \Box_{\mathrm{I}\Sigma_1}(\mathrm{RFN}_{\Pi_3}(\mathrm{EA}))$ we also have

$$\mathrm{EA} \vdash \Box_{\mathrm{EA}}(\mathrm{I}\Sigma_1 \to \mathrm{RFN}_{\Pi_3}(\mathrm{EA})). \quad (*)$$

Now reason in EA and assume $\Box_{\mathrm{PRA}}(\neg \mathrm{I}\Sigma_1 \to B)$ where B is some Π_3-sentence.[4] We get

$$\begin{array}{rll} \Box_{\mathrm{PRA}}(\neg \mathrm{I}\Sigma_1 \to B) & \to & \\ \Box_{\mathrm{PRA}}(\neg B \to \mathrm{I}\Sigma_1) & \to & \\ \exists\, \pi{\in}\Pi_2\ \Box_{\mathrm{EA}}(\neg B \wedge \pi \to \mathrm{I}\Sigma_1) & \to & \text{by } (*) \\ \exists\, \pi{\in}\Pi_2\ \Box_{\mathrm{EA}}(\neg B \wedge \pi \to \mathrm{RFN}_{\Pi_3}(\mathrm{EA})) & \to & \text{as } B \vee \neg\pi \in \Pi_3 \\ \exists\, \pi{\in}\Pi_2\ \Box_{\mathrm{EA}}(\neg B \wedge \pi \to (\Box_{\mathrm{EA}}(B \vee \neg\pi) \to B \vee \neg\pi)) & & (**) \end{array}$$

But, by simple propositional logic, we also have

$$\Box_{\mathrm{EA}}(\neg(\neg B \wedge \pi) \to (\Box_{\mathrm{EA}}(B \vee \neg\pi) \to B \vee \neg\pi))$$

which combined with $(**)$ yields $\Box_{\mathrm{EA}}(\Box_{\mathrm{EA}}(B \vee \neg\pi) \to (B \vee \neg\pi))$. By Löb's axiom we get $\Box_{\mathrm{EA}}(B \vee \neg\pi)$ which is the same as $\Box_{\mathrm{EA}}(\pi \to B)$. Thus certainly we have $\Box_{\mathrm{PRA}} B$, as π was just a part of PRA. QED

Note that Corollary 2.2 and Lemma 2.5 actually hold for a wider class of formulas than just boolean combinations of $\Box^\alpha \bot$ formulas. For example $\neg(A \rhd B)$ is always Π_3. Beklemishev also isolated a class of modal formulas which are always Π_2, see section 9.3 of [Vis97], or Subsection 4.2 of this paper. In our axioms of **PGL** we do not need the principles in full generality though.

[4] This B might be non-standard.

2.3 Arithmetical Completeness of PGL

Theorem 2.6 *For all A in F we have that if* $\mathrm{PRA} \vdash A$ *then* $\mathbf{PGL} \vdash A$.

PROOF OF THEOREM 2.6. The completeness of **PGL** actually boils down to an exercise in normal forms in modal logic. The only arithmetical ingredients are the soundness of **PGL**, the fact that $\mathrm{PRA} \vdash \Box A$ whenever $\mathrm{PRA} \vdash A$, and the fact that $\mathrm{PRA} \nvdash \Box^\alpha \bot$ for $\alpha \in \omega$.

In Lemma 2.8 we will show that $\Box A$ is always equivalent in **PGL** to $\Box^\alpha \bot$ for some $\alpha \in \omega{+}1$. Then, in Lemma 2.9 we show that if $\mathbf{PGL} \vdash \Box A$ then $\mathbf{PGL} \vdash A$. So, if $\mathbf{PGL} \nvdash A$ then $\mathbf{PGL} \nvdash \Box A$. As $\mathbf{PGL} \vdash \Box A \leftrightarrow \Box^\alpha \bot$ for some $\alpha \in \omega$ (not $\omega{+}1$ as we assumed $\mathbf{PGL} \nvdash \Box A$!) and **PGL** is sound we also have $\mathrm{PRA} \vdash \Box A \leftrightarrow \Box^\alpha \bot$. Hence $\mathrm{PRA} \nvdash \Box A$ and also $\mathrm{PRA} \nvdash A$. QED

We work out the exercise in modal normal forms. Although this is already carried out in the literature (see e.g. Boolos [Boo93], or Visser [Vis92]) we repeat it here to obtain some subsidiary information which we shall need later on.

Recall that we will in this subsection reserve the letters $B, B_0, B_1, \ldots$ for boolean combinations of $\Box^\alpha \bot$-formulas. Thus, a sentence B can be written in conjunctive normal form, that is, $\bigwedge_i (\bigvee_j \neg \Box^{a_{ij}} \bot \vee \bigvee_k \Box^{b_{ik}} \bot)$. Each conjunct $\bigvee_j \neg \Box^{a_{ij}} \bot \vee \bigvee_k \Box^{b_{ik}} \bot$ can be written as $\Box^{\alpha_i} \bot \to \Box^{\beta_i} \bot$ where $\alpha_i := \min(\{a_{ij}\})$ and $\beta_i := \max(\{b_{ik}\})$.

By convention the empty conjunction is just $\top$ and the empty disjunction is just $\bot$. In order to have this convention in concordance with our normal forms we define $\min(\varnothing){=}0$ and $\max(\varnothing){=}\omega$. In $\bigwedge_i (\Box^{\alpha_i} \bot \to \Box^{\beta_i} \bot)$ we can leave out the conjuncts whenever $\alpha_i \leq \beta_i$, for, in that case, $\mathbf{PGL} \vdash \Box^{\alpha_i} \bot \to \Box^{\beta_i} \bot$.

So, if we say that some formula B is in conjunctive normal form we will in the sequel assume that B is written as $\bigwedge_i (\Box^{\alpha_i} \bot \to \Box^{\beta_i} \bot)$ with $\alpha_i > \beta_i$. The empty conjunction gives $\top$ and if we take $\alpha_0{=}\omega > 0{=}\beta_0$, we get with one conjunct just $\bot$. Further restrictions on the α_i and β_i yield strong normal forms. They will be introduced in Definition 4.11.

Lemma 2.7 *If a formula B can be written in the form $\bigwedge_i (\Box^{\alpha_i} \bot \to \Box^{\beta_i} \bot)$ with $\alpha_i{>}\beta_i$, then we have that* $\mathbf{PGL} \vdash \Box B \leftrightarrow \Box^{\beta+1} \bot$ *where* $\beta{=}\min(\{\beta_i\})$.

PROOF OF LEMMA 2.7. The proof is actually carried out in **GL**. We have in **PGL** that $\Box(\bigwedge_i (\Box^{\alpha_i} \bot \to \Box^{\beta_i} \bot)) \leftrightarrow \bigwedge_i \Box(\Box^{\alpha_i} \bot \to \Box^{\beta_i} \bot)$. We will see that $\Box(\Box^{\alpha_i} \bot \to \Box^{\beta_i} \bot)$ is equivalent in **PGL** to $\Box^{\beta_i+1} \bot$.

So, reason in **PGL** and assume $\Box B$. As $\beta_i < \alpha_i$ we know that $\beta_i + 1 \leq \alpha_i$ and thus $\Box^{\beta_i+1} \bot \to \Box^{\alpha_i} \bot$. Now $\Box(\Box^{\alpha_i} \bot \to \Box^{\beta_i} \bot) \to \Box(\Box^{\beta_i+1} \bot \to \Box^{\beta_i} \bot)$. One application of L_3 yields $\Box(\Box^{\beta_i} \bot)$ i.e. $\Box^{\beta_i+1} \bot$.

On the other hand we easily see that $\Box(\Box^{\beta_i}\bot) \rightarrow \Box(\Box^{\alpha_i}\bot \rightarrow \Box^{\beta_i}\bot)$ hence we have shown the equivalence. Finally we remark that $(\bigwedge\!\!\!\bigwedge_i \Box^{\beta_i+1}\bot) \leftrightarrow \Box^{\beta+1}\bot$ where $\beta = \min(\{\beta_i\})$. QED

Lemma 2.8 *For any formula A in F we have that A is equivalent in* **PGL** *to a boolean combination of formulas of the form* S *or $\Box^\beta\bot$, and that $\Box A$ is equivalent in* **PGL** *to $\Box^\alpha\bot$ for some $\alpha \in \omega + 1$.*

Proof of Lemma 2.8. By induction on the complexity of formulas in F. The base cases are trivial. The only interesting case in the induction is where we consider the case that $A = \Box C$. Note that C, by induction being a boolean combination of $\Box^\alpha\bot$ formulas and S, can be written as $(\mathtt{S} \rightarrow B_0) \wedge (\neg\mathtt{S} \rightarrow B_1)$. So, by Lemma 2.7 we have that for suitable indices β, β', β'':

$$\begin{array}{cc} \Box C & \leftrightarrow \\ \Box((\mathtt{S} \rightarrow B_0) \wedge (\neg\mathtt{S} \rightarrow B_1)) & \leftrightarrow \\ \Box(\mathtt{S} \rightarrow B_0) \wedge \Box(\neg\mathtt{S} \rightarrow B_1) & \leftrightarrow \\ \Box B_0 \wedge \Box B_1 & \leftrightarrow \\ \Box^{\beta'+1}\bot \wedge \Box^{\beta''+1}\bot & \leftrightarrow \\ \Box^\beta\bot. & \end{array}$$

QED

Lemma 2.9 *If* **PGL** $\vdash \Box A$ *then* **PGL** $\vdash A$.

Proof of Lemma 2.9. By Lemma 2.8 we can write A as a boolean combination of formulas of the form S or $\Box^\beta\bot$. Thus let $A \leftrightarrow (\mathtt{S} \rightarrow B_0) \wedge (\neg\mathtt{S} \rightarrow B_1)$ with B_0 and B_1 in conjunctive normal form and assume $\vdash \Box A$. For appropriate indices we have $B_0 = \bigwedge\!\!\!\bigwedge_i(\Box^{\alpha_i}\bot \rightarrow \Box^{\beta_i}\bot)$ and $B_1 = \bigwedge\!\!\!\bigwedge_j(\Box^{\alpha'_j}\bot \rightarrow \Box^{\beta'_j}\bot)$. Using S_2, S_3 and Lemma 2.7 we get that $\Box A \leftrightarrow \Box^{\beta+1}\bot$ with $\beta = \min(\{\beta_i, \beta'_j\})$. By assumption $\beta = \omega$, thus all the β_i and β'_j were ω and hence $\vdash A$. QED

We can also define a logic, say **PGLS**, that captures all the *true* sentences of F. The logic **PGLS** is defined as follows. The axioms of **PGLS** are all the theorems of **PGL** together with S and $\{\Diamond^\alpha\top \mid \alpha \in \omega\}$. It's sole rule of inference is modus ponens.

Theorem 2.10 **PGLS** $\vdash A \Leftrightarrow \mathbb{N} \models A$

Proof of Theorem 2.10. By induction on the length of **PGLS** $\vdash A$ we see that **PGLS** $\vdash A \Rightarrow \mathbb{N} \models A$.

To see the converse, we reason as follows. Consider $A \in F$ such that $\mathbb{N} \models A$. By Lemma 2.8 we can find an A' which is a boolean combination of S and $\Diamond^\alpha\top$ ($\alpha \in \omega + 1$), such that **PGL** $\vdash A \leftrightarrow A'$. Thus PRA $\vdash A \leftrightarrow A'$ and also $\mathbb{N} \models A \leftrightarrow A'$. Consequently $\mathbb{N} \models A'$.

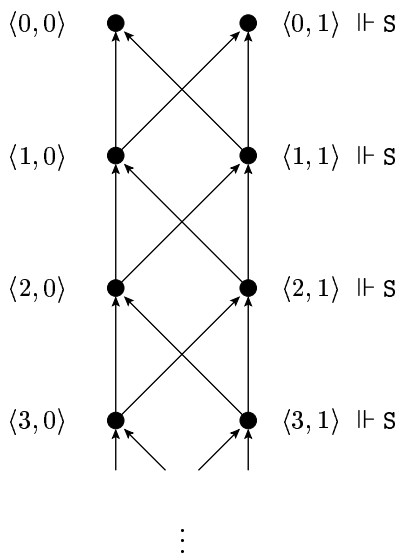

Figure 1: Modal semantics

Moreover, as A' is a boolean combination of $\mathtt{S}$ and $\Diamond^\alpha \top$ ($\alpha \in \omega+1$), for some $m \in \omega$, $\mathtt{S} \wedge \bigwedge_{i=1}^{m} \Diamond^i \top \rightarrow A'$ is a propositional logical tautology whence A' is provable in **PGLS**. Also **PGLS** $\vdash A \leftrightarrow A'$ whence **PGLS** $\vdash A$. QED

Clearly the theorems of **PGLS** are recursively enumerable. As **PGLS** is a complete logic in the sense that it either refutes a formula or proves it, we see that theoremhood of **PGLS** is actually decidable.

2.4 Modal Semantics for PGL, Decidability

In this subsection we will provide a modal semantics for **PGL**. Actually we will give a model $\mathcal{M}$ as depicted in Figure 1 which in some sense displays all there is to know about closed sentences with a constant for $\mathrm{I}\Sigma_1$ in **PGL**.

Definition 2.11 *We define the model $\mathcal{M}$ as follows, $\mathcal{M} := \langle M, R, \Vdash \rangle$. Here $M := \{\langle n, i\rangle \mid n \in \omega,\ i \in \{0,1\}\}$ and $\langle n,i\rangle R\langle m,j\rangle \Leftrightarrow m < n$. Furthermore $\langle n, i\rangle \Vdash \mathtt{S} \Leftrightarrow i = 1$.*

Theorem 2.12 $\forall m\ \mathcal{M}, m \Vdash A \Leftrightarrow$ **PGL** $\vdash A$

Proof of Theorem 2.12.

$\Leftarrow$ This direction is obtained by induction on the complexity of proofs in **PGL**. As $\mathcal{M}$ is a reflexive and upwards well-founded model, it is indeed a model of all instantiations of the axioms L_1, L_2 and L_3. Thus, consider S_1.

So, suppose at some world $\mathbf{m}$ $(= \langle m, i\rangle)$, we have that $\langle m, i\rangle \Vdash \Box(\mathtt{S} \to B)$. Then $\langle n, 1\rangle \Vdash B$ for $n < m$. Recall that B does not contain $\mathtt{S}$. It is well-known that the forcing of B depends solely on the depth of the world, so, we also have $\langle n, 0\rangle \Vdash B$. Thus $\mathbf{m}R\mathbf{n}$ yields $\mathbf{n} \Vdash B$. Consequently $\mathbf{m} \Vdash \Box B$, which gives us the validity of S_1.

The S_2-case is treated completely similarly. It is also clear that this direction of the theorem remains valid under applications of both modus ponens and the necessitation rule.

$\Rightarrow$ Suppose $\mathbf{PGL} \nvdash A$. By Lemma 2.9 $\mathbf{PGL} \nvdash \Box A$, thus $\mathbf{PGL} \vdash \Box A \leftrightarrow \Box^\alpha \bot$ for a certain $\alpha \in \omega$. By the first part of this proof we may conclude that $\mathbf{m} \Vdash \Box A \leftrightarrow \Box^\alpha \bot$ for any $\mathbf{m}$. As $\langle \alpha, i\rangle \nVdash \Box^\alpha \bot$, we automatically get $\langle \alpha, i\rangle \nVdash \Box A$. So, for some $\langle \beta, j\rangle$ with $\langle \alpha, i\rangle R \langle \beta, j\rangle$ we have $\langle \beta, j\rangle \Vdash \neg A$ showing the "non-validity" of A.

QED

The set of theorems of **PGL** is clearly recursively enumerable. If a formula is not provable in **PGL**, then, by Theorem 2.12, in some node of the model $\mathcal{M}$ it is refuted. Thus the theoremhood of **PGL** is actually decidable.

3 The Closed Fragment of the Interpretability Logic of PRA with a Constant for $\mathrm{I}\Sigma_1$.

In this section we will introduce a modal logic that generates the closed fragment of the interpretability logic of PRA with a constant for $\mathrm{I}\Sigma_1$. We call this logic **PIL** and discuss its most important properties.

In Subsection 3.1 the logic is introduced and the main theorem, comprising the arithmetical soundness and completeness, is formulated. In 3.2 the arithmetical soundness of the logic is proved. Subsection 3.3 provides a proof of the fact that $\mathrm{I}\Sigma_1$ proves the consistency of PRA on a definable cut. This is an interesting result on its own but it also provides an alternative proof of the arithmetical soundness of **PIL**. The arithmetical completeness is dealt with in 3.11.

Also the logic **PILS** is treated in this section. This is the logic **PIL** plus reflection. It generates all true (on the standard model) interpretability principles of PRA formulated in the closed fragment together with a constant for $\mathrm{I}\Sigma_1$. The **S** in **PILS** stands for Solovay.

In 3.5, the last subsection, a modal semantics is provided for **PIL** which gives us the decidability of the logic.

