

PROVABILITY AND INTERPRETABILITY LOGICS WITH RESTRICTED REALIZATIONS

THOMAS F. ICARD AND JOOST J. JOOSTEN

ABSTRACT. The provability logic of a theory T is the set of modal formulas, which under any arithmetical realization are provable in T . We slightly modify this notion by requiring the arithmetical realizations to come from a specified set Γ . We make an analogous modification for interpretability logics.

We first studied provability logics with restricted realizations, and show that for various natural candidates of theory T and restriction set Γ , where each sentence in Γ has a well understood (meta)-mathematical content in T , the result is the logic of linear frames. However, for the theory Primitive Recursive Arithmetic (PRA), we define a fragment that gives rise to a more interesting provability logic, by capitalizing on the well-studied relationship between PRA and IS_1 .

We then study interpretability logics, obtaining some upper bounds for $\mathbf{IL}(\text{PRA})$, whose characterization remains a major open question in interpretability logic. Again this upper bound is closely related to linear frames. The technique is also applied to yield the non-trivial result that $\mathbf{IL}(\text{PRA}) \subset \mathbf{ILM}$.

1. INTRODUCTION

In a recent discussion on a mailing list on the foundations of mathematics¹ Joe Shipman asked for important theorems that have essentially only one proof. In reply, Giovanni Sambin provided the example of Solovay's arithmetical completeness theorem of the provability logic \mathbf{GL} ([32]).

This paper deals with restricted cases of Solovay's theorem where alternative proof-methods are available. One of the broad motivations for this paper is the hope of obtaining an alternative proof of Solovay's Theorem (see Section 7). However, the method of provability logics with restricted realizations, we feel, merits interest in its own right, as we shall explain shortly. Let us first briefly restate Solovay's Completeness Theorem, which is the cornerstone result in the study of provability logics.

1.1. Provability Logics. The propositional modal logic \mathbf{GL} , known as *Gödel-Löb Logic*, captures exactly the behavior of the standard provability predicate in arithmetic. For a given theory T (e.g. Peano Arithmetic), formulas $\Box A$ are interpreted as, “ A is provable in T ”. It is defined by extending the basic modal logic \mathbf{K} with a schematic formalization of Löb's Theorem (\mathbf{L} in the following definition).

¹FOM mailing list, 8/19/2009, <http://cs.nyu.edu/pipermail/fom/2009-August/013994.html>

Definition 1.1. **GL** is given by all boolean tautologies, in addition to all instances of the following schemata

$$\begin{aligned} \mathbf{K} & : \Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B); \\ \mathbf{L} & : \Box(\Box A \rightarrow A) \rightarrow \Box A. \end{aligned}$$

The logic is closed under modus ponens and necessitation.

GL enjoys modal completeness with respect to a simple class of frames, in particular the class of finite, irreflexive, and transitive frames, which we henceforth refer to as **GL**-frames. The logic is linked to formalized provability via *arithmetical realizations*. An arithmetical realization is a function $*$ that maps propositional variables to sentences in the language of (a given) arithmetic, sending \perp to $0 = 1$. A realization $*$ can be extended uniformly so that we can interpret an arbitrary modal formula as an arithmetical formula by stipulating,

$$\begin{aligned} (A \rightarrow B)^* & = A^* \rightarrow B^* \\ (\Box A)^* & = \text{Bew}_T(\ulcorner A^* \urcorner). \end{aligned}$$

Here $\ulcorner \cdot \urcorner$ is a function that maps a formula φ to its code $\ulcorner \varphi \urcorner$ and $\text{Bew}_T(\cdot)$ is a predicate in the language of T formalizing provability in T , so that $T \vdash \varphi$ just in case $\mathbb{N} \models \text{Bew}_T(\ulcorner \varphi \urcorner)$.

We define $\mathbf{PL}(T)$, the *provability logic* of a theory T , as follows

$$\mathbf{PL}(T) := \{A \mid \forall * T \vdash A^*\}.$$

Since Löb ([26]) it is known that $\mathbf{GL} \subseteq \mathbf{PL}(T)$ for a large class² of theories T . The reverse inclusion is Solovay's completeness result.

Theorem 1.2 (Solovay's Theorem). $\mathbf{PL}(T) = \mathbf{GL}$ for a wide range of theories T .

Solovay proved that whenever $\mathbf{GL} \not\vdash A$, there is a realization $*$ so that $\text{PA} \not\vdash A^*$. An outline of the proof runs as follows. First, a modal countermodel \mathcal{M} in the form of a rooted tree is taken that witnesses $\mathbf{GL} \not\vdash A$. Next, a new root is added to this model. A primitive recursive function f on this model is defined in terms of its own provable limit behavior. This definition is made using an arithmetical fixed point. The function f starts in the newly added root and $f(x)$ remains where it is unless x is a proof that the function does not have the node y , which is accessible from x , as a limit, in which case the function jumps to y . If T is a sound theory, the function must stay where it started, in the newly added root. The realization $*$ is defined as a disjunction of the limit-statements λ_y of the function f , where λ_y says “ y is not the limit of f ”. More specifically $p^* := \bigvee_{\mathcal{M}, y \Vdash p} \lambda_y$.

1.2. Restricted Realizations. This ingenious proof thus gives us the concrete realization $*$. However, the arithmetical content of this realization $*$ is not exactly transparent.³ A natural question to ask is whether we can find translations with more clear arithmetical and proof theoretic content. And conversely, given a set of

²Arithmetical completeness, i.e. that $\mathbf{PL}(T) \subseteq \mathbf{GL}$, is known to hold for any theory extending $\text{I}\Delta_0 + \text{exp}$ (see [21]). For soundness, i.e. $\mathbf{GL} \subseteq \mathbf{PL}(T)$, the theory can be as weak as $\text{I}\Delta_0 + \Omega_1$ (see [7] and [11]).

³There is a paper by de Jongh, Jumelet and Montagna [21] where an alternative proof of Solovay's theorem is given. In that proof, using the diagonal lemma, one finds some sentences with the required properties rather than defining the sentences and then proving the necessary properties. However, the obtained sentences are essentially the same as the ones defined in Solovay's original proof.

arithmetical sentences with a clear arithmetical content, what modal logics results from restricting realizations to this particular set? These questions motivate the following definition. We shall write, *par abus de langage*, $* \in \Gamma$ to mean that the realization $*$ takes on all its values within the set of sentences Γ .

Definition 1.3. $\mathbf{PL}_\Gamma(T) := \{A : \forall * \in \Gamma, T \vdash A^*\}$

From the definition the following lemma is evident.

Lemma 1.4. *If $\Gamma \subseteq \Delta$, then $\mathbf{PL}_\Delta(T) \subseteq \mathbf{PL}_\Gamma(T)$.*

Clearly, by taking Γ to be the set of *all* arithmetical sentences⁴ we get $\mathbf{PL}_\Delta(T) = \mathbf{PL}(T)$. Thus, the simple Lemma 1.4 can be used to establish upperbounds for a provability logic if one is unable to find the full provability logic. For example, it is a long standing open question what the provability logic is of bounded arithmetics such as S_2^1 .⁵

1.3. Applications and plan of the paper. One can thus use Γ to study the provability logic of T . On the other hand, we shall see that $\mathbf{PL}_\Gamma(T)$ can also be used to *characterize* the fragment Γ . For example, in Theorem 2.1 below we consider the closed fragment \mathcal{B} of provability logic, which consists of boolean combinations of iterated (in)consistency statements. This fragment is given by the following grammar.

$$\mathcal{B} := \perp \mid \mathcal{B} \rightarrow \mathcal{B} \mid \Box \mathcal{B}.$$

We shall see that the modal formulas valid under all realizations from this fragment are exactly the formulas valid on all finite strict linear orders. This can be said to provide yet further evidence that reflection principles and iterated consistency statements are inherently linearly ordered.

Moreover, this fact also gives us information on what kind of arithmetical fixed point constructions are needed in the proof of Theorem 1.2. By the modal Fixed Point Theorem, independently due to de Jongh and Sambin (see [29] (de Jongh actually never published his proof)), we know that certain applications of the arithmetical fixed point theorem can be dispensed with. More precisely, if we have a formula $A(x)$ where the x only occurs directly under the scope of a Bew_T predicate then applying the fixed point to this formula does not give us new expressive power. That is, if we can prove $B \leftrightarrow A(\ulcorner B \urcorner)$ then B is actually provably equivalent to a formula in the language of provability logic. Thus, these sort of applications of the arithmetical fixed point theorem only yield formulas in \mathcal{B} , whence, *pace* Theorem 3.2, cannot suffice for a proof of Solovay's completeness result, Theorem 1.2.

Another example of restricting the substitutions is known in the literature. In [34] Visser studied the provability logic that arises when restricting substitutions to Σ_1 sentences.

In Section 4 we shall consider a fragment \mathcal{D} which contains infinitely many copies of \mathcal{B} for increasingly strong provability predicates. It turns out that even for this richer fragment we do not move beyond linear frames (cf. Theorem 4.5). However,

⁴By inspection of Solovay's proof we actually see that $\mathbf{PL}_{\Sigma_2}(T) = \mathbf{PL}(T)$. This is so as all the limit statements λ_y can be taken to be Σ_2 and we stay within that class taking disjunctions.

⁵This question has been studied in depth in [7]. The question also has important connections to matters in computational complexity. For example, it is shown in [11] that if S_2^1 proves Π_1^b -completeness with parameters (Π_1^b is the set of formulas $(\forall x \leq t) \theta$ with θ sharply bounded), then $\text{NP} = \text{coNP}$.

in Section 5 we shall see that there is a natural fragment for PRA whose associated provability logic lies strictly in between the logic of linear frames and **GL**.

In Section 8 we shall see how restricted realizations can also be applied to interpretability logics.

2. FRAGMENTS AND LOGICS

In this section we show that certain conditions on a given fragment translate to a semantic characterization of the corresponding restricted provability logics. First, some preliminaries on basic relational semantics for provability logics.

Recall that a *frame* \mathbb{F} for **GL** is an ordered pair $\langle W, R \rangle$, where W is a set of points and $R \subseteq W \times W$ is a finite, irreflexive, and transitive relation. Given a set Prop of propositional variables, a *model* \mathcal{M} based on \mathbb{F} is a triple $\langle W, R, V \rangle$, where $V : \text{Prop} \rightarrow \wp(W)$ is a *valuation function* assigning to each variable the set of the points where it is true. We shall also write V for the straightforward extension of V to arbitrary modal formulas. We then write $\langle W, R, V \rangle, w \models A$, just in case $w \in V(A)$. We write $\langle W, R, V \rangle \models A$ if $A \in V(w)$ for all $w \in W$. Overloading notation, we also write $\langle W, R \rangle \models A$, if $\langle W, R, V \rangle \models A$ for all V . We say A is *valid* in the model and in the frame, respectively.