3.1 The Logic PIL.

In this subsection we present the logic **PIL**. The set of formulas, I, of this logic is defined inductively as

$$I := \quad \bot \mid \top \mid \mathsf{S} \mid I \wedge I \mid I \vee I \mid I \to I \mid \neg I \mid \Box I \mid I \rhd I.$$

The constants of the language are $\bot, \top$ and S. Again we will reserve in this section the symbols $B, B_0, B_1, \ldots$ to denote boolean combinations of $\Box^\alpha \bot$ ($\alpha \in \omega{+}1$) formulas. We will write $C \equiv D$ as short for $(C \rhd D) \wedge (D \rhd C)$ and we say that they are equi-interpretable. The axioms of **PIL** are all tautologies over I and all instances of the following axiom schemes.[5]

$$\begin{array}{ll}
\mathsf{L}_1: & \Box(C \to D) \to (\Box C \to \Box D) \\
\mathsf{L}_2: & \Box A \to \Box\Box A \\
\mathsf{L}_3: & \Box(\Box A \to A) \to \Box A \\
\mathsf{S}_1: & \Box(\mathsf{S} \to B) \to \Box B \\
\mathsf{S}_2: & \Box(\neg\mathsf{S} \to B) \to \Box B \\
\mathsf{J1}: & \Box(C \to D) \to C \rhd D \\
\mathsf{J2}: & (C \rhd D) \wedge (D \rhd E) \to C \rhd E \\
\mathsf{J3}: & (C \rhd E) \wedge (D \rhd E) \to C \vee D \rhd E \\
\mathsf{J4}: & C \rhd D \to (\Diamond C \to \Diamond D) \\
\mathsf{J5}: & \Diamond A \rhd A \\
\mathsf{W}: & C \rhd D \to C \rhd D \wedge \Box\neg C \\
\mathsf{S}_3: & \neg\mathsf{S} \wedge B \equiv B \\
\mathsf{S}_4: & (B \rhd \mathsf{S} \wedge B) \to \Box\neg B
\end{array}$$

We recall that the B in S_1-S_4 are boolean combinations of $\Box^\alpha \bot$ formulas. The rules of **PIL** are modus ponens and necessitation.

Again we can see any sentence in I as an arithmetical statement translating $\rhd$ as the intended arithmetization of interpretability over PRA and $\Box$ as an arithmetization of provability in PRA and propagating this inductively along the structure of the formulas as usual. With this convention we can formulate the main result of this section.

Theorem 3.1 *For all sentences $A \in I$ we have* $\mathrm{PRA} \vdash A \Leftrightarrow$ **PIL** $\vdash A$.

[5] **PIL** contains some redundancy. Our aim however is not to find an axiomatization without redundancy. L_2 for example is doubly redundant as it can be deduced from L_3 and L_1 but also from J5 and J4. Also S_2 follows from S_3. If $B \rhd \neg\mathsf{S} \wedge B$ by J4, $\Diamond\neg B \to \Diamond(\neg\mathsf{S} \wedge \neg B)$ from which S_2 follows. The axiom scheme W could actually be replaced by the somewhat weaker scheme $\mathsf{F}: \; A \rhd \Diamond A \to \Box\neg A$.

The "$\Leftarrow$" direction, that is, the arithmetical soundness, is proved in Subsection 3.2. In Subsection 3.3 the arithmetical soundness is proved by completely different means. The "$\Rightarrow$" direction, the arithmetical completeness, is proved in Subsection 3.4.

As the interpretability logic **ILW** is a part of **PIL** we have access to all known reasoning in **IL** and **ILW**. In this section, unless mentioned otherwise $\vdash$ refers to provability in **PIL**.

Fact 3.2

(1.) $\vdash \Box A \leftrightarrow \neg A \rhd \bot$

(2.) $\vdash \Box^{\alpha+1}\bot \rightarrow \Diamond^{\beta}\top \rhd A \quad$ *if* $\alpha \leq \beta$

(3.) $\vdash A \equiv A \vee \Diamond A$

(4.) $\vdash A \rhd \Diamond A \rightarrow \Box\neg A$

(1.) tells us that in our language the $\Box$ could actually be dispensed with or could be seen as an abbreviation. We choose not to do so for the sake of readability.

As an example we prove (2.). We reason in **PIL** and use our notational conventions. It is sufficient to prove the case when $\alpha = \beta$. Thus, $\Box^{\alpha+1}\bot \rightarrow \Box(\Box^{\alpha}\bot) \rightarrow \Box(\neg A \rightarrow \Box^{\alpha}\bot) \rightarrow \Box(\Diamond^{\alpha}\top \rightarrow A) \rightarrow \Diamond^{\alpha}\top \rhd A$.

Fact (4.) is Feferman's principle and can be seen as a "coordinate free" version of Gödel's second incompleteness theorem. It follows immediately from W realizing that $A \rhd \bot$ is by (1.) nothing but $\Box\neg A$.

3.2 Arithmetical Soundness of PIL

This subsection is mainly dedicated to prove the following lemma.

Lemma 3.3 *For all A in I we have that if* **PIL** $\vdash A$ *then* PRA $\vdash A$.

In [Vis91] it has been shown that **ILW** is sound for any reasonably formulated theory extending $\mathrm{I}\Delta_0 + \Omega_1$. So, to check for soundness of **PIL** with respect to PRA we only need to see that all translations of S_3 and S_4 are provable in PRA. As always we have that PRA is closed under the rules of modus ponens and of necessitation.

In the proof of the following lemma we will use a characterization for interpretability due to Orey and Hájek for reflexive theories. The characterization says that in some basic theory, say EA, we have $T \rhd S \leftrightarrow \forall n\ \Box_T \mathsf{Con}(S\!\upharpoonright\! n)$ whenever T is a reflexive theory. (Orey [Ore61], Hájek [Háj71], [Háj72].) As we work in PRA, whence in the absence of $\mathsf{B}\Sigma_1$, we use smooth interpretation. In [Vis91] it is shown that $\mathrm{I}\Delta_0 + \Omega_1 \vdash \forall n\ \Box_T(\mathsf{Con}(S\!\upharpoonright\! n)) \rightarrow T \rhd_s S$, where the $\rhd_s$ denotes smooth interpretability (see Subsection 1.4), and this is what we use in the reasoning below. Lemma 3.4 provides the arithmetical soundness of axiom scheme S_3.

Lemma 3.4 $\mathrm{PRA} \vdash B \rhd_{\mathrm{PRA}} B \wedge \neg \mathrm{I}\Sigma_1$ *for* $B \in \Sigma_2$, *so, certainly for* B *as in* S_3.

Proof of Lemma 3.4. We want to show that $\mathrm{PRA} + B \rhd \mathrm{PRA} + B + \neg \mathrm{I}\Sigma_1$. At the end of Section 1.3 we have shown that every finite Σ_2-extension of PRA is reflexive. Thus, in the light of the Orey-Hájek characterization, we are ready if we can prove

$$\mathrm{PRA} \vdash \forall n \, \Box_{\mathrm{PRA}+B}(\Diamond_{\mathrm{PRA}\restriction n+B+\neg \mathrm{I}\Sigma_1} \top). \quad (\alpha)$$

We will set out to prove that

(i) $\quad \mathrm{EA} \vdash \forall n \, \Box_{\mathrm{PRA}+B}(\Box_{\mathrm{PRA}\restriction n+B+\neg \mathrm{I}\Sigma_1} \bot \to \Box_{\mathrm{PRA}\restriction n+B} \bot)$,

(ii) $\quad \mathrm{EA} \vdash \forall n \, \Box_{\mathrm{PRA}+B}(\Box_{\mathrm{PRA}\restriction n+B} \bot \to \bot)$,

from which (α) immediately follows.

The proof of (i) is just a slight modification of the proof of Lemma 2.5. We reason in EA and fix some n:

$$\begin{array}{lll}
\Box_{\mathrm{PRA}+B} & (& \Box_{\mathrm{PRA}\restriction n+B+\neg \mathrm{I}\Sigma_1} \bot \\
& \to & \Box_{\mathrm{PRA}\restriction n+B} \mathrm{I}\Sigma_1 \\
& \to & \Box_{\mathrm{PRA}\restriction n+B} \mathrm{RFN}_{\Pi_3}(\mathrm{EA}) \\
& \to & \Box_{\mathrm{EA}}(\mathrm{PRA}\restriction n \wedge B \to \mathrm{RFN}_{\Pi_3}(\mathrm{EA})) \\
& \to & \Box_{\mathrm{EA}}(\mathrm{PRA}\restriction n \wedge B \to (\Box_{\mathrm{EA}} \neg(\mathrm{PRA}\restriction n \wedge B) \to \neg(\mathrm{PRA}\restriction n \wedge B))) \\
& \to & \Box_{\mathrm{EA}}(\Box_{\mathrm{EA}} \neg(\mathrm{PRA}\restriction n \wedge B) \to \neg(\mathrm{PRA}\restriction n \wedge B)) \\
& \to & \Box_{\mathrm{EA}} \neg(\mathrm{PRA}\restriction n \wedge B) \\
& \to & \Box_{\mathrm{EA}}(\mathrm{PRA}\restriction n \to \neg B) \\
& \to & \Box_{\mathrm{PRA}\restriction n} \neg B \\
& \to & \Box_{\mathrm{PRA}\restriction n+B} \bot \quad).
\end{array}$$

The proof of (ii) is just a formalization of the fact that every finite Σ_2-extension of PRA is reflexive. Recall that we fixed our axiomatization of PRA in Section 1.3 such that $\mathrm{PRA}\restriction n = (\mathrm{EA})^2_n$. Thus, by definition, $\mathrm{PRA}\restriction(n+1) \vdash \Box_{\mathrm{PRA}\restriction n} \pi \to \pi$ for $\pi \in \Pi_2$.

If we fix some $\neg B \in \Pi_2$, $\mathrm{PRA}\restriction(n+1) \vdash \Box_{\mathrm{PRA}\restriction n} \neg B \to \neg B$ becomes a true Σ_1-sentence, and thus is verifiable in EA:

$$\mathrm{EA} \vdash \Box_{\mathrm{PRA}\restriction(n+1)}(\Box_{\mathrm{PRA}\restriction n} \neg B \to \neg B).$$

Obviously we also have $\mathrm{EA} \vdash \Box_{\mathrm{PRA}\restriction(n+1)+B} B$. Combining, this yields a proof of (ii) in EA:

$$\begin{array}{lll}
\Box_{\mathrm{PRA}\restriction(n+1)+B} & (& \Box_{\mathrm{PRA}\restriction n+B} \bot \\
& \to & \Box_{\mathrm{PRA}\restriction n} \neg B \\
& \to & \neg B \\
& \to & \bot \quad).
\end{array}$$

QED

As we will make ample use of the principle W in a more general form, we find it convenient to explicitly state and prove it here. Below, we will assume that U and V contain $I\Delta_0 + \Omega_1$.

Theorem 3.5 $\mathrm{EA} \vdash U \rhd V \to U \rhd V + \Box_U \bot$

Proof of Theorem 3.5. We reason in EA and assume $j : U \rhd V$. It is well-known (cf. Pudlák [Pud85], Lemma 3.3) that we can find, corresponding to our interpretation j, a U-cut I and a mapping f that is an isomorphism between I and its image $f[I]$, where $f[I]$ is an initial segment of the V-numbers. We fix the interpretation j, the cut I and the isomorphism f and set out to prove

$$U \rhd U + \Box_U^I \bot. \quad (\&)$$

Obviously $U \rhd (U + (\Box_U^I \bot \vee \Diamond_U^I \top))$, so, we have proved (&) if we have shown $U + \Diamond_U^I \top \rhd U + \Box_U^I \bot$.

By Löb's theorem relativized to I we get $\Diamond_U^I \top \to \Diamond_U^I (\top \wedge \Box_U^I \bot)$. Consequently $U + \mathsf{Con}_U^I \top \rhd U + \mathsf{Con}_U^I (\top \wedge \Box_U^I \bot)$. By the formalized model-existence lemma we get $I\Delta_0 + \Omega_1 + \mathsf{Con}_U^I (\top \wedge \Box_U^I \bot) \rhd U + \Box_U^I \bot$, and thus we have showed (&). By our interpretation j we now get $j : U + \Box_U^I \bot \rhd V$.

But, as j comes with the isomorphism f on I and $\Box_U^I \bot$ is a Σ_1-formula, we actually get $j : U + \Box_U^I \bot \rhd V + \Box_U^I \bot$. (Remember that $f[I]$ is an initial part of the V-numbers.) We resume and see that we now have come to a proof of our theorem:

$$U \rhd U + \Box_U^I \bot \rhd V + \Box_U^I \bot \rhd V + \Box_U \bot.$$

QED

In the following lemma we need to use the fact that Feferman's principle F, $\vdash A \rhd \Diamond A \to \Box \neg A$, is provable for any reasonably formulated arithmetical theory U of some minimal strength, in the sense that

$$\mathrm{EA} \vdash U \rhd (U + \mathsf{Con}(U)) \to \Box_U \bot.$$

This however, is a direct consequence of Theorem 3.5.[6]

[6]Here is a nice direct proof of Feferman's principle in case U is a finite sequential theory. It was told to me by Volodya Shavrukov. Consider in U the fixed point $\varphi \leftrightarrow \neg(\top \rhd_U \varphi)$. We show in U that $\varphi \leftrightarrow \Diamond_U \top$. For the "$\to$" direction we assume φ and $\Box_U \bot$. But $\Box_U \bot \to \top \rhd_U \varphi$ (see Fact 3.2.(2.)) whence $\neg\varphi$. Together with our assumption φ we get $\bot$, whence $\Diamond_U \top$. To show the $\leftarrow$ direction we assume $\Diamond_U \top$ and $\neg\varphi$. The latter is nothing but $\top \rhd_U \varphi$. As U is finite, this is a Σ_1-sentence, whence $\Box(\top \rhd_U \varphi)$, that is, $\Box(\neg\varphi)$. From $\top \rhd_U \varphi$ we get $\Diamond_U \top \to \Diamond_U \varphi$ and under our assumption $\Diamond_U \top$ we get $\Diamond_U \varphi$. This yields a contradiction with $\Box_U \neg\varphi$ and thus we have $\Diamond_U \top \to \varphi$. Substituting φ back in the fixed point equation yields the required $\top \rhd_U \Diamond_U \top \to \Box_U \bot$. Note that in the proof of Lemma 3.6 we need Feferman's principle for $U = \mathrm{PRA}\restriction k + B$ which is finite (although possibly non-standard).

Lemma 3.6 $\mathrm{PRA} \vdash B \rhd_{\mathrm{PRA}} B \wedge \mathrm{I}\Sigma_1 \rightarrow \Box_{\mathrm{PRA}}\neg B$ *for* $B{\in}\Sigma_2$*, so, certainly for* B *as in* S_4

Proof of Lemma 3.6. The theory $\mathrm{PRA} + B + \mathrm{I}\Sigma_1$ is just $\mathrm{I}\Sigma_1 + B$ and hence finitely axiomatizable and this is verifiable in PRA. Now we will reason in PRA.

We suppose that $\mathrm{PRA} + B \rhd \mathrm{PRA} + B + \mathrm{I}\Sigma_1$. As $\mathrm{PRA} + B + \mathrm{I}\Sigma_1$ is finitely axiomatizable we have that $\mathrm{PRA}\restriction k + B \rhd \mathrm{PRA} + B + \mathrm{I}\Sigma_1$ for some natural number k. $\mathrm{PRA} + B$ is reflexive as it is a finite Σ_2-extension of PRA and thus $\mathrm{PRA} + B \vdash \mathsf{Con}(\mathrm{PRA}\restriction k + B)$.

So, certainly $\mathrm{PRA} + B + \mathrm{I}\Sigma_1 \vdash \mathsf{Con}(\mathrm{PRA}\restriction k + B)$ and thus also $\mathrm{PRA} + B + \mathrm{I}\Sigma_1 \rhd \mathrm{PRA}\restriction k + B + \mathsf{Con}(\mathrm{PRA}\restriction k + B)$. By transitivity we get $\mathrm{PRA}\restriction k + B \rhd \mathrm{PRA}\restriction k + B + \mathsf{Con}(\mathrm{PRA}\restriction k + B)$. By Feferman's principle we get that $\Box_{\mathrm{PRA}\restriction k+B}\bot$ and thus $\Box_{\mathrm{PRA}+B}\bot$ and also $\Box_{\mathrm{PRA}}(B \rightarrow \bot)$, i.e., $\Box_{\mathrm{PRA}}\neg B$. QED

Lemma 3.6 certainly proves the correctness of axiom scheme S_4. The proof also yields the following insights.

Corollary 3.7 *A* Σ_1*-sound reflexive theory* U *does not interpret any finitely axiomatized theory extending it. In particular* PRA *does not interpret* $\mathrm{I}\Sigma_1$ *nor any other finitely axiomatized theory extending it.*

Corollary 3.8 $\mathrm{PRA} + \neg\mathrm{I}\Sigma_1$ *is not finitely axiomatizable.*

Corollary 3.9 *(*$\mathrm{EA} \vdash$*) No consistent* Σ_2*-extension of* PRA *is finitely axiomatizable.*

Proof of Corollary 3.9. The proof is a slight modification of the proof of Lemma 3.6. Let $\sigma{\in}\Sigma_2$. If $\mathrm{PRA} + \sigma$ is finitely axiomatizable then $\mathrm{PRA}\restriction k + \sigma \rhd \mathrm{PRA} + \sigma$ for some natural number k. Thus $\mathrm{PRA}\restriction k + \sigma \rhd \mathrm{PRA}\restriction k + \sigma + \mathsf{Con}(\mathrm{PRA}\restriction k + \sigma)$, whence $\Box_{\mathrm{PRA}}\neg\sigma$.[7] QED

3.3 $\mathrm{I}\Sigma_1$ Proves the Consistency of PRA on a Cut

The main result of this subsection is formulated in Theorem 3.10. As an immediate consequence of this theorem we see that we can find an $\mathrm{I}\Sigma_1$-cut J such that for every Σ_2-sentence B, $\mathrm{I}\Sigma_1$ proves the consistency of $\mathrm{PRA} + B$ on this cut J. We will denote $\forall x\ (J(x) \rightarrow \neg\mathsf{Prf}_{\mathrm{PRA}+B}(x, \bot))$ by $\mathsf{Con}^J(\mathrm{PRA} + B)$. As we will see, this gives us alternative proofs of Lemmas 3.4 and 3.6. If the reader is mainly interested in the logic **PIL**, this subsection can be skipped.

The phenomenon of two types of proofs of the soundness of interpretability principles shows up time and again. We call the proof-style

[7] The corollary also follows directly from the fact that $\mathrm{PRA} + \sigma$ is reflexive and Gödel's second incompleteness theorem.

that makes use of finite approximations of a theory P-style proofs as to refer to the principle P that holds for any finitely axiomatized theory. The other style of proofs considers the theory as one entity but makes use of definable cuts. We call those sort of proofs M-style proofs as to refer to the principle M that holds in any essentially reflexive theory. In this subsection we thus present the M-style proofs of the soundness of S_3 and S_4.