When dealing with fragments, however, arbitrary variables will not be present. All of the fragments we shall consider in this paper will extend the fragment \mathcal{B} defined above, by adding constants $\sigma_1, \sigma_2, \sigma_3, \dots$, with some clear arithmetical content. As these constants will be fixed, and as we would like to characterize the sentences in this fragment modally, we shall add constants s_1, s_2, s_3, \dots , to our modal language, and correspondingly extend the definition of a realization to ensure that $(s_i)^* = \sigma_i$. In fact, given this convention, we will be able to define our fragments in a single language and throughout treat each constant simultaneously as a constant in the modal language and as a specified arithmetical formula, disambiguating whenever the distinction is not clear from context. In other words, we will usually not distinguish between A and A^* .

On the other hand, as far as the relational semantics is concerned, the constants s_1, s_2, s_3, \dots are simply treated as variables. Therefore the above notation is extended in the obvious way to this setting.

Suppose we would like to obtain a modal characterization of $\mathbf{PL}_{\mathcal{F}}(T)$. Under certain circumstances, it suffices to know how \mathcal{F} is characterized according to T . To be precise, if we have a model \mathcal{M} based on a frame \mathbb{F} , such that for each $A \in \mathcal{F}$, the following condition holds:

$$(1) \quad T \vdash A \Leftrightarrow \mathcal{M} \models A,$$

then, as is shown in Theorem 2.1 below, $\mathbf{PL}_{\mathcal{F}}(T) = \mathcal{L}(\mathbb{F})$. Here, $\mathcal{L}(\mathbb{F})$ is the set of formulas in the basic modal language (with propositional variables) valid on the frame \mathbb{F} .

There are two side conditions to our theorem. One of them involves image-finiteness. We call a model *image-finite* if $\{y : xRy\}$ is finite for each x . We shall denote the set $\{y : xRy\} \cup \{x\}$ by $x \uparrow$. Our theorem thus reads as follows:

Theorem 2.1. *Suppose that (1) holds for a model \mathcal{M} based on frame \mathbb{F} . Suppose moreover that \mathcal{M} is image-finite and that each point $x \in \mathcal{M}$ is uniquely definable by a formula $D_x \in \mathcal{F}$. Then, we have that $\mathbf{PL}_{\mathcal{F}}(T) = \mathcal{L}(\mathbb{F})$.*

Proof. In the light of (1) it suffices to prove that

$$\forall * \in \mathcal{F}, \mathcal{M} \vDash B^* \Leftrightarrow \mathbb{F} \vDash B.$$

\Leftarrow Consider some arbitrary $* \in \mathcal{F}$ and define $V_*(p) := \{i : \mathcal{M}, i \vDash p^*\}$. By induction on A we see that for each $i \in \mathbb{F}$

$$\langle \mathbb{F}, V_* \rangle, i \Vdash A \Leftrightarrow \mathcal{M}, i \vDash A[p/p^*]$$

and we are done.

\Rightarrow Given some $i \in \mathbb{F}$ and some arbitrary valuation V we define $*$ by

$$p^* := \bigvee_{x \in V(p) \cap i \uparrow} D_x.$$

As the frame is image-finite, the disjunction is finite. By an induction⁶ on C we see again that

$$\langle \mathbb{F}, V \rangle, i \vDash C \Leftrightarrow \mathcal{M}, i \vDash C^*.$$

As i was arbitrary, we see that $\mathbb{F} \vDash C$. □

As we shall see below, in many occasions we actually will have something stronger than (1). In particular we shall often find ourselves in a situation where we have, apart from the frame, also a modal logic \mathbf{L} for which we have

$$T \vdash A \Leftrightarrow \mathbf{L} \vdash A \Leftrightarrow \mathcal{M} \vDash A.$$

This logic \mathbf{L} will facilitate our calculations considerably.

3. THE CLOSED FRAGMENT

With Theorem 2.1 we can calculate our first provability logic with restricted substitutions. Recall the definition of the closed fragment \mathcal{B} in Subsection 1.3.

Definition 3.1. $\mathbf{GL.3}$ is the logic \mathbf{GL} together with the linearity axiom:

$$\Box(\Box A \rightarrow B) \vee \Box(\Box^+ B \rightarrow A).$$

Here and below, $\Box^+ A$ is short for $A \wedge \Box A$.

Theorem 3.2. $\mathbf{PL}_{\mathcal{B}}(T) = \mathbf{GL.3}$ for a large class⁷ of theories T .

Proof. It is well known that the truth of a closed formula at a particular point in a model depends solely on the rank of that point. Here, the rank of a point x is defined as the supremum of lengths of paths leading from x to a leaf. See for example Chapter 7 from [10].

Thus, the linear frame $\langle \omega, > \rangle$ is universal for \mathcal{B} in the sense that if a formula $\varphi \in \mathcal{B}$ is false at some point in some frame, then it is actually false at some point in $\langle \omega, > \rangle$. Thus, by Theorem 1.2, we have $T \vdash A \Leftrightarrow \langle \omega, > \rangle \vDash A$.

⁶In order to get the inductive step for the \Box operator going we should prove the slightly stronger statement that for all $j \in i \uparrow$ we have $\langle \mathbb{F}, V \rangle, j \vDash C \Leftrightarrow \mathcal{M}, j \vDash C^*$.

⁷See Footnote 2 on conditions on theories. The current proof of this theorem invokes Solovay's completeness result, Theorem 1.2, in full. However, in [22] it is shown how we can substitute the use of Solovay's completeness result by the proof of Theorem 2.1. Thus, Theorem 3.2 actually holds for a larger class of theories including $\mathbf{I}\Delta_0 + \Omega_1$.

Furthermore, it is known that the logic of the frame $\langle \omega, > \rangle$ is axiomatized by **GL.3**. (See, for example, Chapter 13 of [10].) Thus, $\langle \omega, > \rangle \models A \Leftrightarrow \mathbf{GL.3} \vdash A$ and Condition 1 is satisfied for any model based on $\langle \omega, > \rangle$.

Note that $\langle \omega, > \rangle$ is image-finite and that the point n is defined by $\diamond^n \top \wedge \square^{n+1} \perp$. Thus, by Theorem 2.1 we have our result. \square

4. SUBSTITUTIONS FROM THE CLOSED FRAGMENT OF **GLP**

Japaridze's Logic **GLP** ([19]) describes all of the universally valid schemata for reflection principles of restricted logical complexity in arithmetic. It is formulated in a language with infinitely many modalities, where $[n]A$ is read arithmetically as,

A is provable from T along with all true Π_n sentences.

Arithmetical completeness with respect to this interpretation was proven in [18], for sound theories containing only a modest amount of arithmetic.

Definition 4.1. **GLP** is given by the following axiom schemata,

- (i) All boolean tautologies;
- (ii) $[n]([n]A \rightarrow A) \rightarrow [n]A$, for all n ;
- (iii) $[m]A \rightarrow [n]A$, for $m \leq n$;
- (iii) $\langle m \rangle A \rightarrow [n]\langle m \rangle A$, for $m < n$;

in addition to the rules of modus ponens and necessitation for each $[n]$.

While **GLP** does not admit of any frame semantics, various other models have been given (see, e.g. [3] and [4]). In particular, Ignatiev [18] has defined a *universal frame* for the closed fragment of **GLP**, denoted **GLP**₀, which will be of use.⁸

Define \mathcal{D} to be the fragment given by the following infinite grammar:

$$\mathcal{D} := \perp \mid \mathcal{D} \rightarrow \mathcal{D} \mid [0]\mathcal{D} \mid [1]\mathcal{D} \mid [2]\mathcal{D} \mid \dots$$

That is, **GLP**₀ is simply **GLP** restricted to the fragment \mathcal{D} , with no variables.

We can describe Ignatiev's universal frame for **GLP**₀ as follows. Let Ω consist of the set of ω -sequences of ordinals $(\alpha_0, \alpha_1, \alpha_2, \dots)$, where each $\alpha_i < \epsilon_0$. Recall ϵ_0 is the least fixed point of the equation $\omega^\alpha = \alpha$. If the Cantor Normal Form of α is $\omega^{\lambda_n} + \dots + \omega^{\lambda_1}$, then let $e(\alpha) := \lambda_1$ and set $e(0) = 0$.

Definition 4.2. Ignatiev's universal frame is defined as $\mathcal{U} := \langle U, \{R_n\}_{n < \omega} \rangle$, with,

$$U := \{ \vec{\alpha} \in \Omega : \forall i < \omega, \alpha_{i+1} \leq e(\alpha_i) \};$$

$$\vec{\alpha} R_n \vec{\beta} :\Leftrightarrow (\forall m < n, \alpha_m = \beta_m \ \& \ \alpha_n > \beta_n).$$

Notice that each point in U can be seen as a finite, strictly decreasing sequence of ordinals less than ϵ_0 , as each sequence ends in an infinite tail of zeros. For a visualization of the frame, see Figure 1.

A point of the form $(\alpha, e(\alpha), e(e(\alpha)), \dots)$, where $\alpha_{i+1} = e(\alpha_i)$ for all i , is called a *root point*, and is denoted by $\hat{\alpha}$ when α is the first coordinate. Thus every coordinate of $\hat{\alpha}$ is uniquely determined by α . The following lemma is then obvious, given the definition of \mathcal{U} .

Lemma 4.3. *If $\hat{\alpha}$ and $\hat{\beta}$ are root points, then either $\hat{\alpha} R_0 \hat{\beta}$, $\hat{\beta} R_0 \hat{\alpha}$, or $\hat{\alpha} = \hat{\beta}$.*

⁸This frame is studied in detail in [5] and [16].