Theorem 3.10 *For each $n\in\omega$ there exists some $\mathrm{I}\Sigma_n$-cut J_n such that for all Σ_{n+1}-sentences σ, $\mathrm{I}\Sigma_n + \sigma \vdash \mathsf{Con}^{J_n}(\mathrm{I}\Sigma_n^R + \sigma)$.*

Before we will begin a proof of the theorem we first show how the theorem provides alternative proofs of the arithmetical soundness of S_3 and S_4.

Proof of Lemma 3.4. We have $B\in\Sigma_2$ and we want to show in EA that $\mathrm{PRA} + B \rhd \mathrm{PRA} + B + \neg\mathrm{I}\Sigma_1$. Clearly

$$\mathrm{PRA} + B \rhd (\mathrm{PRA} + B + (\mathrm{I}\Sigma_1 \vee \neg\mathrm{I}\Sigma_1)).$$

So, we are done if we can show that $\mathrm{PRA}+B+\mathrm{I}\Sigma_1 \rhd \mathrm{PRA}+B+\neg\mathrm{I}\Sigma_1$. By Theorem 3.10 we get that $\mathrm{PRA}+B+\mathrm{I}\Sigma_1 = \mathrm{I}\Sigma_1+B \vdash \mathsf{Con}^J(\mathrm{PRA}+B)$. By Solovay's shortening techniques we can find a cut $J' \subseteq J$ with J' closed under the ω_2 function.

Using this cut J' to relativize the identity translation, we find an interpretation that witnesses $\mathrm{I}\Sigma_1+B\rhd\mathrm{I}\Delta_0+\Omega_1+\Diamond_{\mathrm{PRA}}B$. By Theorem 3.5 we see that

$$\begin{array}{ll} \mathrm{I}\Sigma_1 + B & \rhd \\ \mathrm{I}\Delta_0 + \Omega_1 + \Diamond_{\mathrm{PRA}}B + \Box_{\mathrm{I}\Sigma_1+B}\bot & \rhd \\ \mathrm{I}\Delta_0 + \Omega_1 + \Diamond_{\mathrm{PRA}}B + \Box_{\mathrm{PRA}}(B \to \neg\mathrm{I}\Sigma_1). & \end{array}$$

Thus $\mathrm{I}\Sigma_1+B\rhd\mathrm{I}\Delta_0+\Omega_1+\Diamond_{\mathrm{PRA}}(B\wedge\neg\mathrm{I}\Sigma_1)$. By the formalized Henkin construction we get $\mathrm{I}\Sigma_1 + B \rhd \mathrm{PRA} + B + \neg\mathrm{I}\Sigma_1$. QED

Proof of Lemma 3.6. We have $B\in\Sigma_2$ and assume in EA that $\mathrm{PRA}+B \rhd \mathrm{PRA} + B + \mathrm{I}\Sigma_1$. We have already seen in the above proof that $\mathrm{PRA} + B + \mathrm{I}\Sigma_1 \rhd \mathrm{I}\Delta_0 + \Omega_1 + \Diamond_{\mathrm{PRA}}B$.

Thus, by transitivity $\mathrm{PRA} + B \rhd \mathrm{I}\Delta_0 + \Omega_1 + \Diamond_{\mathrm{PRA}}B$. Theorem 3.5 now yields

$$\begin{array}{ll} \mathrm{PRA} + B & \rhd \\ \mathrm{I}\Delta_0 + \Omega_1 + \Diamond_{\mathrm{PRA}}B + \Box_{\mathrm{PRA}+B}\bot & \rhd \\ \bot & \end{array}$$

which is the same as $\Box_{\mathrm{PRA}+B}\bot$, i.e., $\Box_{\mathrm{PRA}}\neg B$. QED

Proof of Theorem 3.10. From [Bek97] it is known that $\mathrm{I}\Sigma_n^R \equiv (\mathrm{EA})_\omega^{n+1}$. Let ϵ be the arithmetical sentence axiomatizing EA. We fix the following axiomatization $\{i_m^n\}_{m\in\omega}$ of $\mathrm{I}\Sigma_n^R$:

$i^n_0 := \epsilon$,

$i^n_{m+1} := i^n_m \wedge \forall x\, (\Box_{i^n_m} \mathsf{True}_{\Pi_{n+1}}(\dot{x}) \to \mathsf{True}_{\Pi_{n+1}}(x))$.

The map that sends m to the code of i^n_m is clearly primitive recursive. We will assume that the context makes clear if we are talking about the formula or its code when writing i^n_m. Similarly for other formulas. An $\mathrm{I}\Sigma_n$-cut J_n is defined as follows:

$$J_n(x) := \forall\, y{\leq}x\ \mathsf{True}_{\Pi_{n+1}}(i^n_y).$$

We will now see that J_n indeed is an $\mathrm{I}\Sigma_n$-cut. Clearly $\mathrm{I}\Sigma_n \vdash J_n(0)$. It remains to show that $\mathrm{I}\Sigma_n \vdash J_n(m) \to J_n(m{+}1)$.

So, we reason in $\mathrm{I}\Sigma_n$ and assume $J_n(m)$. We need to show that $\mathsf{True}_{\Pi_{n+1}}(i^n_{m+1})$, that is,

$$\mathsf{True}_{\Pi_{n+1}}(i^n_m \wedge \forall x\, (\Box_{i^n_m} \mathsf{True}_{\Pi_{n+1}}(\dot{x}) \to \mathsf{True}_{\Pi_{n+1}}(x))).$$

The induction hypothesis gives us $\mathsf{True}_{\Pi_{n+1}}(i^n_m)$ thus we need to show

$$\mathsf{True}_{\Pi_{n+1}}(\forall x\, (\Box_{i^n_m} \mathsf{True}_{\Pi_{n+1}}(\dot{x}) \to \mathsf{True}_{\Pi_{n+1}}(x))). \qquad (\%)$$

As $\mathrm{I}\Sigma_n \equiv \mathrm{RFN}_{\Pi_{n+2}}(\mathrm{EA})$, we see that for arbitrary x

$$\Box_{\mathrm{EA}}(\mathsf{True}_{\Pi_{n+1}}(i^n_m) \to \mathsf{True}_{\Pi_{n+1}}(\dot{x})) \to (\mathsf{True}_{\Pi_{n+1}}(i^n_m) \to \mathsf{True}_{\Pi_{n+1}}(x)).$$

Again by the induction hypothesis this simplifies to

$$\Box_{\mathrm{EA}}(\mathsf{True}_{\Pi_{n+1}}(i^n_m) \to \mathsf{True}_{\Pi_{n+1}}(\dot{x})) \to \mathsf{True}_{\Pi_{n+1}}(x)$$

which implies (%).

To finish the proof, we reason in $\mathrm{I}\Sigma_n + \sigma$ and suppose $\Box^{J_n}_{\mathrm{I}\Sigma^R_n + \sigma}\bot$. Thus for some $m{\in}J_n$ we have $\Box_{i^n_m \wedge \sigma}\bot$ and also $\Box_{i^n_m}\neg\sigma$. Now $m{\in}J_n$, so also $m{+}1{\in}J_n$ and thus $\mathsf{True}_{\Pi_{n+1}}(\forall x\, (\Box_{i^n_m} \mathsf{True}_{\Pi_{n+1}}(\dot{x}) \to \mathsf{True}_{\Pi_{n+1}}(x)))$. As $\forall x\, (\Box_{i^n_m} \mathsf{True}_{\Pi_{n+1}}(\dot{x}) \to \mathsf{True}_{\Pi_{n+1}}(x))$ is a standard formula (with possibly non-standard parameters) we see that we have the required Π_{n+1} reflection whence $\Box_{i^n_m}\neg\sigma$ yields us $\neg\sigma$. This contradicts with σ. Thus we get $\mathsf{Con}^{J_n}(\mathrm{I}\Sigma^R_n + \sigma)$.

QED

3.4 Arithmetical Completeness of PIL

This subsection is mainly dedicated to prove the next lemma.

Lemma 3.11 *For all A in I we have that if* $\mathrm{PRA} \vdash A$ *then* $\mathbf{PIL} \vdash A$.

The reasoning is completely analogous to that in the proof of Theorem 2.6. We thus need to prove a Lemma 3.18 stating that for any formula A in I we have that $\Box A$ is equivalent over **PIL** to a formula

of the form $\Box^\alpha\bot$, and a Lemma 3.19 which tells us that **PIL** $\vdash A$ whenever **PIL** $\vdash \Box A$. In a series of rather technical lemmas we will work up to these results.

Lemma 3.12 $\mathtt{S} \wedge B \equiv (\mathtt{S} \wedge \Diamond^\beta\top) \vee \Diamond^{\beta+1}\top$ *for some* $\beta \in \omega + 1$.

Proof of Lemma 3.12.
$\mathtt{S} \wedge B \equiv (\mathtt{S} \wedge B) \vee \Diamond(\mathtt{S} \wedge B) \equiv \neg(\neg(\mathtt{S} \wedge B) \wedge \Box\neg(\mathtt{S} \wedge B))$, but $\neg(\mathtt{S} \wedge B) \wedge \Box\neg(\mathtt{S} \wedge B) \leftrightarrow (\mathtt{S} \to \neg B) \wedge \Box(\mathtt{S} \to \neg B) \leftrightarrow (\mathtt{S} \to \neg B) \wedge \Box\neg B$. Now we consider a conjunctive normal form of $\neg B$. Thus, $\neg B$ is equivalent to $\bigwedge\!\!\!\!\bigwedge_i(\Box^{\alpha_i}\bot \to \Box^{\beta_i}\bot)$ for certain $\alpha_i > \beta_i$ (possibly none). So, by Lemma 2.7, $\Box\neg B \leftrightarrow \bigwedge\!\!\!\!\bigwedge_i \Box^{\beta_i+1}\bot \leftrightarrow \Box^{\beta+1}\bot$ for $\beta = \min(\{\beta_i\})$. So,

$$\begin{array}{ll}
(\mathtt{S} \to \neg B) \wedge \Box\neg B & \leftrightarrow \\
(\mathtt{S} \to \neg B) \wedge \Box^{\beta+1}\bot & \leftrightarrow \\
(\mathtt{S} \to \neg B) \wedge (\mathtt{S} \to \Box^{\beta+1}\bot) \wedge \Box^{\beta+1}\bot & \leftrightarrow \\
(\mathtt{S} \to (\bigwedge\!\!\!\!\bigwedge_i(\Box^{\alpha_i}\bot \to \Box^{\beta_i}\bot) \wedge \Box^{\beta+1}\bot)) \wedge \Box^{\beta+1}\bot \qquad (1) &
\end{array}$$

As $\alpha_i > \beta_i \geq \beta$ we have $\beta + 1 \leq \alpha_i$ whence $\Box^{\beta+1}\bot \to \Box^{\alpha_i}\bot$. Thus,

$$\bigwedge\!\!\!\!\bigwedge_i(\Box^{\alpha_i}\bot \to \Box^{\beta_i}\bot) \wedge \Box^{\beta+1}\bot \leftrightarrow \bigwedge\!\!\!\!\bigwedge_i \Box^{\beta_i}\bot \leftrightarrow \Box^\beta\bot,$$

and (1) reduces to $(\mathtt{S} \to \Box^\beta\bot) \wedge \Box^{\beta+1}\bot$. Consequently,

$$\begin{array}{ll}
(\mathtt{S} \wedge B) \vee \Diamond(\mathtt{S} \wedge B) & \leftrightarrow \\
\neg(\neg(\mathtt{S} \wedge B) \wedge \Box\neg(\mathtt{S} \wedge B)) & \leftrightarrow \\
\neg((\mathtt{S} \to \Box^\beta\bot) \wedge \Box^{\beta+1}\bot) & \leftrightarrow \\
(\mathtt{S} \wedge \Diamond^\beta\top) \vee \Diamond^{\beta+1}\top. &
\end{array}$$

QED

By a proof similar to that of Lemma 3.12 we get the following lemma.

Lemma 3.13 $B \equiv \Diamond^{\gamma'}\top$ *for certain* $\gamma' \in \omega + 1$.

In **PIL** we have a substitution lemma in the sense that $\vdash F(C) \leftrightarrow F(D)$ whenever $\vdash C \leftrightarrow D$. We do not have a substitution lemma for equi-interpretable formulas[8] but we do have a restricted form of it.

Lemma 3.14 *If* $C \equiv C'$, $D \equiv D'$, $E \equiv E'$ *and* $F \equiv F'$, *then* $\vdash C \vee D \rhd E \vee F \leftrightarrow C' \vee D' \rhd E' \vee F'$.

[8]We have that $\neg\mathtt{S} \equiv \top$. If the substitution lemma were to hold for equi-interpretable formulas then $\mathtt{S} \equiv \neg(\neg\mathtt{S}) \equiv \bot$ which will turn out not to be the case.

We reason in **PIL**. Suppose that $C \vee D \rhd E \vee F$. We have for any G that $C' \vee D' \rhd G \leftrightarrow (C' \rhd G) \wedge (D' \rhd G)$. As $C' \rhd C \rhd (C \vee D)$ and $D' \rhd D \rhd (C \vee D)$ we have that $C' \vee D' \rhd C \vee D$. Likewise we obtain $E \vee F \rhd E' \vee F'$ thus $C' \vee D' \rhd C \vee D \rhd E \vee F \rhd E' \vee F'$. The other direction is completely analogous.

Lemma 3.15 $\mathtt{S} \wedge \Diamond^\alpha \top \rhd (\mathtt{S} \wedge \Diamond^\beta \top) \vee \Diamond^\gamma \top$ *is provably equivalent to*

$$\begin{cases} \Box^\omega \bot & \text{if } \alpha \geq \min(\{\beta, \gamma\}) \\ \Box^{\alpha+1} \bot & \text{if } \alpha < \beta, \gamma \end{cases}$$

Proof of Lemma 3.15. The case when $\alpha \geq \min(\{\beta, \gamma\})$ is trivial. The identity interpretation always works as $\Diamond^\alpha \top \rightarrow \Diamond^\delta \top$ whenever $\alpha \geq \delta$. So, we consider the case when $\neg(\alpha \geq \min(\{\beta, \gamma\}))$, that is, $\alpha < \beta, \gamma$.

Then we have $\Diamond^\beta \top \rhd \Diamond^{\alpha+1} \top \rhd \Diamond(\Diamond^\alpha \top) \rhd \Diamond(\mathtt{S} \wedge \Diamond^\alpha \top)$ and likewise for $\Diamond^\gamma \top$. Thus, together with our assumption, we get $\mathtt{S} \wedge \Diamond^\alpha \top \rhd (\mathtt{S} \wedge \Diamond^\beta \top) \vee \Diamond^\gamma \top \rhd \Diamond(\mathtt{S} \wedge \Diamond^\alpha \top)$. By Feferman's principle we get $\Box\neg(\mathtt{S} \wedge \Diamond^\alpha \top)$ whence $\Box^{\alpha+1} \bot$. The implication in the other direction is immediate by Fact 3.2. QED

Lemma 3.16 $\Diamond^\alpha \top \rhd (\mathtt{S} \wedge \Diamond^\beta \top) \vee \Diamond^\gamma \top$ *is provably equivalent to*

$$\begin{cases} \Box^\omega \bot & \text{if } \alpha \geq \min(\{\beta + 1, \gamma\}) \\ \Box^{\alpha+1} \bot & \text{if } \alpha < \beta + 1, \gamma \end{cases}$$

Proof of Lemma 3.16. The proof is completely analogous to that of Lemma 3.15 with the sole exception in the case that $\alpha = \beta < \gamma$. In this case

$$\Diamond^\gamma \top \rhd \Diamond^{\alpha+1} \top \rhd \Diamond(\Diamond^\alpha \top) \rhd \Diamond(\mathtt{S} \wedge \Diamond^\alpha \top) \rhd \mathtt{S} \wedge \Diamond^\alpha \top$$

and thus $(\mathtt{S} \wedge \Diamond^\alpha \top) \vee \Diamond^\gamma \top \rhd \mathtt{S} \wedge \Diamond^\alpha \top$. An application of $\mathsf{S_4}$ yields the desired result, i.e. $\Box^{\alpha+1} \bot$.

In case $\alpha \geq \beta + 1$ it is useful to realize that $\Diamond^\alpha \top \rhd \Diamond^{\beta+1} \top \rhd \Diamond(\Diamond^\beta \top) \rhd \Diamond(\mathtt{S} \wedge \Diamond^\beta \top) \rhd \mathtt{S} \wedge \Diamond^\beta \top$. QED

Axiom $\mathsf{S_3}$ tells us that in our logic occurences of $\neg\mathtt{S}$ "under the scope of a $\rhd$" can easily be dispensed with. We don't seem to be in the luxurious position (as is the case in Visser [Vis92]) where we have the same for $\mathtt{S}$. For that reason some more modal work has to be done to show that we *can* get rid of the occurences of $\mathtt{S}$ in compound statements as is shown in the next lemma.