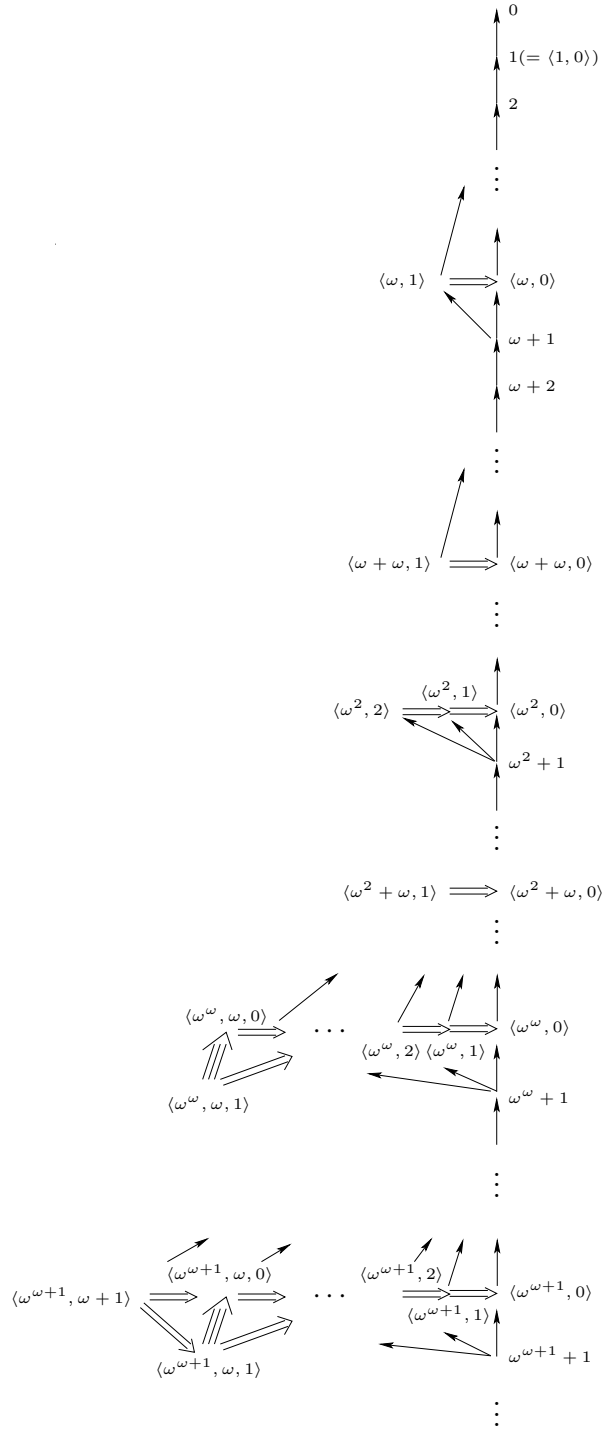


FIGURE 1. The universal model for \mathbf{GLP}_0

In addition to the more routine soundness, the following strong completeness theorem has also been proven using several different methods in the works cited above.

Theorem 4.4. *If $\mathbf{GLP}_0 \not\vdash A$, then there is a root point $\hat{\alpha} \in U$, such that $\mathcal{U}, \hat{\alpha} \not\vdash A$.*

With these results we can now show that even with this much richer fragment the resulting provability logic is exactly the same as for the fragment with only the single \Box -operator (c.f. Theorem 3.2).

Theorem 4.5. $\mathbf{PL}_{\mathcal{D}}(\mathbf{PRA}) = \mathbf{GL.3}$.

Proof. By Theorem 3.2, by Lemma 1.4 and by observing that \Box is just $[0]$, it is clear that $\mathbf{PL}_{\mathcal{D}}(\mathbf{PRA}) \subseteq \mathbf{GL.3}$. For the other inclusion, we must show, under the arithmetical interpretation,

$$\mathbf{PRA} \vdash \Box(\Box A \rightarrow B) \vee \Box(\Box^+ B \rightarrow A),$$

for any $A, B \in \mathcal{D}$. However, this follows by arithmetical completeness and by the universality of Ignatiev's frame.

For, suppose $\mathcal{U}, \vec{\alpha} \vDash \Diamond(\Box A \wedge \neg B) \wedge \Diamond(\Box^+ B \wedge \neg A)$, for some $\vec{\alpha}$. By Theorem 4.4 there are root points $\hat{\beta}$ and $\hat{\gamma}$, such that $\mathcal{U}, \hat{\beta} \vDash \Box A \wedge \neg B$, and $\mathcal{U}, \hat{\gamma} \vDash \Box^+ B \wedge \neg A$. By Lemma 4.3, either $\hat{\beta} R_0 \hat{\gamma}$, $\hat{\gamma} R_0 \hat{\beta}$, or $\hat{\beta} = \hat{\gamma}$. All three lead to contradiction. \square

5. NON-LINEAR \mathbf{GL} -FRAMES

Theorems 3.2 and 4.5 suggest that it may not be straightforward to define a fragment whose associated restricted provability logic is anything other than $\mathbf{GL.3}$ or just \mathbf{GL} . In this section we fill in this gap by giving some sufficient conditions on constants, so that we obtain logics of non-linear \mathbf{GL} -frames. We will be working with generic fragments \mathcal{F}_n , with some finite number n of constants:

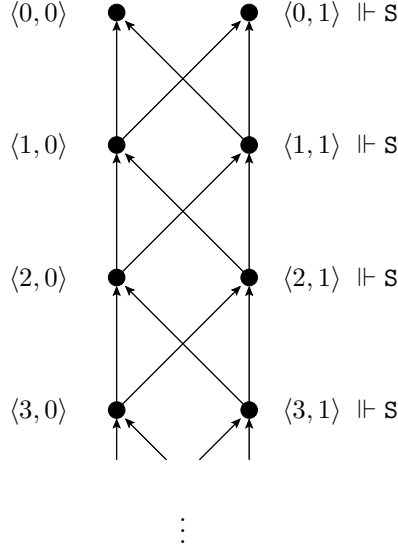
$$\mathcal{F}_n := s_1 \mid s_2 \mid \dots \mid s_n \mid \perp \mid \mathcal{F}_n \rightarrow \mathcal{F}_n \mid \Box \mathcal{F}_n$$

As before, we will be viewing formulas in \mathcal{F}_n simultaneously as arithmetical formulas, where each s_i is a specified formula in the language of arithmetic and \Box is the standard provability predicate, and as modal formulas, where each s_i is interpreted as a constant and \Box is a normal modal operator.

5.1. Fragments, Logics and Models. Let \vec{s}_i stand for the sentence $\bigwedge_{j \in J} s_{j+1} \wedge \bigwedge_{k \in K} \neg s_{k+1}$, where J is the set of places in the binary expansion for i with value 1, and K is the complement of J in $\{0, \dots, i-1\}$. Then we define the following class of logics.

Definition 5.1. *The logic \mathbf{FGL}_n is formulated in the language \mathcal{F}_n and thus, contains no propositional variables. The axioms and rules are specified by the axioms and rules of \mathbf{GL} together with the list of the 2^n many axioms below, one axiom for each Boolean combination of the s_i . The B in these axioms stands for any formula that is a Boolean combination of formulas of the form $\Box^\alpha \perp$, where $\alpha < \omega + 1$ and $\Box^\omega \perp := \top$.*

$$\begin{aligned} & \Box(\vec{s}_0 \rightarrow B) \rightarrow \Box B; \\ & \vdots \\ & \Box(\vec{s}_{2^n-1} \rightarrow B) \rightarrow \Box B. \end{aligned}$$


 FIGURE 2. The model \mathfrak{G}_1^\bullet

These logics \mathbf{FGL}_n come with an associated model, based on the following frames:

Definition 5.2. *The frame $\mathfrak{G}_n := \langle G_n, R_n \rangle$, where $G_n := \{ \langle m, i \rangle : m \in \omega, i < 2^n \}$, and $\langle m, i \rangle R_n \langle p, j \rangle$ just in case $p < m$.*

The associated model defined on this frame is given *via* the binary expansion, where J_j is given as above, relative to j .