Lemma 3.17 *If C and D are both boolean combinations of $\mathtt{S}$ and sentences of the form $\Box^\gamma \bot$ then we have that* **PIL** $\vdash (C \rhd D) \leftrightarrow \Box^\delta \bot$ *for some $\delta \in \omega + 1$.*

PROOF OF LEMMA 3.17. So, let C and D meet the requirements of the lemma and reason in **PIL**. We get that

$$C \rhd D \leftrightarrow (\mathtt{S} \wedge B_0) \vee (\neg\mathtt{S} \wedge B_1) \rhd (\mathtt{S} \wedge B_2) \vee (\neg\mathtt{S} \wedge B_3)$$

for some B_0, B_1, B_2 and B_3. The righthand side of this bi-implication is equivalent to

$$((\mathtt{S}\wedge B_0)\rhd(\mathtt{S}\wedge B_2)\vee(\neg\mathtt{S}\wedge B_3))\wedge((\neg\mathtt{S}\wedge B_1)\rhd(\mathtt{S}\wedge B_2)\vee(\neg\mathtt{S}\wedge B_3)). \quad (*)$$

We will show that each conjunct of $(*)$ is equivalent to a formula of the form $\Box^\epsilon\bot$. Starting with the left conjunct we get by repeatedly applying Lemma 3.14 that

$\mathtt{S} \wedge B_0 \rhd (\mathtt{S} \wedge B_2) \vee (\neg\mathtt{S} \wedge B_3)$	$\leftrightarrow$	Lemma 3.12
$(\mathtt{S} \wedge \Diamond^\alpha\top) \vee \Diamond^{\alpha+1}\top \rhd (\mathtt{S} \wedge B_2) \vee (\neg\mathtt{S} \wedge B_3)$	$\leftrightarrow$	S_3
$(\mathtt{S} \wedge \Diamond^\alpha\top) \vee \Diamond^{\alpha+1}\top \rhd (\mathtt{S} \wedge B_2) \vee B_3$	$\leftrightarrow$	Lemma 3.13
$(\mathtt{S} \wedge \Diamond^\alpha\top) \vee \Diamond^{\alpha+1}\top \rhd (\mathtt{S} \wedge B_2) \vee \Diamond^{\gamma'}\top$	$\leftrightarrow$	Lemma 3.12
$(\mathtt{S} \wedge \Diamond^\alpha\top) \vee \Diamond^{\alpha+1}\top \rhd (\mathtt{S} \wedge \Diamond^\beta\top) \vee \Diamond^{\beta+1}\top \vee \Diamond^{\gamma'}\top$	$\leftrightarrow$	
$(\mathtt{S} \wedge \Diamond^\alpha\top) \vee \Diamond^{\alpha+1}\top \rhd (\mathtt{S} \wedge \Diamond^\beta\top) \vee \Diamond^\gamma\top$	$\leftrightarrow$	
$(\mathtt{S} \wedge \Diamond^\alpha\top \rhd (\mathtt{S} \wedge \Diamond^\beta\top) \vee \Diamond^\gamma\top) \quad \wedge$		
$(\Diamond^{\alpha+1}\top \rhd (\mathtt{S} \wedge \Diamond^\beta\top) \vee \Diamond^\gamma\top)$	$\leftrightarrow$	Lemma 3.15
$\Box^\mu\bot \wedge (\Diamond^{\alpha+1}\top \rhd (\mathtt{S} \wedge \Diamond^\beta\top) \vee \Diamond^\gamma\top)$	$\leftrightarrow$	Lemma 3.16
$\Box^\mu\bot \wedge \Box^\lambda\bot$	$\leftrightarrow$	
$\Box^\delta\bot$		

for suitable indices $\alpha, \beta, \ldots$. For the right conjunct of $(*)$ we get a similar reasoning. QED

Lemma 3.17 is the only new ingredient needed to prove the next two lemmas in complete analogy to their counterparts 2.8 and 2.9 in **PGL**.

Lemma 3.18 *For any formula A in I we have that A is equivalent in* **PIL** *to a boolean combination of formulas of the form* $\mathtt{S}$ *or $\Box^\beta\bot$, and that $\Box A$ is equivalent in* **PIL** *to $\Box^\alpha\bot$ for some $\alpha \in \omega + 1$.*

Lemma 3.19 *For all A in I we have that* **PIL** $\vdash A$ *whenever* **PIL** $\vdash \Box A$.

The logic **PILS** is defined as follows. The axioms of **PILS** are all the theorems of **PIL** together with $\mathtt{S}$ and $\{\Diamond^\alpha\top \mid \alpha \in \omega\}$. It's sole rule of inference is modus ponens.

Theorem 3.20 **PILS** $\vdash A \Leftrightarrow \mathbb{N} \models A$

Clearly **PILS** is a decidable logic.

3.5 Modal Semantics for PIL, Decidability

As in Section 2.4 we will define a model for **PIL**.

The model $\mathcal{N} = \langle M, R, \{S_m\}_{m \in M}, \Vdash \rangle$ is obtained from the model $\mathcal{M} = \langle M, R, \Vdash \rangle$ as defined in Definition 2.11 as follows. We define $\langle m, 1 \rangle S_{\mathbf{n}} \langle m, 0 \rangle$ for $\mathbf{n} R \langle m, 1 \rangle$ and close off as to have the $S_{\mathbf{n}}$ relations reflexive, transitive and containing R the amount it should.

Theorem 3.21 $\forall n \quad \mathcal{N}, n \Vdash A \Leftrightarrow \mathbf{PIL} \vdash A$

PROOF OF THEOREM 3.21. The proof is completely analogous to that of Theorem 2.12. We only should check that all the instantiations of S_3 and S_4 hold in all the nodes of $\mathcal{N}$.

We first show that S_3 holds at any point $\mathbf{n}$. So, for any B, consider any point $\langle m, i \rangle$ such that $\mathbf{n} R \langle m, i \rangle \Vdash B$. As $\langle m, i \rangle S_{\mathbf{n}} \langle m, 0 \rangle$, we see that $\mathbf{n} \Vdash B \rhd B \wedge \neg \mathsf{S}$.

To see that any instantiation of S_4 holds at any world $\mathbf{n}$ we reason as follows. If $\mathbf{n} \Vdash \Diamond B$ we can pick the minimal $m \in \omega$ such that $(m, 0) \Vdash B$. It is clear that no $S_{\mathbf{n}}$-transition goes to a world where $B \wedge \mathsf{S}$ holds, hence $\mathbf{n} \Vdash \neg(B \rhd B \wedge \mathsf{S})$. QED

The modal semantics gives us the decidability of the logic **PIL**.

4 Logics with Restricted Substitution.

In this section we prove some results concerning the full interpretability logic of PRA and of related systems. Invoking results from the previous sections we obtain upperbounds for our logics.

In Subsection 4.1 we restrict the possible substitutions in Solovay's theorem for the theories $\mathrm{I}\Sigma_n^R$, for $n \geq 2$. This gives us upperbounds for the interpretability logics $\mathbf{IL}(\mathrm{I}\Sigma_{\mathrm{n}}^{\mathrm{R}})$. We also isolate some modal principles and relate them to $\mathbf{IL}(\mathrm{I}\Sigma_{\mathrm{n}}^{\mathrm{R}})$. These modal principles can be considered in a lowerbound as is done in Corollary 4.3.

In Subsection 4.2 we make some general remarks on $\mathbf{IL}(\mathrm{PRA})$ and provide a lowerbound. We also consider a new principle which we call Zambella's principle.

In Subsection 4.3 we isolate a class $\mathcal{B}$ of arithmetical sentences with a "clear mathematical content". We provide logics that generate all provability respectively all interpretability principles of PA that are provable in PA under all substitutions in $\mathcal{B}$. This result yields a rough upperbound for $\mathbf{IL}(\mathrm{PRA})$.

In Definition 4.11 we introduce the notion of a strong normal form of a closed formula and in Lemma 4.13 we give an application of it.

4.1 Theories Strongly Related to PRA

The main interest of this subsection is in Theorem 4.2 where we show that the interpretability logic of $\mathrm{EA} + \Sigma_n\text{-IR}$, we will write $\mathrm{I}\Sigma_n^R$ from now on[9], is strictly contained in **ILM** for $n \geq 2$. For PRA, which is $\mathrm{I}\Sigma_1^R$, the situation is still unsettled.

ILM is known to be the interpretability logic for any essentially reflexive theory T in the sense that $\mathbf{ILM} = \{A \mid \forall * \, T \vdash A^*_{\triangleright}\}$. Here we identify **ILM** with the set of its theorems. The $*$ ranges over realizations and the subscript $\triangleright$ is adopted to emphasize that the binary modal symbol $\triangleright$ is translated as formalized interpretability.

We also consider translations where $\triangleright$ is translated as formalized Π_1-conservativity. So, $T \rhd_{\Pi_1} S$ is a formalization of $\forall\, \pi{\in}\Pi_1 \ (S \vdash \pi \Rightarrow T \vdash \pi)$. By a result of Hájek and Montagna ([HM90], [HM92]) it is known that **ILM** is the Π_1-conservativity logic of any theory S containing $\mathrm{I}\Sigma_1$ in the sense that $\mathbf{ILM} = \{A \mid \forall * \, S \vdash A^*_{\Pi_1}\}$.

With our notation as fixed in Section 1.4 we can write $\mathbf{ILM} = \mathbf{IL}(T)$ for essentially reflexive T. The $\subseteq$ inclusion is clear and the $\supseteq$ inclusion is the completeness result of Berarducci and Shavrukov which says $\mathbf{ILM} \nvdash A \Rightarrow \exists * \, T \nvdash A^*$. By inspection of the proofs of this completeness result it follows that the specific translation is always a disjunction of statements that a certain primitive recursive function has a certain limit. Thus, the realization can always be chosen to be Σ_2 and we get $\mathbf{ILM} = \mathbf{IL}(T) = \mathbf{IL}_{\Sigma_2}(T)$. Likewise, we see that **ILM** is the Π_1-conservativity logic for realizations restricted to Σ_2-sentences over any theory S containing $\mathrm{I}\Sigma_1$.

Beklemishev has shown in ([Bek97]) that $\mathrm{I}\Sigma_n^R$ is equivalent to $(\mathrm{EA})_\omega^{\Sigma_n}$ and $(\mathrm{EA})_\omega^{\Pi_{n+1}}$, and it is easily seen that every Σ_{n+1}-extension (an extension by Σ_{n+1} sentences) of this theory is reflexive. If T is a reflexive theory we have a nice characterization of interpretability, the Orey-Hájek characterization (already provable in EA): $\vdash T \rhd S \leftrightarrow$ "T is Π_1 conservative over S" or stated otherwise $\vdash T \rhd S \leftrightarrow T \rhd_{\Pi_1} S$. These ingredients combine in the following theorem.

Lemma 4.1 $\mathbf{IL}_{\Sigma_2}(\mathrm{I}\Sigma_n^{\mathrm{R}}) = \mathbf{IL}_{\Delta_{n+1}}(\mathrm{I}\Sigma_n^{\mathrm{R}}) = \mathbf{ILM}$ *whenever* $n \geq 2$.

Proof of Lemma 4.1. If, for two classes of sentences we have $X \subseteq Y$, then $\mathbf{IL}_Y(\mathrm{T}) \subseteq \mathbf{IL}_X(\mathrm{T})$. We will thus show that $\mathbf{IL}_{\Sigma_2}(\mathrm{I}\Sigma_n^{\mathrm{R}}) \subseteq \mathbf{ILM}$ and $\mathbf{ILM} \subseteq \mathbf{IL}_{\Delta_{n+1}}(\mathrm{I}\Sigma_n^{\mathrm{R}})$.

First, we prove by induction on the complexity of a modal formula A that $\forall * {\in} \mathsf{Sub}(\Delta_{n+1})$ $\mathrm{I}\Sigma_n^R \vdash A^*_{\Pi_1} \leftrightarrow A^*_{\triangleright}$ and that the logical complexity of $A^*_{\Pi_1}$ is at most Δ_{n+1}. The basis is trivial and the only interesting induction step is whenever $A = (B \rhd C)$. We reason in $\mathrm{I}\Sigma_n^R$:

[9] We will in this section identify a theory with its set of theorems. So, for example, it is not at all clear whether $\mathrm{I}\Sigma_n^R + C^*$ satisfies the Σ_n-induction rule.

$$
\begin{array}{cl}
(B \rhd C)^*_{\rhd} & \leftrightarrow \text{def.} \\
\mathrm{I}\Sigma^R_n + B^*_{\rhd} \rhd \mathrm{I}\Sigma^R_n + C^*_{\rhd} & \leftrightarrow \text{i.h.} \\
\mathrm{I}\Sigma^R_n + B^*_{\Pi_1} \rhd \mathrm{I}\Sigma^R_n + C^*_{\Pi_1} & \leftrightarrow \text{Orey-Hájek} \\
\mathrm{I}\Sigma^R_n + B^*_{\Pi_1} \rhd_{\Pi_1} \mathrm{I}\Sigma^R_n + C^*_{\Pi_1} & \leftrightarrow \text{def.} \\
(B \rhd C)^*_{\Pi_1} &
\end{array}
$$

Note that we have access to the Orey-Hájek characterization as $B^*_{\Pi_1}$ is at most of complexity Δ_{n+1} and thus $\mathrm{I}\Sigma^R_n + B^*_{\Pi_1}$ is a reflexive theory. Also note that $(B \rhd C)^*_{\Pi_1}$ is a Π_2 sentence and thus certainly Δ_{n+1} whenever $n \geq 2$.

If now $\mathbf{ILM} \vdash A$ then $\mathrm{I}\Sigma^R_n \vdash A^*_{\Pi_1}$ and thus whenever $* \in \mathsf{Sub}(\Delta_{n+1})$, $\mathrm{I}\Sigma^R_n \vdash A^*_{\rhd}$ and $\mathbf{ILM} \subseteq \mathbf{IL}_{\Delta_{n+1}}(\mathrm{I}\Sigma^{\mathrm{R}}_{\mathrm{n}})$.

If $\mathbf{ILM} \nvdash A$ then for some $* \in \mathsf{Sub}(\Sigma_2)$ we have $\mathrm{I}\Sigma^R_n \nvdash A^*_{\Pi_1}$ whence $\mathrm{I}\Sigma^R_n \nvdash A^*_{\rhd}$. We may conclude that $\mathbf{IL}_{\Sigma_2}(\mathrm{I}\Sigma^{\mathrm{R}}_{\mathrm{n}}) \subseteq \mathbf{ILM}$. QED

Theorem 4.2 *For $n \geq 2$, $\mathbf{IL}(\mathrm{I}\Sigma^{\mathrm{R}}_{\mathrm{n}}) \subsetneq \mathbf{ILM}$.*

Proof of Theorem 4.2. From Lemma 4.1 we see that $\mathbf{IL}(\mathrm{I}\Sigma^{\mathrm{R}}_{\mathrm{n}}) \subseteq \mathbf{ILM}$. To see that this inclusion is strict, we will expose a realization $*$ such that $\mathrm{I}\Sigma^R_n \nvdash (p \rhd q \to p \wedge \Box r \rhd q \wedge \Box r)^*$.

It is well-known that $\mathrm{I}\Sigma^R_n \subsetneq \mathrm{I}\Sigma_n \subsetneq \mathrm{I}\Sigma^R_{n+1}$ and that, for every $n \geq 1$, $\mathrm{I}\Sigma_n$ is finitely axiomatized. Let σ_n be the single sentence axiomatizing $\mathrm{I}\Sigma_n$. It is also known that $\mathbf{IL}(\mathrm{I}\Sigma_{\mathrm{n}}) = \mathbf{ILP}$ and $\mathbf{ILP} \nvdash p \rhd q \to p \wedge \Box r \rhd q \wedge \Box r$. Thus, for any $n \geq 1$ one can find α_n, β_n and γ_n such that

$$\mathrm{I}\Sigma_n \nvdash \alpha_n \rhd \beta_n \to \alpha_n \wedge \Box\gamma_n \rhd \beta_n \wedge \Box\gamma_n.$$

Note that $\mathrm{EA} \vdash \alpha_n \rhd_{\mathrm{I}\Sigma_1} \beta_n \leftrightarrow \sigma_n \wedge \alpha_n \rhd_{\mathrm{I}\Sigma^R_n} \sigma_n \wedge \beta_n$ and $\vdash \Box_{\mathrm{I}\Sigma_n}\gamma_n \leftrightarrow \Box_{\mathrm{I}\Sigma^R_n}(\sigma_n \to \gamma_n)$.[10] Thus, we have

$$\mathrm{I}\Sigma^R_n \nvdash \sigma_n \wedge \alpha_n \rhd \sigma_n \wedge \beta_n \to \sigma_n \wedge \alpha_n \wedge \Box(\sigma_n \to \gamma_n) \rhd \sigma_n \wedge \beta_n \wedge \Box(\sigma_n \to \gamma_n)$$

and we can take $p^* = \sigma_n \wedge \alpha_n$, $q^* = \sigma_n \wedge \beta_n$ and $r^* = \sigma_n \to \gamma_n$. QED

Let $\mathcal{A}$ denote here the class of all modal interpretability formulas. We define three sets of modal formulas. For example ES_2 will be the set of so-called *essentially* Σ_2*-formulas* for the theories $\mathrm{I}\Sigma^R_n$, $n \geq 1$. The ES_2 formulas will be Σ_2 in $\mathrm{I}\Sigma^R_n$ under any realization $*$.

[10] It seems that we ignore the fact here that $\mathrm{I}\Sigma^R_n$ has no deduction theorem. It is good to recall that in a formalized setting we will always take the axiomatic equivalent of $\mathrm{I}\Sigma^R_n$, that is, $(\mathrm{EA})^{\Pi_{n+1}}_{\omega}$.

$$
\begin{array}{lcl}
ES_2 & := & \Box\mathcal{A} \mid \neg\Box\mathcal{A} \mid ES_2 \wedge ES_2 \mid ES_2 \vee ES_2 \mid \neg(ES_2 \rhd \mathcal{A}) \\
ES_3 & := & \Box\mathcal{A} \mid \neg\Box\mathcal{A} \mid \mathcal{A} \rhd \mathcal{A} \mid ES_3 \wedge ES_3 \mid ES_3 \vee ES_3 \mid \neg(ES_2 \rhd \mathcal{A}) \\
ES_4 & := & \Box\mathcal{A} \mid \mathcal{A} \rhd \mathcal{A} \mid \neg ES_4 \mid ES_4 \wedge ES_4 \mid ES_4 \vee ES_4 \mid ES_4 \to ES_4
\end{array}
$$

Let $\mathsf{M}^{\mathsf{ES_n}}$ be the schema $A \rhd B \to A \wedge \Box C \rhd B \wedge \Box C$ with $A \in ES_n$.[11] If we combine the results from above, we obtain the following picture.