Definition 5.3. \mathfrak{G}_n^\bullet is the triple $\langle G_n, R_n, V_n \rangle$, where $V_n(s_j) = \{ \langle m, i \rangle : i \in J_j \}$.

For a visualization of \mathfrak{G}_1^\bullet , see Figure 2.

Theorem 5.4. *For all formulas $A \in \mathcal{F}_n$, $\mathbf{FGL}_n \vdash A$, if and only if $\mathfrak{G}_n^\bullet \models A$.*

Proof Sketch. The full proof for the case of \mathcal{F}_1 is established in [23]. Here we give a sketch for the general case. Soundness is routine. For completeness, we use the following two lemmata.

Lemma 5.5. *Each $A \in \mathcal{F}_n$ is equivalent in \mathbf{FGL}_n to a Boolean combination of formulas of the form s_1, \dots, s_n , or $\Box^\alpha \perp$. In particular $\mathbf{FGL}_n \vdash \Box A \leftrightarrow \Box^\alpha \perp$ for some $\alpha < \omega + 1$.*

Lemma 5.6. *If $\mathbf{FGL}_n \vdash \Box A$, then $\mathbf{FGL}_n \vdash A$.*

These lemmata are straightforwardly proven by manipulation of modal normal forms. Completeness is then clear. If $\mathbf{FGL}_n \not\vdash A$, then by Lemma 5.6, $\mathbf{FGL}_n \not\vdash \Box A$, and by Lemma 5.5, $\mathbf{FGL}_n \vdash \Box A \leftrightarrow \Box^\alpha \perp$, for some $\alpha < \omega$ (in particular $\alpha \neq \omega$). By soundness, for any point $\langle m, i \rangle \in G_n$ we know $\mathfrak{G}_n^\bullet, \langle m, i \rangle \models \Box A \leftrightarrow \Box^\alpha \perp$. Certainly $\mathfrak{G}_n^\bullet, \langle \alpha, 0 \rangle \not\models \Box^\alpha \perp$, so $\mathfrak{G}_n^\bullet, \langle \alpha, 0 \rangle \not\models \Box A$. That, in turn, means for some $\langle \beta, j \rangle$ with $\beta < \alpha$, we have $\mathfrak{G}_n^\bullet, \langle \beta, j \rangle \models \neg A$. So A is falsified on \mathfrak{G}_n^\bullet . \square

5.2. Conditions for completeness. Suppose we have a given theory T and some fragment \mathcal{F}_n , and we would like a characterization of $\mathbf{PL}_{\mathcal{F}_n}(T)$. In Section 2 we showed that if condition (1) holds for some logic \mathbf{L} and model \mathcal{M} , then Theorem 2.1 will follow. Recall Condition (1):

$$T \vdash A \Leftrightarrow \mathbf{L} \vdash A \Leftrightarrow \mathcal{M} \vDash A.$$

To show (1) holds for this case, one merely needs to show arithmetical soundness and completeness of \mathbf{L} for T . However, given Lemmata 5.5 and 5.6, arithmetical completeness of \mathbf{L} depends only on arithmetical soundness of \mathbf{L} .

To see this, suppose $\mathbf{FGL}_n \not\vdash A$. Then by Lemma 5.5, $\mathbf{FGL}_n \not\vdash \Box A$. Since $\mathbf{FGL}_n \vdash \Box A \Leftrightarrow \Box^\alpha \perp$, for $\alpha \neq \omega$, as long as we have soundness of \mathbf{L} , $T \vdash \Box A \Leftrightarrow \Box^\alpha \perp$, under the arithmetical interpretation. Now, if moreover T is a sound theory in the sense that it does not prove any false statements we get $T \not\vdash \Box A$, from which it follows $T \not\vdash A$.

Consequently, the following is a corollary of Theorem 2.1 and Theorem 5.4. Note that both image finiteness and definability of the states in the model \mathfrak{G}_n^\bullet are evident.

Corollary 5.7. $\mathbf{PL}_{\mathcal{F}_n}(T) = \mathcal{L}(\mathfrak{G}_n)$ whenever $[\mathbf{FGL}_n \vdash A \Rightarrow T \vdash A]$.

In Section 6, we shall see that each of these frames \mathfrak{G}_n has a simple axiomatization. For the rest of this section, we exhibit a suitable constant for the case of \mathcal{F}_1 .

5.3. A Constant for $\mathbf{I}\Sigma_1$. Recall $\mathbf{I}\Sigma_1$ is the theory \mathbf{Q} ([33]) along with induction over Σ_1 formulas. This theory is finitely axiomatizable, so let σ stand for the sentence axiomatizing it. We then define the fragment \mathcal{Q} as a special case of \mathcal{F}_1 :

$$\mathcal{Q} := \sigma \mid \perp \mid \mathcal{Q} \rightarrow \mathcal{Q} \mid \Box \mathcal{Q}$$

Our theory T will be Primitive Recursive Arithmetic (PRA), essentially just \mathbf{Q} with function symbols for all of the primitive recursive functions and induction over Δ_0 formulas. The relationship between $\mathbf{I}\Sigma_1$ and PRA is well studied and understood ([27], [28], [1]). By Corollary 5.7, we need to show that \mathbf{FGL}_n is sound with respect to PRA. It is already well known that $\mathbf{PL}(\text{PRA}) = \mathbf{GL}$, so certainly all the axioms and rules of \mathbf{GL} are sound. We need only observe the following also hold:

- (i) $\text{PRA} \vdash \Box(\sigma \rightarrow B) \rightarrow \Box B$,
- (ii) $\text{PRA} \vdash \Box(\neg\sigma \rightarrow B) \rightarrow \Box B$.