Corollary 4.3

$$
\begin{array}{rcl}
\mathbf{ILB} = \mathbf{ILM}^{\mathsf{ES_2}} & \subseteq & \mathbf{IL}(\mathrm{I}\Sigma_1^{\mathrm{R}}) = \mathbf{IL}(\mathrm{PRA}) \\
\cap & & \\
\mathbf{ILM}^{\mathsf{ES_3}} & \subseteq & \mathbf{IL}(\mathrm{I}\Sigma_2^{\mathrm{R}}) \subsetneq \mathbf{ILM} \\
\cap & & \\
\mathbf{ILM}^{\mathsf{ES_4}} & \subseteq & \mathbf{IL}(\mathrm{I}\Sigma_{\mathrm{n}}^{\mathrm{R}}) \subsetneq \mathbf{ILM}, \quad n \geq 3
\end{array}
$$

It is unknown if the converse of the left to right inclusions also hold. The B in $\mathbf{ILB}$ stands for Beklemishev who first formulated the principle $\mathsf{M}^{\mathsf{ES_2}}$.

4.2 Back to PRA

We have just shown that the interpretability logic of $\mathrm{I}\Sigma_n^R$ is included in $\mathbf{ILM}$ for every $n \geq 2$. For $n = 1$, c.q. PRA, this inclusion is unknown. The proof of Theorem 4.2 also shows that $\mathbf{ILM} \not\subseteq \mathbf{IL}(\mathrm{PRA})$.

Albert Visser has shown in [Vis97] that also $\mathbf{ILP} \not\subseteq \mathbf{IL}(\mathrm{PRA})$. He provides a special instantiation of the principle P, namely $A \rhd \Diamond B \to \Box(A \rhd \Diamond B)$, and shows that this is not generally valid for PRA. The main ingredient of the proof is a result on the Σ_3-completeness of interpretability of reflexive theories in finitely many axioms by Shavrukov ([Sha97]).

As $A \rhd \Diamond B \to \Box(A \rhd \Diamond B)$ is also provable in $\mathbf{ILM}$ we also see that $\mathbf{ILM} \cap \mathbf{ILP} \not\subseteq \mathbf{IL}(\mathrm{PRA})$.

We do have some positive results though. In the previous subsection we have seen that $\mathbf{ILB} \subseteq \mathbf{IL}(\mathrm{PRA})$. As PRA is definitely a reasonable arithmetical theory we have access to all principles that are known to hold in any reasonable arithmetical theory (see [JV00]). We thus see that $\mathbf{ILBM_0P_0W} \subseteq \mathbf{IL}(\mathrm{PRA})$ where with these letters we refer to the corresponding schemata:

$$
\begin{array}{l}
\mathsf{M_0} : A \rhd B \to \Diamond A \wedge \Box C \rhd B \wedge \Box C, \\
\mathsf{W} : A \rhd B \to A \rhd B \wedge \Box\neg A, \\
\mathsf{P_0} : A \rhd \Diamond B \to \Box(A \rhd B).
\end{array}
$$

[11] Note that $\mathsf{M}^{\mathsf{ES_n}\prime} : A \rhd B \to A \wedge (\Box C \vee \Box C') \rhd B \wedge (\Box C \vee \Box C')$ is also valid in $\mathrm{I}\Sigma_n^R$ for $A \in ES_2$. However over $\mathbf{IL}$ it is derivable from $\mathsf{M}^{\mathsf{ES_n}}$.

It is easy to see that M_0 follows from **ILB**: Reason in **ILB** and suppose $A \rhd B$. Then certainly $\Diamond A \rhd B$ whence by Beklemishev's principle $\Diamond A \wedge \Box C \rhd B \wedge \Box C$. Thus $\mathbf{ILBM_0WP_0} = \mathbf{ILBWP_0}$.

It is up to now unclear what the semantic counterpart of B is. Without proof we state such a semantic correspondence for one instance of B. A frame F validates $\Box A \rhd B \to \Box(A) \wedge \Box(C) \rhd B \wedge \Box C$ if and only if for all u, v, w, x in F the following holds $\forall x\ (\neg \mathsf{End}(x) \to \exists u\ (xRu \wedge \mathsf{End}(u) \wedge \forall v\ (uS_x v \to \mathsf{End}(v))))$. In this formula $\mathsf{End}(x)$ is an abbreviation of the formula that says that x is an R-end point, that is, the formula $\neg \exists y\ xRy$.

The Orey-Hájek theorem tells us that for essentially reflexive theories T three arithmetical notions do provably coincide:

- $\alpha \rhd_T \beta$
- $\forall x\ \Box_{T+\alpha}(\mathsf{Con}((T+\beta)\restriction x))$
- $\forall\, \pi \in \Pi^0_1\ (\Box_{T+\beta}\pi \to \Box_{T+\alpha}\pi)$

As PRA is not essentially reflexive these three different notions can be studied independently and can (and probably will) yield different logics. Interaction between these logics does exist whenever α is Σ_2 and consequently PRA $+\ \alpha$ reflexive. Zambella proved in his dissertation ([Zam94]) a fact that might be useful in the study of the Π_1-conservativity logic (the third notion in our list) of PRA. His Lemma 14 on Page 55 reads as follows.

Lemma 4.4 *(Zambella)*[12] *Let T and S be two theories axiomatized by Π_2-axioms. If T and S have the same Π_1 consequences then $T+S$ has no more Π_1 consequences than T or S.*

Zambella notes that the formalization of this theorem, which is proven by model-theoretic means, seems non-trivial. Here we will show that at first sight Zambella's lemma does not add to our knowledge of the interpretability logic of PRA in case it were formalizable.[13]

A translation of the formalized lemma reads $(A \rhd_{\Pi_1} B) \wedge (B \rhd_{\Pi_1} A) \to A \rhd_{\Pi_1} A \wedge B$, where $\neg A, \neg B \in ES_2$. (Recall that PRA is Π_2 axiomatizable.) If, moreover, A and B are both also in ES_2 we have that interpretability and Π_1 conservativity coincide. This gives rise to a new interpretability principle. Let ED_2, the set of *essentially Δ_2-formulas* be defined as

$$ED_2 \quad := \quad \Box\mathcal{A} \mid \neg ED_2 \mid ED_2 \wedge ED_2 \mid ED_2 \vee ED_2 \quad .$$

[12] We have omitted the phrase "is consistent and" here for obvious reasons.

[13] G. Mints has sketched how such a formalization would proceed. Beklemishev has worked out this sketch in an unpublished note [Bek02].

The principle, which we call Zambella's principle Z here, would thus read

$$\mathsf{Z}: \qquad (A \rhd B) \wedge (B \rhd A) \to A \rhd A \wedge B \quad \text{whenever } A, B \in ED_2.$$

Thus, with a formalization of Lemma 4.4 we would get $\mathbf{ILBWP}_0\mathsf{Z} \subseteq \mathbf{IL}(\mathrm{PRA})$. It is noteworthy that B proves many instances of Z.[14] As an example we show that

$$\mathbf{ILB} \vdash (\Box A \wedge \Diamond C \rhd B) \wedge (B \rhd \Box A \wedge \Diamond C) \to \Box A \wedge \Diamond C \rhd B \wedge \Box A \wedge \Diamond C.$$

if $B \in ED_2$. We reason in $\mathbf{ILB}$.

As $B \rhd \Box A \wedge \Diamond C$ and $B \in ES_2$ we have $B \wedge \Box\neg C \rhd \Box A \wedge \Diamond C \wedge \Box\neg C \rhd \bot$, i.e. , $\Box(B \to \Diamond C)$. Together with $\Box A \wedge \Diamond C \rhd B$ this yields $\Box A \wedge \Diamond C \rhd B \wedge \Diamond C$. One more application of B gives the desired result, that is, $\Box A \wedge \Diamond C \rhd B \wedge \Box A \wedge \Diamond C$.

4.3 Two Logics with Restricted Substitution

We first present an easy example of a provability logic with restricted substitution. Since Solovay ([Sol76]) we know that $\mathbf{GL}$ is the provability logic of PA. Also we know that $\mathbf{GL}$ is decidable. If some provability logical principle is unprovable in $\mathbf{GL}$, we can find an arithmetical instantiation of it which is not provable in PA. In symbols: $\mathbf{GL} \nvdash A \Rightarrow \exists * \ \mathrm{PA} \nvdash A^*$. This instantiation $*$ is given by Solovay's proof of the arithmetical completeness of $\mathbf{GL}$.

The arithmetical content of the instantiation $*$ however is not at all clear as Solovay's proof works with limit statements of a primitive recursive function which incorporates the modal countermodel and is defined in terms of its own code. In this subsection we will determine when a non-derivable (in $\mathbf{GL}$) provability principle does have a non-provable (in PA) instantiation with a "clear (meta-)arithmetical content".

We first define the set $\mathcal{B}$ of arithmetical sentences with a "clear arithmetical content".

$$\mathcal{B} := \bot \mid \top \mid \mathsf{Bew}(\mathcal{B}) \mid \mathsf{Con}(\mathcal{B}) \mid \mathcal{B} \to \mathcal{B} \mid \mathcal{B} \vee \mathcal{B} \mid \mathcal{B} \wedge \mathcal{B}$$

Note that $\mathcal{B}$ is just the arithmetical counterpart of the closed fragment of $\mathbf{GL}$. We are thus interested in the following logic $\{A \mid \forall * \in \mathsf{Sub}(\mathcal{B})\ \mathrm{PA} \vdash A^*\}$ which we will denote, in accordance with the notation for interpretability/provability logics with restricted substitution, by $\mathbf{PL}_{\mathcal{B}}(\mathrm{PA})$.

[14] Certainly B proves all instances of Z where one formula contains either conjunctions or disjunctions.

Definition 4.5 *The logic* **RGL** *is obtained by adding the linearity axiom schema*[15] $\Box(\Box A \to B) \vee \Box(\boxdot B \to A)$ *to* **GL**. *Here* $\boxdot B$ *is an abbreviation of* $B \wedge \Box B$.

The logic **RGL** (the **R** stands here for restricted) has been considered before in the literature. It is system J in Chapter 13 of Boolos' book [Boo93]. In the book of Chagrov and Zakhariaschev [CZ97], in Exercise 5.4 of Chapter 5 an equivalent system **GL.3** is treated.

Theorem 4.6 $\mathbf{PL}_{\mathcal{B}}(\mathrm{PA}) = \mathbf{RGL}$

Proof of Theorem 4.6. Let L_n be the linear frame with n elements. For convenience we call the bottom world $n{-}1$ and the top world 0. It is well known that $\mathbf{RGL} \vdash A \Leftrightarrow \forall n\ (L_n \models A)$. Our proof will thus consist of showing that $\forall n\ (L_n \models A) \Leftrightarrow \forall * \in \mathsf{Sub}(\mathcal{B})\ \mathrm{PA} \vdash A^*$.

For the $\Rightarrow$ direction we assume that $\exists * \in \mathsf{Sub}(\mathcal{B})\ \mathrm{PA} \nvdash A^*$ and show that for some $m \in \omega$, $L_m \not\models A$. So, fix a $*$ for which $\mathrm{PA} \nvdash A^*$. The arithmetical formula A^* can be seen as a formula in the closed fragment of **GL**. By the completeness of **GL** we can find a **GL**-model such that $M, x \vdash \neg A^*$. By $\rho(y)$ we denote the rank of y, that is, the length of the longest R-chain that starts in y. Let $\rho(x) = n$. As the valuation of $\neg A^*$ at x solely depends on the rank of x,(see for example [Boo93], Chapter 7, Lemma 3) we see that $L_{n+1}, n \Vdash \neg A^*$ for every possible valuation on L_{n+1}(this is also denoted by $L_{n+1}, n \models \neg A^*$). We define $\mathbf{L}_{n+1}, m \Vdash p \Leftrightarrow L_{n+1}, m \models p^*$. It is clear that that $\mathbf{L}_{n+1}, n \Vdash \neg A$.

An alternative proof of this direction consists of showing that everything provable in **RGL** is provable in PA under any translation in $\mathsf{Sub}(\mathcal{B})$. The only novelty is the linearity axiom. Let $A^* = \bigwedge\!\!\!\bigwedge_i (\Box^{\alpha_i} \bot \to \Box^{\beta_i} \bot)$ and $B^* = \bigwedge\!\!\!\bigwedge_j (\Box^{\gamma_j} \bot \to \Box^{\delta_j} \bot)$. Let β be the minimum of all the β_i and let δ be the minimum of the δ_j. By techniques from Section 2 we see that $\Box(\Box A^* \to B^*) \vee \Box(\boxdot B^* \to A^*)$ is equivalent to $\Box(\Box^{\beta+1} \bot \to \bigwedge\!\!\!\bigwedge_j (\Box^{\gamma_j} \bot \to \Box^{\delta_j} \bot)) \vee \Box(\Box^{\delta} \bot \to \bigwedge\!\!\!\bigwedge_i (\Box^{\alpha_i} \bot \to \Box^{\beta_i} \bot))$.

[15] Equivalently we could add a trichotomy axiom scheme $\Box(A \leftrightarrow B) \vee \Box(A \to \Diamond B) \vee \Box(B \to \Diamond A)$. We can also consider the logic $\mathbf{RGL}_0$ that arises from adding for all $m \in \omega$ the following schema $\Box(F_m \to A) \vee \Box(F_m \to \neg A)$ to **GL**, where $F_m := \Box^{m+1} \bot \wedge \Diamond^m \top$. It is clear that $\mathbf{RGL}_0$ and **RGL** have the same class of characteristic frames. Also, we can not expect that $\mathbf{RGL}_0 \vdash \Box(\Box A \to B) \vee \Box(\boxdot B \to A)$ due to incompactness phenomena in modal logic. One way to circumvent this incompactness is to allow for one occurrence of an existential quantifier ranging over the natural numbers. To make $\mathbf{RGL}_0$ complete, we should thus add $\Diamond A \to \exists n \in \omega\ \Diamond(A \wedge F_n)$. We call this logic $\mathbf{RGL}'$. Indeed $\mathbf{RGL}' \vdash \Box(\Box A \to B) \vee \Box(\boxdot B \to A)$. To see this, we reason in $\mathbf{RGL}'$ and assume $\Diamond(\Box A \wedge \neg B)$ and $\Diamond(\boxdot B \wedge \neg A)$. For some $n, m \in \omega$ we get $\Diamond(\Box A \wedge \neg B \wedge F_m)$ and $\Diamond(\boxdot B \wedge \neg A \wedge F_n)$. Realizing that $m < n \Rightarrow \mathbf{RGL}' \vdash F_n \to \Diamond F_m$ we see that none of the $(n{=}m) \vee (n{<}m) \vee (m{>}n)$ can hold. We conjecture $\mathbf{RGL}'$ to be conservative over **RGL**. It is not clear what a natural semantics for **RGL** should look like.

The first disjunct is provable if $\beta<\delta$ and the second disjunct whenever $\delta\leq\beta$.

For the $\Leftarrow$ direction we fix some $n\in\omega$ such that $L_n \not\models A$ and construct a $*$ in $\mathsf{Sub}(\mathcal{B})$ such that $\mathrm{PA} \nvdash A^*$. Let $\mathbf{L}_n$ be a model such that $\mathbf{L}_n, n{-}1 \Vdash \neg A$. Instead of applying the Solovay construction we can assign to each world m the arithmetical sentence

$$\varphi_m := \mathsf{Bew}^{m+1}_{\mathrm{PA}}(\ulcorner\bot\urcorner) \wedge \mathsf{Con}^m_{\mathrm{PA}}.$$

(The $\mathsf{Bew}^{m+1}_{\mathrm{PA}}(\ulcorner\bot\urcorner)$ are defined in the obvious way being the arithmetical counterpart of $\Box^{m+1}\bot$. Similarly for $\mathsf{Con}^m_{\mathrm{PA}}$.) It is easy to see that

1. $\mathrm{PA} \vdash \varphi_n \to \neg\varphi_m \quad$ if $n \neq m$
2. $\mathrm{PA} \vdash \varphi_n \to \Box(\bigvee_{m<n} \varphi_m)$
3. $\mathrm{PA} \vdash \varphi_n \to \bigwedge_{m<n} \Diamond\varphi_m$

We set $p^* := \bigvee_{\mathbf{L}_n, m \Vdash p} \varphi_m$. Notice that $*$ is in $\mathsf{Sub}(\mathcal{B})$. Using 1, 2 and 3 we can prove a truth-lemma, that is,

$$\mathbf{L}_n, m \Vdash B \Rightarrow \mathrm{PA} \vdash \varphi_m \to B^*.$$

As always, the truth-lemma is obtained by showing by induction on C that for all m

$$\begin{array}{ll} \mathbf{L}_n, m \Vdash C \Rightarrow \mathrm{PA} \vdash \varphi_m \to C^* & \text{and} \\ \mathbf{L}_n, m \not\Vdash C \Rightarrow \mathrm{PA} \vdash \varphi_m \to \neg C^*. & \end{array}$$

For the basic case we need 1. The boolean connectives are trivial. We treat the case when $C = \Box D$.

If $\mathbf{L}_n, m \Vdash \Box D$, then for any $m'<m$, we have $\mathbf{L}_n, m' \Vdash D$, and by the induction hypothesis $\mathrm{PA} \vdash \varphi_{m'} \to D^*$. Thus also $\mathrm{PA} \vdash \Box\varphi_{m'} \to \Box D^*$ and by 2 we get $\mathrm{PA} \vdash \varphi_m \to \Box D^*$.

If $\mathbf{L}_n, m \not\Vdash \Box D$, then for some $m'<m$, we have $\mathbf{L}_n, m' \Vdash \neg D$. By the induction hypothesis $\mathrm{PA} \vdash \varphi_{m'} \to \neg D^*$. Thus $\mathrm{PA} \vdash \Diamond\varphi_{m'} \to \neg\Box D^*$. By 3 we get $\mathrm{PA} \vdash \varphi_m \to \neg\Box D^*$.

So, by our truth-lemma, $\mathbf{L}_n, n-1 \Vdash \neg A \Rightarrow \mathrm{PA} \vdash \varphi_{n-1} \to (\neg A)^*$ and consequently $\mathrm{PA} \vdash \mathsf{Con}(\ulcorner\varphi_{n-1}\urcorner) \to \neg\mathsf{Bew}(\ulcorner A^*\urcorner)$. Thus $\mathbb{N} \models \mathsf{Con}(\ulcorner\varphi_{n-1}\urcorner) \to \neg\mathsf{Bew}(\ulcorner A^*\urcorner)$. As φ_{n-1} is consistent with PA we see that $\mathbb{N} \models \mathsf{Con}(\ulcorner\varphi_{n-1}\urcorner)$ whence $\mathbb{N} \models \neg\mathsf{Bew}(\ulcorner A^*\urcorner)$ and thus $\mathrm{PA} \nvdash A^*$. QED

Instead of giving the φ_m in the proof of the above theorem one could also construct sentences ψ_m so that they satisfy 2 and 3. The ψ_m could be defined recursively top-down by $\psi_m := \mathsf{Bew}(\ulcorner\bigvee_{l<m} \psi_l\urcorner) \wedge \bigwedge_{l<m} \mathsf{Con}(\ulcorner\psi_l\urcorner)$. As our model is linear we get 1 for free, so again we obtain the truth-lemma. The thus obtained ψ_m are actually provably

equivalent to the φ_m but it is interesting to see that the ψ_m are constructed without any application of the recursion theorem or not even the fixed point theorem.

The recursive construction of the ψ_m can also be seen as a degenerate case of a method applied by de Jongh, Montagna and Jumulet ([JJM91]) where they assign arithmetical statements to worlds in a model not by means of limit statements but rather by a direct fixed point construction. Their method will yield sentences equivalent to our φ_m's on linear models.

Another provability logic with restricted substitution has been considered by Visser (see Boolos Chapter 9 Page 136, [Boo93], or Visser [Vis81]) when he demanded the substitutions to be Σ_1. His logic $\mathbf{PL}_{\Sigma_1}(\mathrm{PA})$ (or **GLV** as Boolos calls it) is not a normal modal logic.

We define for a model M and a propositional variable p a set of natural numbers $\mathcal{D}_M(p)$ as follows.

$$\mathcal{D}_M(p) := \{n \in \omega \mid \exists x{\in}M\ [\rho(x){=}n\ \&\ M, x \Vdash p]\}$$

It is easy to see that any model M with $\forall p\ \mathcal{D}_M(p) \cap \mathcal{D}_M(\neg p) = \varnothing$, is bisimilar to some $\mathbf{L}_n$ model. With Theorem 4.6 we can find "natural" counterexamples for non-provable provability principles A of PA whenever $\mathbf{RGL} \nvdash A$. In this case we can thus find a countermodel M with $\forall p\ \mathcal{D}_M(p) \cap \mathcal{D}_M(\neg p) = \varnothing$. In practice this turns out to happen quite often.

We see that all of the above discussion actually holds for any theory T containing $\mathrm{I}\Delta_0 + \mathsf{exp}$. So, certainly we have our next corollary.

Corollary 4.7 $\mathbf{PL}_{\mathcal{B}}(\mathrm{PRA}) = \mathbf{RGL}$

We have thus singled out a class of not generally valid provability principles that have "natural" counterexamples. The same enterprise can be done for the not generally valid interpretability principles of PA with "natural" counterexamples. This class of formulas will be characterized by the complement of $\mathbf{IL}_{\mathcal{B}}(\mathrm{PA})$.

Definition 4.8 *The logic* **RIL** *is obtained by adding the linearity axiom schema* $\Box(\Box A \to B) \vee \Box(\boxdot B \to A)$ *to* **ILM**. *The* M *in* **ILM** *is Montagna's principle:* $A \rhd B \to A \wedge \Box C \rhd B \wedge \Box C$.

Theorem 4.9 $\mathbf{RIL} = \mathbf{IL}_{\mathcal{B}}(\mathrm{PA})$

Proof of Theorem 4.9. We will expose a translation from formulas φ in the language of **RIL** to formulas φ^{tr} in the language of **RGL** such that

$$\mathbf{RIL} \vdash \varphi \Leftrightarrow \mathbf{RGL} \vdash \varphi^{\mathsf{tr}} \quad (*)$$

and

$$\mathbf{RIL} \vdash \varphi \leftrightarrow \varphi^{\mathsf{tr}}. \quad (**)$$

If we moreover know $(***)$: $\mathbf{RIL} \vdash \varphi \Rightarrow \forall * \in \mathsf{Sub}(\mathcal{B})\ \mathrm{PA} \vdash \varphi^*$ we would be done, for then we have

$$\begin{array}{ll} \forall * \in \mathsf{Sub}(\mathcal{B})\ \mathrm{PA} \vdash \varphi^* & \Leftrightarrow \\ \forall * \in \mathsf{Sub}(\mathcal{B})\ \mathrm{PA} \vdash (\varphi^{\mathsf{tr}})^* & \Leftrightarrow \\ \mathbf{RGL} \vdash \varphi^{\mathsf{tr}} & \Leftrightarrow \\ \mathbf{RIL} \vdash \varphi. & \end{array}$$

We first see that $(***)$ holds. It is clear that $\mathbf{ILM} \subseteq \mathbf{IL}_{\mathcal{B}}(\mathrm{PA})$. Thus it remains to show that $\mathrm{PA} \vdash \Box(\Box A^* \to B^*) \vee \Box(\boxdot B^* \to A^*)$ for any $\mathbf{ILM}$ formulas A and B and any $* \in \mathsf{Sub}(\mathcal{B})$. As any formula in the closed fragment of $\mathbf{ILM}$ is equivalent to a formula in the closed fragment of $\mathbf{GL}$, Theorem 4.6 gives us that indeed the linearity axiom holds for the closed fragment of $\mathbf{GL}$.

Our translation will be the identity translation except for $\rhd$. In that case we define

$$(A \rhd B)^{\mathsf{tr}} := \Box(A^{\mathsf{tr}} \to (B^{\mathsf{tr}} \vee \Diamond B^{\mathsf{tr}})).$$

We first see that we have $(**)$. It is sufficient to show that $\mathbf{RIL} \vdash p \rhd q \to \Box(p \to (q \vee \Diamond q))$. An instantiation of the linearity axiom gives us $\Box(\Box\neg q \to (\neg p \vee q)) \vee \boxdot((\neg p \vee q) \wedge \Box(\neg p \vee q) \to \neg q)$. The first disjunct would yield us immediately $\Box(p \to (q \vee \Diamond q))$.

In case of the second disjunct we get by propositional logic $\Box(q \to \Diamond(p \wedge \neg q))$ and thus also $\Box(q \to \Diamond p)$. Now we assume $p \rhd q$. By W, which is provable in $\mathbf{ILM}$, we get $p \rhd q \wedge \Box\neg p$. Together with $\Box(q \to \Diamond(p \wedge \neg q))$, this gives us $p \rhd \bot$, that is $\Box\neg p$. Consequently we have $\Box(p \to (q \vee \Diamond q))$.

We now prove $(*)$. By induction on $\mathbf{RIL} \vdash \varphi$ we see that $\mathbf{RGL} \vdash \varphi^{\mathsf{tr}}$. All the specific interpretability axioms turn out to be provable under our translation in $\mathbf{K4}$. The only axioms where the $\Box A \to \Box\Box A$ axiom scheme is really used is in $\mathsf{J2}$, $\mathsf{J4}$ and M. Alternatively one can reason that

$$\begin{array}{ll} \mathbf{RIL} \vdash \varphi & \Rightarrow \\ \forall * \in \mathsf{Sub}(\mathcal{B})\ \mathrm{PA} \vdash \varphi^* & \Rightarrow \\ \forall * \in \mathsf{Sub}(\mathcal{B})\ \mathrm{PA} \vdash (\varphi^{\mathsf{tr}})^* & \Rightarrow \\ \mathbf{RGL} \vdash \varphi^{\mathsf{tr}}. & \end{array}$$

If $\mathbf{RGL} \vdash \varphi^{\mathsf{tr}}$ then certainly $\mathbf{RIL} \vdash \varphi^{\mathsf{tr}}$ and by $(**)$, $\mathbf{RIL} \vdash \varphi$.

QED

Corollary 4.10 $\mathbf{RIL} = \mathbf{IL}_{\mathcal{B}}(\mathrm{PRA})$. *Consequently* $\mathbf{RIL}$ *is an upper-bound for* $\mathbf{IL}(\mathrm{PRA})$.

Proof of Corollary 4.10. We know that $\mathbf{ILW} \subseteq \mathbf{IL}_{\mathcal{B}}(\mathrm{PRA})$. Also we know that the linearity axiom is contained in $\mathbf{IL}_{\mathcal{B}}(\mathrm{PRA})$. The proof of Theorem 4.9 now can be copied as $\varphi \leftrightarrow \varphi^{\mathsf{tr}}$ is provable in $\mathbf{ILW}$ together with the linearity axiom. Consequently also M can be proved.

QED

In Lemma 4.13 we will give a direct proof of the fact that $A \rhd B \to \Box(A \to (B \vee \Diamond B)) \in \mathbf{IL}_{\mathcal{B}}(\mathrm{PA})$. We include this proof to give a demonstration of how one can work with strong normal forms.

Definition 4.11 *A formula* $A := \bigwedge_{i=1}^{n}(\Box^{\alpha_i}\bot \to \Box^{\beta_i}\bot)$ *is in* strong normal form, *we write s.n.f., if* $0 \leq \beta_1 < \alpha_1 < \ldots < \beta_n < \alpha_n$ *with* $\alpha_n \leq \omega$. *In this case we define* $\beta(A) := \beta_0$ *(ω in case of the empty conjunction) and* $\alpha(A) := \alpha_0$ *(0 in case of the empty conjunction).*

As the trace (as defined in for example [Boo93], Chapter 7) of every closed formula is a finite union of left-closed, right-open intervals, we see that indeed every closed formula of **GL** is equivalent to a unique formula in s.n.f. Thus, the α and the β functions can be uniquely defined to all closed formulas of **GL**. We define the signum of a natural number as follows: $\sigma(n) := 0$ if $n{=}0$, and $\sigma(n) := 1$ if $n{>}0$. Furthermore we define $\overline{\sigma}(n) := 1 - \sigma(n)$.

If $A := \bigwedge_{i=1}^{n}(\Box^{\alpha_i}\bot \to \Box^{\beta_i}\bot)$ is in s.n.f., then

$$\neg A \leftrightarrow \bigwedge_{i=1}^{n+1}(\Box^{\beta_i}\bot \to \Box^{\alpha_i - 1}\bot)$$

where we define $\alpha_0 := 0$ and $\beta_{n+1} := \omega$. This equivalent of $\neg A$ is almost again in s.n.f.. Techniques from Section 2 can be applied to yield the following fact.

Fact 4.12 *In* **GL** *we can prove the following statements for any closed formula A:*

1. $\Box^{\beta(A)}\bot \to A$,
2. $\Box A \leftrightarrow \Box^{\beta(A)+1}\bot$,
3. $A \wedge \Box A \leftrightarrow \Box^{\beta(A)}\bot$,
4. $\beta(\neg A) = \alpha(A) \cdot \overline{\sigma}(\beta(A))$,
5. $\beta(A) = 0 \leftrightarrow \beta(\neg A) \neq 0$,
6. $A \wedge \Box\neg A \leftrightarrow \Box^{\alpha(A)\cdot\overline{\sigma}(\beta(A))+1}\bot$,
7. $A \vee \Diamond A \leftrightarrow \Diamond^{\alpha(A)\cdot\overline{\sigma}(\beta(A))}\top$.

Lemma 4.13 $\mathbf{ILF} \vdash A \rhd B \to \Box(A \to (B \vee \Diamond B))$ *for any closed formulas A,B of* **ILF**.

Proof of Lemma 4.13. So, reason in **ILF** and suppose $A \rhd B$. In **IL** we have $A \equiv A \vee \Diamond A$ thus our assumption reduces, with Fact 4.12.7 to

$$\Diamond^{\alpha(A)\cdot\overline{\sigma}(\beta(A))}\top \rhd \Diamond^{\alpha(B)\cdot\overline{\sigma}(\beta(B))}\top. \quad (+)$$

If $\alpha(A) \cdot \overline{\sigma}(\beta(A)) \geq \alpha(B) \cdot \overline{\sigma}(\beta(B))$, $(+)$ is provable and we should thus show $\Box(A \to (B \vee \Diamond B))$. Again Fact 4.12.7 reduces this to

$\Box(A \to \Diamond^{\alpha(B)\cdot\overline{\sigma}(\beta(B))}\top)$ or equivalently $\Box(\Box^{\alpha(B)\cdot\overline{\sigma}(\beta(B))}\bot \to \neg A)$. This immediately follows from Facts 4.12.1 and 4.12.4 and our assumption that $\alpha(A)\cdot\overline{\sigma}(\beta(A)) \geq \alpha(B)\cdot\overline{\sigma}(\beta(B))$.

If $\alpha(A)\cdot\overline{\sigma}(\beta(A)) < \alpha(B)\cdot\overline{\sigma}(\beta(B))$, $(+)$ is actually equivalent (here we use F) to $\Box^{\alpha(A)\cdot\overline{\sigma}(\beta(A))+1}\bot$. Consequently (using Fact 4.12.1) we have $\Box(\neg A)$ and we are done. QED

5 Appendix, $I\Sigma_1$ Proves the Consistency of PRA on a Cut

The result we present here has been known, be it in some different formulation, for the last two decades, or so. In Subsection 3.3 we have already provided a proof of a more general theorem. We have chosen to also insert our other proof since it is based on different methods. The present proof employs the method of formalized tableaux proofs of inconsistency.

Ignjatovic has shown that $I\Sigma_1$ proves the consistency of PRA on a cut in his dissertation ([Ign90]). He used this result to show that the length of PRA-proofs can be roughly superexponentially larger than the length of the corresponding $I\Sigma_1$ proofs.

His reasoning was based on a paper by Pudlák ([Pud86]). Pudlák showed in this paper by model-theoretic means that GB proves the consistency of ZF on a cut. The cut that Ignjatovic exposes is actually an RCA_0-cut. (See for example Simpson's book [Sim99] for a definition of RCA_0.)

The elements of the cut correspond to complexities of formulas for which a sort of truth-predicate is available. By an interpretability argument it is shown that a corresponding cut can be defined in $I\Sigma_1$. It seems straight-forward to generalize his result to obtain Theorem 5.1.

The proof we present here is a simplification of an argument by Visser. In an unpublished note [Vis90b], Visser adapted a proof of Paris and Wilkie from [WP87] to obtain Theorem 5.1. Lemma 8.10 from the paper *Induction for bounded arithmetic formulas*, [WP87], implies that for every $r\in\omega$ there is an $I\Delta_0 + \mathsf{exp}$-cut such that for every $\sigma\in\Sigma_2$, $I\Delta_0 + \sigma + \mathsf{exp}$ proves the consistency of $I\Delta_0 + \sigma + \Omega_r$ on that cut.

Theorem 5.1 *There exists an $I\Sigma_1$-cut J such that for all $B\in\Sigma_2$ we have $I\Sigma_1 + B \vdash \mathsf{Con}^J(\mathrm{PRA} + B)$*

In our proof of Theorem 5.1 we find it convenient to work with a different definition of PRA from the one we fixed in Section 1.3. We

do not extend the language and will add totality statements to $I\Delta_0$ for an envelope of provably total recursive functions of PRA in a way that is reminiscent of Definition 3.10 from [Joo02].

We define a sequence of functions on the natural numbers as follows

- $\mathsf{Sup}_0(x) = 2{\cdot}x$
- $\mathsf{Sup}_{z+1}(0) = 1$
- $\mathsf{Sup}_{z+1}(x+1) = \mathsf{Sup}_z(\mathsf{Sup}_{z+1}(x))$

The functions Sup_i describe on the standard model a well-known hierarchy; Sup_0 is the doubling function, Sup_1 is the exponentiation function, Sup_2 is superexponentiation, Sup_3 is superduperexponentiation and so on. It is also known that the Sup_i form an envelope for PRA, that is, every provably total recursive function of PRA gets eventually majorized by some Sup_i. (Essentially this is Parikh's theorem.)