In fact, item (i) is a direct consequence of what is known as Parson's Theorem (named after Charles Parsons, but discovered independently by Grigori Mints and Gaisi Takeuti), which says that $\mathbf{I}\Sigma_1$ is Π_2 -conservative over PRA. In [1] it is shown that this theorem is in fact formalizable in PRA, which gives us (i).

Theorem 5.8 (Parson's Theorem). $\text{PRA} \vdash \forall^{\Pi_2} B (\Box(\sigma \rightarrow B) \rightarrow \Box B)$.

So this certainly holds for $\mathcal{B}(\Sigma_1)$ formulas consisting of Boolean combinations of formulas of the form $\Box^\alpha \perp$. As for (ii), it is shown in [23] that the negation of the sentence axiomatizing $\mathbf{I}\Sigma_1$ is Π_3 -conservative over PRA. That is, we have the following lemma:

Lemma 5.9. $\text{PRA} \vdash \forall^{\Pi_3} B (\Box(\neg\sigma \rightarrow B) \rightarrow \Box B)$.

Thus, we can state the following corollary:

Corollary 5.10. $\mathbf{PL}_{\mathcal{Q}}(\text{PRA}) = \mathcal{L}(\mathfrak{G}_1)$.

While the logic **GL.3** of the linear frame \mathfrak{G}_0 is well known, that of \mathfrak{G}_1 is not. Therefore in the following section we provide a simple axiomatization. Our work can then be generalized to arbitrary \mathfrak{G}_n .

6. THE LOGIC OF \mathfrak{G}_1

6.1. The Modal Logic GL.4 and its corresponding class of frames. We define **GL.4** to be the normal modal logic obtained by adding to **GL** the following two axiom schemata:

- Q1.* $\Box(\Box A \rightarrow (B \vee C)) \vee \Box(\Box^+ B \rightarrow (A \vee C)) \vee \Box(\Box^+ C \rightarrow (A \vee B));$
Q2. $\Diamond(\Diamond A \wedge \Box B) \rightarrow \Box(\Diamond A \vee B).$

GL.4 in fact defines a natural class of frames. We define \mathcal{C} to be the class satisfying the following properties:

- C1.* Finite, irreflexive and transitive;
C2. Non-triple branching: $(xRy \ \& \ xRz \ \& \ xRw) \Rightarrow$
 $(wRy \vee yRw \vee zRw \vee wRz \vee yRz \vee zRy \vee w = y \vee z = y \vee w = z);$
C3. Strongly confluent: $(xRy \ \& \ xRz \ \& \ yRw) \Rightarrow (zRw \vee wRz \vee yRz).$

Theorem 6.1. *GL.4 is sound and complete with respect to \mathcal{C} .*

Soundness is proven as usual by induction on complexity of proofs. As for completeness, we shall appeal to the canonical model of **GL.4** (see Definition 4.18 of [9]). In particular we use the finite filtration method to transform the canonical model into a model in the class \mathcal{C} .

Recall the canonical model \mathfrak{M} of **GL.4** is the triple, $\langle W^{GL.4}, R^{GL.4}, V^{GL.4} \rangle$ with

- $W^{GL.4}$ is the set of maximal **GL.4**-consistent sets;
- For $\Gamma, \Delta \in W^{GL.4}$, define $\Gamma R^{GL.4} \Delta$ if for all $\phi \in \Delta$ we have $\Diamond \phi \in \Gamma$;
- $V(p) = \{\Gamma : p \in \Gamma\}$, for propositional variables p .

First, we make some key observations about this model, the verifications of which are straightforward.

Lemma 6.2. *C2 holds on \mathfrak{M} .*

Lemma 6.3. *C3 holds on \mathfrak{M} .*

In fact, these follow by the fact that axiom *Q1* is *canonical* for property *C2*, as is axiom *Q2* for *C3* (see [9], Definition 4.31). Thus, it remains to show that we can transform the underlying frame of \mathfrak{M} into a finite partial order, while preserving validity of formulas.

Proof of Theorem 6.1. Suppose that **GL.4** $\not\vdash A$, for some formula A . We would like to find a maximal consistent set Γ such that $(\Box A \wedge \neg A) \in \Gamma$, so that Γ is an ‘irreflexive’ point in the canonical model.

By the fact that A is not a theorem, we are guaranteed of some $\Delta \in W^{GL.4}$ such that $A \notin \Delta$. If $\Box A \in \Delta$, then set $\Gamma := \Delta$. Otherwise, since $\neg \Box A \in \Delta$, by the contrapositive form of Löb’s Theorem $\Diamond(\Box A \wedge \neg A) \in \Delta$. Thus by the so-called ‘Existence Lemma’ ([9], Lemma 4.20) for normal modal logics, Δ is $R^{GL.4}$ -related to some Σ for which $(\Box A \wedge \neg A) \in \Sigma$. In that case, set $\Gamma := \Sigma$.

Either way we have some Γ with $(\Box A \wedge \neg A) \in \Gamma$. Notice also, if $\Box C$ is a subformula of A , and $\Box C \notin \Gamma$, then by the same argument there is some ‘irreflexive’

Δ such that $\Gamma R^{GL.4} \Delta$ and that $(\Box C \wedge \neg C) \in \Delta$. Moreover, by Lemma 6.2 there are at most two distinct such Δ .