We see that $\mathsf{Sup}_z(x) = y$ can be expressed by a Σ_1-formula:[16]

$$(\mathsf{Sup}_z(x) = y) := (\exists s\ \widetilde{\mathsf{Sup}}(s, z, x, y))$$

where $\widetilde{\mathsf{Sup}}(s, z, x, y)$ is the following Δ_0-formula:

$$\begin{gathered}\mathsf{Finseq}(s) \wedge \mathsf{lh}(s){=}z{+}1 \wedge \\ \mathsf{lh}(s_z){=}x{+}1 \wedge \forall\, i{\leq}z\ (\mathsf{Finseq}(s_i) \wedge [(i{<}z) \rightarrow \mathsf{lh}(s_i) = (s_{i+1})_{|\mathsf{lh}(s_{i+1})-2|}]) \\ \wedge\ \forall\, j{<}\mathsf{lh}(s_0)\ (s_0)_j = 2{\cdot}j \wedge \\ \forall\, i{<}\mathsf{lh}(s){-}1\ ((s_{i+1})_0 = 1 \wedge \forall\, j{<}\mathsf{lh}(s_{i+1}){-}1\ ((s_{i+1})_{j+1} = (s_i)_{(s_{i+1})_j})) \\ \wedge\ (s_z)_x = y.\end{gathered}$$

The intuition behind the formula $\widetilde{\mathsf{Sup}}(s, z, x, y)$ is very clear. The s is a sequence of sufficient large parts of the graphs of the $\mathsf{Sup}_{z'}$'s. Thus,

[16] By close inspection of the defining formula we see that $\mathsf{Sup}_z(x){=}z$ can actually be regarded as a $\Delta_0(\mathsf{exp})$-formula. All the elements in the s_{i+1} occur in the s_i. The s_0 is a very easy sequence, namely $[0, 2, 4, \ldots, 2{\cdot}(\mathsf{lh}(s_0){-}1)]$. The $\widetilde{\mathsf{Sup}}(s, z, x, y)$ is defined in such a way that s is the smallest sequence that builds up $\mathsf{Sup}_z(x)$, thus, we see that $2{\cdot}(\mathsf{lh}(s_0) - 1) = \mathsf{Sup}_z(x) = y$. Consequently $\mathsf{lh}(s_0) = \frac{y+2}{2} \leq y$(for $y \geq 2$). We can roughly estimate (bound by above) s by $[[y, \ldots, y], \ldots, [y, \ldots, y]]$, that is, $\mathsf{lh}(s){=}z + 1 \wedge \forall\, i{\leq}z\ (\mathsf{lh}(s_i) = y \wedge \forall\, j{<}y\ (s_i)_j = y)$. A bound on $\overbrace{[y, \ldots, y]}^{y \text{ times}}$ is given by $2^{2^y}{+}c$ (of course this is dependent on our specific coding protocol, but this bound is reasonable to have). A bound on $\overbrace{[2^{2^y}{+}c, \ldots, 2^{2^y}{+}c]}^{y \text{ times}}$ can also be given by some elementary function. Similar considerations lead to the following "equality": $\mathsf{Sup}_z(x + 1) = \mathsf{Sup}_0(\mathsf{Sup}_1(\ldots \mathsf{Sup}_{z-1}(\mathsf{Sup}_z(x) - 1) \ldots - 1) - 1)$. It is not clear if better bounds or "smart calculations" can lower the complexity of a formula defining the graph of $\mathsf{Sup}_z(x)$.

$$s = \begin{cases} [[\mathsf{Sup}_0(0), \mathsf{Sup}_0(1), \ldots, \mathsf{Sup}_0(\mathsf{lh}(s_0) - 1)], \\ \quad [\mathsf{Sup}_1(0), \mathsf{Sup}_1(1), \ldots, \mathsf{Sup}_1(\mathsf{lh}(s_1) - 1)], \\ \qquad\qquad \vdots \\ \quad [\mathsf{Sup}_z(0), \mathsf{Sup}_z(1), \ldots, \mathsf{Sup}_z(\mathsf{lh}(s_z) - 1)]]. \end{cases}$$

Rather weak theories already prove the main properties of the Sup_z functions (without saying anything about the definedness) like

$$\begin{array}{l} \mathsf{Sup}_n(1) = 2, \\ \mathsf{Sup}_n(2) = 4, \\ 1 \leq \mathsf{Sup}_{n+1}(y), \\ x \leq y \rightarrow \mathsf{Sup}_n(x) \leq \mathsf{Sup}_n(y), \\ (n \leq m \wedge x \leq y) \rightarrow \mathsf{Sup}_n(x) \leq \mathsf{Sup}_m(y), \end{array}$$

and so on.

Definition 5.2 PRA *is the first-order theory in the language* $\{+, \cdot, \leq, 0, 1\}$ *using only the connectives* $\neg, \rightarrow$ *and* $\forall$*, with the following* non-logical axioms.

[*A.*] *Finitely many defining* Π_1*-axioms for* $+$*,* $\cdot$*,* $\leq$*, 0 and 1.*[17]

[*B.*] *Finitely many* Π_1 *identity axioms.*

[*C.*] *For every* $\varphi(x, \vec{a}) \in \Delta_0$ *a* Π_1 *induction axiom of the form:*[18]
$\forall x\ \forall z\ (\varphi(0, z) \wedge \forall y{<}x\ (\varphi(y, z) \rightarrow \varphi(y{+}1, z)) \rightarrow \varphi(x, z))$.

[*D.*] *For all* $z \in \omega$ *a* Π_2 *totality statement for the function* $\mathsf{Sup}_z(x)$ *in the following form:* $\forall x\, \exists s\, \exists y{\leq}s\ \widetilde{\mathsf{Sup}}(s, \overline{z}, x, y)$*. Here and in the sequel* $\overline{z}$ *denotes the numeral corresponding to* z*, that is, the string* $\overbrace{1 + \ldots + 1}^{z \text{ times}}$.

The logical axioms *and* rules *are just as usual.*

We shall need in our proof of Theorem 5.1 a formalization of a proof system that has the sub-formula property. Like Paris and Wilkie we shall use a notion of tableaux proofs rather than some sequent calculus. In our discussion below we consider theories T that are formulated using only the connectives $\rightarrow$, $\neg$ and $\forall$. The other connectives will still be used as abbreviations.

[17] We can take for example Kaye's system PA^- from [Kay91] where in Ax 13 we replace the unbounded existential quantifier by a bounded one.

[18] We mean of course a Π_1-formula using only $\neg, \rightarrow$ and $\forall$, that is logically equivalent to the formula given here. By coding techniques, having just one parameter z in our induction axioms, is no real restriction. It prevents, however, getting a non-standard block of quantifiers in non-standard codes of PRA-axioms.

Definition 5.3 *A* tableau proof of a contradiction *from a set of axioms T containing the identity axioms is a finite sequence $\Gamma_0, \Gamma_1, \ldots, \Gamma_r$ where the Γ_i satisfy the following conditions.*

- *For $0 \leq i \leq r$, Γ_i is a sequence of sequences of labeled formulas. The elements of Γ_i are denoted by Γ_i^j. The elements of the Γ_i^j are denoted by $\varphi_{i,j}^k(l)$ where l is the* label *of $\varphi_{i,j}^k$ and is either 0 or 1. In case $l = 1$ in $\varphi_{i,j}^k(l)$, we call $\varphi_{i,j}^k$ the* active *formula of both Γ_i^j and Γ_i. Only non-atomic formulas can be active.*
- *Γ_0 contains just one finite non-empty sequence of labeled formulas. We require $\varphi_{0,0}^k \in T$ for $k < \mathsf{lh}(\Gamma_0^0)$.*
- *In every Γ_r^j $(j < \mathsf{lh}(\Gamma_r))$ there is an atomic formula that also occurs negated in Γ_r^j.*
- *Every $0 \leq i < r$ contains exactly one sequence Γ_i^j with an active formula in it. This sequence in its turn contains exactly one active formula.*
- *For $0 \leq i < r$, we have $\mathsf{lh}(\Gamma_i) \leq \mathsf{lh}(\Gamma_{i+1}) \leq \mathsf{lh}(\Gamma_i) + 1$.*
- *For $0 \leq i < r$, we have $\mathsf{lh}(\Gamma_i^j) \leq \mathsf{lh}(\Gamma_{i+1}^j) \leq \mathsf{lh}(\Gamma_i^j) + 2$.*
- *For $0 \leq i < r$, we have $\varphi_{i,j}^k = \varphi_{i+1,j}^k$ for $k < \mathsf{lh}(\Gamma_i^j)$.*
- *$\mathsf{lh}(\Gamma_i^j) < \mathsf{lh}(\Gamma_{i+1}^j)$ iff Γ_i^j contains the active formula of Γ_i. In this case, with $n = \mathsf{lh}(\Gamma_i^j)$ and $\varphi_{i,j}^m$ the active formula, one of the following holds.*[19]
  - *(β) $\varphi_{i,j}^m$ is of the form $\neg\neg\theta$ in which case $\Gamma_{i+1,j}^n = \theta$ and $\mathsf{lh}(\Gamma_{i+1}^j) = n + 1$.*
  - *(γ) $\varphi_{i,j}^m$ is of the form $\theta_1 \to \theta_2$. In this case $\Gamma_{i+1,j}^n = \neg\theta_1$ and only in this case $\mathsf{lh}(\Gamma_{i+1}) = \mathsf{lh}(\Gamma_i) + 1$. Let $p := \mathsf{lh}(\Gamma_i)$. Γ_{i+1}^p is defined as follows: $\mathsf{lh}(\Gamma_{i+1}^p) = \mathsf{lh}(\Gamma_{i+1}^j) = n + 1$, $\Gamma_{i+1,p}^k = \Gamma_{i+1,j}^k$ for $k < n$ and $\Gamma_{i+1,p}^n = \theta_2$.*
  - *(δ) $\varphi_{i,j}^m$ is of the form $\neg(\theta_1 \to \theta_2)$. Only in this case $\mathsf{lh}(\Gamma_{i+1}^j) = \mathsf{lh}(\Gamma_i^j) + 2$ and $\Gamma_{i+1,j}^n = \theta_1$ and $\Gamma_{i+1,j}^{n+1} = \neg\theta_2$.*
  - *(ϵ) $\varphi_{i,j}^m$ is of the form $\forall x\ \theta(x)$. In this case $\mathsf{lh}(\Gamma_{i+1}^j) = n + 1$ and $\Gamma_{i+1,j}^n = \theta(t)$ for some term t that is freely substitutable for x in $\theta(x)$.*
  - *(ζ) $\varphi_{i,j}^m$ is of the form $\neg\forall x\ \theta(x)$. In this case $\mathsf{lh}(\Gamma_{i+1}^j) = n + 1$ and $\Gamma_{i+1,j}^n = \neg\theta(y)$ for some variable y that occurs in no formula of Γ_i^j.*

It is well-known that φ is provable from T iff there is a tableau proof of a contradiction from $T \cup \{\neg\varphi\}$. The length of tableaux proofs

[19] We start with (β), so that we have the same labels as in Definition 8.9 from [WP87].

can grow superexponentially larger than their regular counterparts. A pleasant feature of tableaux proofs is the sub-formula property.

We will work with some suitable Δ_1-coding of assignments that are always zero on all but finitely many variables. The constant zero valuation is denoted just by 0. Also do we use well-known satisfaction predicates like $\mathsf{Sat}_{\Pi_1}(\pi, \sigma)$ for formulas $\pi \in \Pi_1$ and valuations σ. By $\mathsf{Val}(t, \sigma)$ we denote some Δ_1 valuation function for terms t and assignments σ. By $\Sigma_1(x)$ we denote the predicate that only holds on the standard model on codes of (syntactical) Σ_1-sentences.

PROOF OF THEOREM 5.1. We will expose an $\mathrm{I}\Sigma_1$-cut and show that $\mathrm{I}\Sigma_1 + B \vdash \mathsf{Con}^J(\mathrm{PRA} + B)$ for any $B \in \Sigma_2$(formulated using only $\neg$, $\rightarrow$ and $\forall$). If we would have a J-proof of $\bot$ from $\mathrm{PRA} + B$ in $\mathrm{I}\Sigma_1 + B$ we can also find a tableau proof of a contradiction (not necessarily in J) from $\mathrm{PRA}^J + B$, as $\mathrm{I}\Sigma_1$ proves the totality of the superexponentiation function. By PRA^J we denote the axiom set of PRA intersected with J.

Thus, it suffices to show that $\mathrm{I}\Sigma_1 + B \vdash \mathsf{TabCon}(\mathrm{PRA}^J + B)$. By TabCon we mean the formalization of the assertion that there is no tableau proof of a contradiction.

The cut that does the job is the following:[20]

$$J(z) := \forall\, z' \leq z\, \forall x\, \exists y\, \mathsf{Sup}_{z'}(x) = y.$$

First we see that $J(z)$ indeed defines a cut in $\mathrm{I}\Sigma_1$. Obviously $\mathrm{I}\Sigma_1 \vdash J(0)$. We now see $\mathrm{I}\Sigma_1 \vdash J(z) \rightarrow J(z{+}1)$. For, reason in $\mathrm{I}\Sigma_1$ and suppose $J(z)$. In order to obtain $J(z{+}1)$ it is sufficient to show that $\forall x\, \exists y\, \mathsf{Sup}_{z+1}(x) = y$. This follows from an easy Σ_1-induction.[21]

As $B \in \Sigma_2$ we may assume that $B = \exists x\, A(x)$ with $A \in \Pi_1$.

We reason in $\mathrm{I}\Sigma_1 + B$ and assume $\neg\mathsf{TabCon}(\mathrm{PRA}^J + B)$. As B holds, for some a we have $A(a)$. We fix this a for the rest of the proof. Let $p = \Gamma_0, \Gamma_1, \ldots, \Gamma_r$ be a hypothetical tableau proof of a contradiction from $\mathrm{PRA}^J + B$.

Via some easy inductions a number of basic properties of p is established, like the sub-formula property and the fact that every $\Sigma_1!$-formula in p comes from an PRA-axiom of the form $[D.]$, etcetera. Inductively we define for every Γ_i^j a valuation $\sigma_{i,j}$.

- $\sigma_{0,0} = 0$.

[20] Formally speaking we should use the $\widetilde{\mathsf{Sup}}(s, z, x, y)$ predicate here.

[21] By Theorem 3.13 from [Joo02] we see that if some theory T proves that J is a cut, then automatically $T \vdash \mathrm{I}\Sigma_1$. Our proof could also work if some theory $T \supseteq \mathrm{PRA}$ proved for some cut J' that $\forall\, z{\in}J'\, \forall\, x{\in}J'\, \exists y\, \mathsf{Sup}_z(x){=}y$. We do not know if for some $\mathrm{PRA} \subset T \subset \mathrm{I}\Sigma_1$ such a J' exists. (Probably in such a case we also need $T \vdash \exists z\, \neg J'(z)$.)

- If Γ_i^j contains no active formula, $\sigma_{i+1,j} = \sigma_{i,j}$.
- If Γ_i^j contains an active formula one of (β)-(ζ) applies. Let $m{=}\mathsf{lh}(\Gamma_i^j)$.
  - (β) $\sigma_{i+1,j} = \sigma_{i,j}$.
  - (γ) $\sigma_{i+1,j} = \sigma_{i+1,m} = \sigma_{i,j}$.
  - (δ) $\sigma_{i+1,j} = \sigma_{i,j}$.
  - (ϵ) $\sigma_{i+1,j} = \sigma_{i,j}$.
  - (ζ) In this case essentially an existential quantifier is eliminated. We treat the three possible eliminations.[22]
    - The first existential quantifier in B is eliminated and B is replaced by $A(y)$. In this case $\sigma_{i+1,j} = \sigma_{i,j}$ for all variables different from y. Furthermore we define $\sigma_{i+1,j}(y) = a$.
    - The first existential quantifier in a formula of the form $\exists s\, \exists y{\leq}s\ \widetilde{\mathsf{Sup}}(s, \overline{z}, t, y)$ for some term t and number $z{\in}J$ is eliminated and replaced by $\exists y{\leq}v\ \widetilde{\mathsf{Sup}}(v, \overline{z}, t, y)$ for some variable v. In this case $\sigma_{i+1,j} = \sigma_{i,j}$ for all variables different from v. Furthermore we define $\sigma_{i+1,j}(v)$ to be the minimal number b such that

      $$\exists y{\leq}b\ \widetilde{\mathsf{Sup}}(b, \mathsf{Val}(\overline{z}, \sigma_{i,j}), \mathsf{Val}(t, \sigma_{i,j}), y).$$

      Note that, as $z \in J$, such a number b must exist. (See also footnote 16.)
    - A bounded existential quantifier in a formula of the form $\exists x{\leq}t\ \theta(x)$ is eliminated and $\exists x{\leq}t\ \theta(x)$ is replaced by $y \leq t \wedge \theta(y)$ for some variable y. In this case $\theta(y)$ is in Δ_0 (yet another induction). We define $\sigma_{i+1,j}(y)$ to be the minimal $c \leq \mathsf{Val}(t, \sigma_{i,j})$ such that $\mathsf{Sat}_{\Delta_0}(\ulcorner \theta(\overline{c}) \urcorner, \sigma_{i,j})$ if such a c exists.[23]In case no such c exists, we define $\sigma_{i+1,j}(y) = 0$. For the other variables we have $\sigma_{i+1,j} = \sigma_{i,j}$.

It is not hard to see that $\sigma_{i,j}(x)$ has a Σ_1 or even Δ_1 graph. The proof is now completed by showing by induction on i:

$$\forall\, i{\leq}r\ \exists\, j{<}\mathsf{lh}(\Gamma_i)\, \forall\, k{<}\mathsf{lh}(\Gamma_i^j)\ (\Sigma_1(\ulcorner \varphi_{i,j}^k \urcorner) \rightarrow \mathsf{Sat}_{\Sigma_1}(\ulcorner \varphi_{i,j}^k \urcorner, \sigma_{i,j})). \quad (\dagger)$$

Note that the statement is indeed Σ_1 as in $\mathrm{I}\Sigma_1$ we have the Σ_1 collection principle which tells us that the bounded universal quantifiers

[22]Again, to see (in $\mathrm{I}\Sigma_1$) that these are the only three possibilities, an induction is executed.

[23]We have the Σ_1 minimal number principle at our disposal in $\mathrm{I}\Sigma_1$. With $\overline{c}$ we mean the numeral corresponding to c.

can be somehow pushed inside the unbounded existential quantifier of the Sat_{Σ_1}.

Once we have shown (†), we have indeed finished the proof as every Γ_r^j ($j<\mathsf{lh}(\Gamma_r)$) contains some atomic formula and its negation. Atomic formulas are certainly Σ_1 which gives for some $j<\mathsf{lh}(\Gamma_r)$ and some atomic formula θ, both $\mathsf{Sat}_{\Sigma_1}(\ulcorner\theta\urcorner, \sigma_{r,j})$ and $\mathsf{Sat}_{\Sigma_1}(\ulcorner\neg\theta\urcorner, \sigma_{r,j})$ and we have arrived at a contradiction.[24] Hence $\mathsf{TabCon}(\mathrm{PRA}^J + B)$.

As announced (†) will be proved by induction on i. If $i=0$, as there are no Σ_1-formulas in Γ_0^0, (†) holds in a trivial way.

For the inductive step, let $i<r$ and $j<\mathsf{lh}(\Gamma_i)$ such that

$$\forall\, k<\mathsf{lh}(\Gamma_i^j)\ (\Sigma_1(\ulcorner\varphi_{i,j}^k\urcorner) \to \mathsf{Sat}_{\Sigma_1}(\ulcorner\varphi_{i,j}^k\urcorner, \sigma_{i,j})).$$

We look for $j'<\mathsf{lh}(\Gamma_{i+1})$ such that

$$\forall\, k<\mathsf{lh}(\Gamma_{i+1}^{j'})\ (\Sigma_1(\ulcorner\varphi_{i+1,j'}^k\urcorner) \to \mathsf{Sat}_{\Sigma_1}(\ulcorner\varphi_{i+1,j'}^k\urcorner, \sigma_{i+1,j'}))\quad (\ddagger).$$

If Γ_i^j contains no active formula, $\Gamma_{i+1}^j=\Gamma_i^j$ and $\sigma_{i+1,j}=\sigma_{i,j}$, and we can just take $j'=j$.

So, we may assume that Γ_i^j contains an active formula, say $\varphi_{i,j}^m$, and one of (β)-(ζ) holds. In the cases (β), (γ) and (δ) it is clear which j' should be taken such that (‡) holds. We now concentrate on the two remaining cases.

(ζ). Here $\varphi_{i,j}^m$ is of the form $\exists x\ \theta(x)$. We only need to consider the case that $\exists x\ \theta(x) \in \Sigma_1$. By an easy induction we see that $\exists x\ \theta(x)$ is either Δ_0 or a subformula (modulo substitution of terms) of an axiom of PRA from group $[D]$.

In case $\varphi_{i,j}^m = \exists x\ \theta(x)$ and $\exists x\ \theta(x) \in \Delta_0$, for some $v \notin \Gamma_i^j$, $\varphi_{i+1,j}^m = \theta(v)$. As we know that $\mathsf{Sat}_{\Sigma_1}(\ulcorner\varphi_{i,j}^m\urcorner, \sigma_{i,j})$, we see that $\sigma_{i+1,j}$ is tailored such that $\mathsf{Sat}_{\Delta_0}(\ulcorner\varphi_{i+1,j}^m\urcorner, \sigma_{i+1,j})$ holds. Clearly also $\mathsf{Sat}_{\Sigma_1}(\ulcorner\varphi_{i+1,j}^m\urcorner, \sigma_{i+1,j})$ and we can take $j=j'$ to obtain (‡).

The other possibility is $\varphi_{i,j}^m = \exists s\, \exists\, y{\leq}s\ \widetilde{\mathsf{Sup}}(s, \overline{z}, t, y)$ for some (possibly non-standard) term t. Consequently $\varphi_{i+1,j}^m = \exists\, y{\leq}v\ \widetilde{\mathsf{Sup}}(v, \overline{z}, t, y)$ for some $v \notin \Gamma_i^j$. Again $\sigma_{i+1,j}$ is tailored such that $\mathsf{Sat}_{\Delta_0}(\ulcorner\varphi_{i+1,j}^m\urcorner, \sigma_{i+1,j})$ holds and we can take $j=j'$ to obtain (‡).

(ϵ). We only need to consider the case $\varphi_{i,j}^m = \forall x\ \theta(x)$ with $\theta(x) \in \Sigma_1$. In case $\forall x\ \theta(x) \in \Sigma_1$, the induction hypothesis and the definition of $\sigma_{i+1,j}$ guarantees us that $j=j'$ yields a solution of (‡). So, we may assume that $\forall x\ \theta(x) \notin \Sigma_1$. By an easy induction we see that thus $\forall x\ \theta(x)$ is $A(a)$ or $\theta(x)$ has one of the following forms:

[24] There seems to be some redundancy in employing both tableaux proofs of a contradiction and the Sat predicates as the internal structure of the Sat predicates is somehow reminiscent to that of tableaux proofs of a contradiction.

1. A subformula (modulo substitution of terms) of an axiom of PRA of the form $[A]$ or $[B]$,
2. A subformula (modulo substitution of terms) of an induction axiom $[C]$,
3. $\exists s\, \exists y{\leq}s\ \widetilde{\mathsf{Sup}}(s, \overline{z}, t, y)$ for some (possibly non-standard) term t and some $z{\in}J$.

Our strategy in all cases but 3 will be to show that[25]

$$\forall\sigma\ \mathsf{Sat}_{\Pi_1}(\ulcorner\forall x\ \theta(x)\urcorner, \sigma). \quad \clubsuit$$

This is sufficient as

$$\begin{array}{ll} \forall\sigma\ \mathsf{Sat}_{\Pi_1}(\ulcorner\forall x\ \theta(x)\urcorner, \sigma) & \Rightarrow \\ \forall\sigma\, \forall x\ \mathsf{Sat}_{\Delta_0}(\ulcorner\theta(v)\urcorner, \sigma[v/x]) & \Rightarrow \\ \forall\sigma'\ \mathsf{Sat}_{\Delta_0}(\ulcorner\theta(v)\urcorner, \sigma') & \Rightarrow \\ \forall\sigma\ \mathsf{Sat}_{\Delta_0}(\ulcorner\theta(t)\urcorner, \sigma) & \Rightarrow \\ \forall\sigma\ \mathsf{Sat}_{\Sigma_1}(\ulcorner\theta(t)\urcorner, \sigma). & \end{array}$$

Here v is some fresh variable, $\theta[v/x]$ denotes the formula where x is substituted for v in $\theta(v)$, and $\sigma[v/x]$ denotes the valuation which (possibly) only differs from σ in that it assigns to the variable v the value x.

The strategy to prove 3 is quite similar. The formula $\forall x\, \exists s\, \exists y{\leq}s\ \widetilde{\mathsf{Sup}}(s, z, x, y)$ is a standard formula that holds if $z \in J$, whence for some variable v we have

$$\forall\sigma\ \mathsf{Sat}_{\Pi_2}(\ulcorner\forall x\ \exists s\, \exists y{\leq}s\ \widetilde{\mathsf{Sup}}(s, v, x, y)\urcorner, \sigma[v/z])$$

and thus also

$$\forall\sigma\ \mathsf{Sat}_{\Pi_2}(\ulcorner\forall x\ \exists s\, \exists y{\leq}s\ \widetilde{\mathsf{Sup}}(s, \overline{z}, x, y)\urcorner, \sigma).$$

We immediately see that

$$\forall\sigma\ \mathsf{Sat}_{\Sigma_1}(\ulcorner\exists s\, \exists y{\leq}s\ \widetilde{\mathsf{Sup}}(s, \overline{z}, t, y)\urcorner, \sigma).$$

The proof is thus finished if we have shown $\clubsuit$ in case $\forall x\ \theta(x)$ is either $A(a)$ or a subformula of an axiom of the groups $[A]$, $[B]$ and $[C]$. The only hard case is whenever $\forall x\ \theta(x)$ is a subformula of a PRA axiom of group $[C]$, as the other cases concern true standard Π_1-sentences only. By an easy induction we see that it is sufficient to show that for every $\varphi \in \Delta_0$

$$\forall x\ \mathsf{Sat}_{\Pi_1}(\ulcorner\forall z\ (\varphi(0, z){\wedge}\forall y{<}v\ (\varphi(y, z) \to \varphi(y{+}1, z)) \to \varphi(v, z))\urcorner, \sigma_{0,0}[v/x]).$$

[25] $\forall\sigma\ \mathsf{Sat}_{\Pi_1}(\ulcorner\varphi\urcorner, \sigma)$ is often denoted by $\mathsf{True}_{\Pi_1}(\varphi)$

This is proved by a Π_1-induction on x. Note that in $\mathrm{I}\Sigma_1$ we have indeed access to Π_1-induction as $\mathrm{I}\Sigma_1 \equiv \mathrm{I}\Pi_1$. The fact that φ can be non-standard urges us to be very precise.

If $x{=}0$ we are done if we have shown

$$\mathsf{Sat}_{\Pi_1}(\ulcorner \forall z\ (\varphi(0,z) \wedge \forall\, y{<}0\ (\varphi(y,z) \to \varphi(y+1,z)) \to \varphi(0,z))\urcorner, \sigma_{0,0})$$

or equivalently

$$\forall z\ \mathsf{Sat}_{\Delta_0}(\ulcorner \varphi(0,w) \to \varphi(0,w)\urcorner, \sigma_{0,0}[w/z]).$$

By an easy induction on the length of φ we can show that for any σ

$$\mathsf{Sat}_{\Delta_0}(\ulcorner \varphi(0,w) \to \varphi(0,w)\urcorner, \sigma).$$

For the inductive step we have to show

$$\mathsf{Sat}_{\Pi_1}(\ulcorner \forall z\ (\varphi(0,z) \wedge \forall\, y{<}v\ (\varphi(y,z) \to \varphi(y{+}1,z)) \to \varphi(v,z))\urcorner, \sigma_{0,0}[v/x{+}1])$$

or equivalently that for arbitrary z

$$\mathsf{Sat}_{\Delta_0}(\ulcorner \varphi(0,w) \wedge \forall\, y{<}v\ (\varphi(y,w) \to \varphi(y{+}1,w)) \to \varphi(v,w)\urcorner, \sigma_{0,0}[v/x{+}1][w/z]).^{26}$$

The reasoning by which we obtain this is almost like φ were standard. So, we suppose

$$\mathsf{Sat}_{\Delta_0}(\ulcorner \varphi(0,w) \wedge \forall\, y{<}v\ (\varphi(y,w) \to \varphi(y{+}1,w))\urcorner, \sigma_{0,0}[v/x{+}1][w/z]) \quad (\natural)$$

and set out to prove

$$\mathsf{Sat}_{\Delta_0}(\ulcorner \varphi(v,w)\urcorner, \sigma_{0,0}[v/x+1][w/z]).$$

The induction hypothesis together with some basic properties of the Sat predicates gives us

$$\mathsf{Sat}_{\Delta_0}(\ulcorner \varphi(0,w) \wedge \forall\, y{<}v\ (\varphi(y,w) \to \varphi(y{+}1,w)) \to \varphi(v,w)\urcorner, \sigma_{0,0}[v/x][w/z]). \quad (\sharp)$$

A witnessing sequence for $(\natural)$ is also a witnessing sequence for

$$\mathsf{Sat}_{\Delta_0}(\ulcorner \varphi(0,w) \wedge \forall\, y{<}v\ (\varphi(y,w) \to \varphi(y+1,w))\urcorner, \sigma_{0,0}[v/x][w/z]).$$

Combing this with $(\sharp)$ gives us $\mathsf{Sat}_{\Delta_0}(\ulcorner \varphi(v,w)\urcorner, \sigma_{0,0}[v/x][w/z])$. Also from $(\natural)$ we get $\mathsf{Sat}_{\Delta_0}(\ulcorner \varphi(v,w) \to \varphi(v+1,w)\urcorner, \sigma_{0,0}[v/x][w/z])$, so that we may conclude $\mathsf{Sat}_{\Delta_0}(\ulcorner \varphi(v+1,w)\urcorner, \sigma_{0,0}[v/x][w/z])$. A witnessing sequence for the latter is also a witnessing sequence for

$$\mathsf{Sat}_{\Delta_0}(\ulcorner \varphi(v,w)\urcorner, \sigma_{0,0}[v/x+1][w/z]).$$

QED

[26] By $\sigma[v/x][w/z]$ we mean sequential substitution. This is not an important detail, as we may assume that we have chosen v and w such that no variable clashes occur.

References

[Bek96] L.D. Beklemishev. Bimodal logics for extensions of arithmetical theories. *Journal of Symbolic Logic*, 61(1):91–124, 1996.

[Bek97] L.D. Beklemishev. Induction rules, reflection principles, and provably recursive functions. *Annals of Pure and Applied Logic*, 85:193–242, 1997.

[Bek02] L.D. Beklemishev. Mints' proof of Zambella's priciple. Unpublished, 2002.

[Ber90] A. Berarducci. The interpretability logic of Peano arithmetic. *Journal of Symbolic Logic*, 55:1059–1089, 1990.

[Boo93] G. Boolos. *The Logic of Provability.* Cambridge University Press, Cambridge, 1993.

[CZ97] A. Chagrov and M. Zakharyaschev. *Modal Logic.* Clarendon Press, Oxford, Oxford Logic Guides 35, 1997.

[dJJ98] D. de Jongh and G. Japaridze. The Logic of Provability. In S.R. Buss, editor, *Handbook of Proof Theory.* Studies in Logic and the Foundations of Mathematics, Vol.137., pages 475–546. Elsevier, Amsterdam, 1998.

[Fef60] S. Feferman. Arithmetization of metamathematics in a general setting. *Fundamenta Mathematicae*, 49:35–92, 1960.

[Goo57] R. L. Goodstein. *Recursive Number Theory.* Studies in Logic and the Foundations of Mathematics. North-Holland, 1957.

[Háj71] P. Hájek. On interpretability in set theories I. *Comm. Math. Univ. Carolinae*, 12:73–79, 1971.

[Háj72] P. Hájek. On interpretability in set theories II. *Comm. Math. Univ. Carolinae*, 13:445–455, 1972.

[HM90] P. Hájek and F. Montagna. The logic of Π_1-conservativity. *Archiv für Mathematische Logik und Grundlagenforschung*, 30:113–123, 1990.

[HM92] P. Hájek and F. Montagna. The logic of Π_1-conservativity continued. *Archiv für Mathematische Logik und Grundlagenforschung*, 32:57–63, 1992.

[HP93] P. Hájek and P. Pudlák. *Metamathematics of First Order Arithmetic.* Springer-Verlag, Berlin, Heidelberg, New York, 1993.

[Hv91] P. Hájek and V. Švejdar. A note on the normal form of closed formulas of interpretability logic. *Studia Logica*, 50:25–38, 1991.

[Ign90] A.D. Ignjatovic. *Fragments of first and Second Order Arithmetic and Length of Proofs.* PhD thesis, University of California, Berkeley, 1990.

[JJM91] de D. Jongh, M. Jumelet, and F. Montagna. On the proof of Solovay's theorem. *Studia Logica*, 50:51–70, 1991.

[Joo02] J.J. Joosten. Two proofs of Parsons' theorem. Logic Group Preprint Series 127, University of Utrecht, November 2002.

[JV00] J.J. Joosten and A. Visser. The interpretability logic of *all* reasonable arithmetical theories. *Erkenntnis*, 53(1–2):3–26, 2000.

[Kay91] R. Kaye. *Models of Peano Arithmetic.* Oxford University Press, Oxford, 1991.

[Lei83] D. Leivant. The optimality of induction as an axiomatization of arithmetic. *Journal of Symbolic Logic*, 48:182–184, 1983.

[Ore61] S. Orey. Relative interpretations. *Zeitschrift f. math. Logik und Grundlagen d. Math.*, 7:146–153, 1961.

[Par70] C. Parsons. On a number-theoretic choice schema and its relation to induction. In A. Kino, J. Myhill, and R.E. Vessley, editors, *Intuitionism and Proof Theory*, pages 459–473. North Holland, Amsterdam, 1970.

[Par72] C. Parsons. On n-quantifier induction. *Journal of Symbolic Logic*, 37(3):466–482, 1972.

[Pud85] P. Pudlák. Cuts, consistency statements and interpretations. *Journal of Symbolic Logic*, 50:423–441, 1985.

[Pud86] P. Pudlák. On the length of proofs of finitistic consistency statements in first-order theories. In J.B. et al Paris, editor, *Logic Colloquium '84*, pages 165–196. North–Holland, Amsterdam, 1986.

[Sch87a] D. G. Schwartz. A free-variable theory of primitive recursive arithmetic. *Zeitschrift f. math. Logik und Grundlagen d. Math.*, 33:147–157, 1987.

[Sch87b] D. G. Schwartz. On the equivalence between logic-free and logic-bearing systems of primitive recursive arithmetic. *Zeitschrift f. math. Logik und Grundlagen d. Math.*, 33:245–253, 1987.

[Sha88] V. Shavrukov. The logic of relative interpretability over Peano arithmetic (in Russian). Technical Report Report No.5, Stekhlov Mathematical Institute, Moscow, 1988.

[Sha97] V.Yu. Shavrukov. Interpreting reflexive theories in finitely many axioms. *Fundamenta Mathematicae*, 152:99–116, 1997.

[Sim99] S. G. Simpson. *Subsystems of Second Order Arithmetic.* Springer-Verlag, 1999.

[Sko67] T. Skolem. The foundations of elementary arithmetic established by means of the recursive mode of thought, without the use of apparent variables ranging over infinite domains. In J. van Heijenoort, editor, *From Frege to Godel*, pages 302–333. Iuniverse, Harvard, 1967.

[Smo77] C. Smoryński. The incompleteness theorems. In J. Barwise, editor, *Handbook of Mathematical Logic*, pages 821–865. North Holland, Amsterdam, 1977.

[Smo85] C. Smoryński. *Self-Reference and Modal Logic.* Springer-Verlag, Berlin, 1985.

[Sol76] R.M. Solovay. Provability interpretations of modal logic. *Israel Journal of Mathematics*, 28:33–71, 1976.

[Tai81] W. Tait. Finitism. *Journal of Philosophy*, 78:524–546, 1981.

[TMR53] A. Tarski, A. Mostowski, and R. Robinson. *Undecidable theories.* North–Holland, Amsterdam, 1953.

[Vis81] A. Visser. A propositional logic with explicit fixed points. *Studia Logica*, 40:155–175, 1981.

[Vis90a] A. Visser. Interpretability logic. In P.P. Petkov, editor, *Mathematical Logic*, pages 175–208. Plenum Press, New York, 1990.

[Vis90b] A. Visser. Notes on $\mathrm{I}\Sigma_1$. Unpublished manuscript, 1990?

[Vis91] A. Visser. The formalization of interpretability. *Studia Logica*, 50(1):81–106, 1991.

[Vis92] A. Visser. An inside view of EXP. the closed fragment of the provability logic of $I\Delta_0 + \Omega_1$ with a propositional constant for EXP. *Journal of Symbolic Logic*, 57:131–165, 1992.

[Vis97] A. Visser. An overview of interpretability logic. In M. Kracht, M. de Rijke, and H. Wansing, editors, *Advances in modal logic '96*, pages 307–359. CSLI Publications, Stanford, CA, 1997.

[WP87] A. Wilkie and J. Paris. On the scheme of induction for bounded arithmetic formulas. *Annals of Pure and Applied Logic*, 35:261–302, 1987.

[Zam94] D. Zambella. *Chapters on bounded arithmetic & on provability logic.* PhD thesis, University of Amsterdam, 1994.