With these observations in place, our filtrated model $\mathfrak{M}' = \langle W, R, V \rangle$ will be defined as a submodel of \mathfrak{M} :

- (i) $W := \{\Gamma\} \cup \{\Delta : \Gamma R^{GL.4} \Delta, \text{ and there is } \Box C \text{ subsentence of } A, \text{ such that } (\Box C \wedge \neg C) \in \Delta \text{ and } \neg \Box C \in \Gamma\}$;
- (ii) R is just $R^{GL.4}$ restricted to points in W ;
- (iii) $V(p) := V^{GL.4}(p) \cap W$.

The model \mathfrak{M}' satisfies $C2$, $C3$, and transitivity simply because \mathfrak{M} does. It is clearly finite. And irreflexivity, as hinted above, follows from the fact that each point in W was chosen to contain some formulas $\Box C$ and $\neg C$, ensuring the point is not related to itself. It follows \mathfrak{M}' is in \mathcal{C} .

The standard ‘Truth Lemma’ is then proven by induction:

Lemma 6.4. *If $\Delta \in W$ and B is a subsentence of A , then $B \in \Delta$ iff $\mathfrak{M}', \Delta \models B$.*

Concluding the proof, since $A \notin \Gamma$, we have that $\mathfrak{M}', \Gamma \not\models A$. \square

6.2. The Class \mathcal{C} and the Frame \mathfrak{G}_1 . We must now show that **GL.4** is the logic of the frame \mathfrak{G}_1 .

Recall a p -morphism from $\mathbb{F} = \langle W, R \rangle$ to $\mathbb{F}' = \langle W', R' \rangle$ is a function $f : W \rightarrow W'$, such that xRy implies $f(x)R'f(y)$; and if $f(x)R'y'$ then there is some $y \in W$ such that $f(y) = y'$ and xRy . The following theorem is standard:⁹

Theorem 6.5. *If there is a p -morphism from \mathbb{F} to \mathbb{F}' , then the existence of a valuation V' and point $w' \in W'$ such that $\langle \mathbb{F}', V' \rangle, w' \not\models A$, ensures the existence of a valuation V and point $w \in W$, such that $\langle \mathbb{F}, V \rangle, w \not\models A$.*

To demonstrate that **GL.4** is the logic of \mathfrak{G}_1 , we use the following proposition:

Proposition 6.6. *For any frame $\mathbb{F} \in \mathcal{C}$ and any point x in \mathbb{F} , there is some point $\langle m, i \rangle$ in \mathfrak{G}_1 , such that there exists a p -morphism from the subframe generated by $\langle m, i \rangle$ to the subframe generated by x .*

In other words, falsifiability is reflected by p -morphisms, which gives us the following corollary of Proposition 6.6 and improvement upon Corollary 5.10.

Corollary 6.7. $\mathbf{PL}_Q(\mathbf{PRA}) = \mathbf{GL.4}$.

It remains only to verify Proposition 6.6.

Proof Sketch of Proposition 6.6. The proof proceeds by induction on the number of points in a frame in \mathcal{C} . The basic case is obvious. Supposing we have a frame with one point, say x , then consider the subframe generated by $\langle 0, 0 \rangle$, and the p -morphism mapping $\langle 0, 0 \rangle$ to x .

Supposing we have a frame in \mathcal{C} with $n + 1$ points, consider the subframe $\mathbb{F} = \langle W, R \rangle$ generated by some point $x \in C$. We would like to use the inductive hypothesis to obtain a p -morphism to some subframe of \mathbb{F} containing $\leq n$ points, and extend it to all of \mathbb{F} . To do this we consider three cases: (i) x has no successors; (ii) x has one immediate successor (i.e. point y such that xRy and there is no z with $xRzRy$); and (iii) x has two immediate successors. More than 2 immediate successors is ruled out by property $C2$.

⁹See, e.g. [9], Definition 3.13, where p -morphisms go under the name *bounded morphism*.

Case (i) is trivial. For case (ii), let \mathbb{F}' be \mathbb{F} without the point x , and let y be the unique immediate successor of x . Then since $\mathbb{F}' \in \mathcal{C}$ and it has n points, we have a p -morphism f from the subframe generated by some point $\langle m, i \rangle$ in \mathfrak{G}_1 to \mathbb{F}' , the subframe generated by y . We then consider the subframe generated by $\langle m+1, i \rangle$ instead, and extend the p -morphism f so that $f(\langle m+1, i \rangle) = x$ and $f(\langle m, i-1 \rangle) = y$.

Verifying case (iii) is similar, except that instead of removing the point x , we must remove the ‘maximal’ points of \mathbb{F} . Then the p -morphism obtained by inductive hypothesis is extended by shifting each point in the morphism by one. Thus, e.g. if $\langle m, i \rangle$ is mapped to y , then in the new mapping $\langle m+1, i \rangle$ is mapped to y . And we let $f(\langle 0, 0 \rangle) = f(\langle 0, 1 \rangle) = x$. The details are straightforward and are left to the reader (or can be found in [17]). \square

Remark 6.8. *The methods in this section carry over to the general case of frames \mathfrak{G}_n for arbitrary n . By an analogous argument, one can prove the logic is simply Q2 (strong confluence) and the axiom corresponding to “non- $n+2$ -ary-branching”, which is just a generalization¹⁰ of non-branching and non-triple-branching:*

$$\bigvee_{i \leq n+1} \Box(\Box^+ A_i \rightarrow \bigvee_{i \neq j} A_j).$$

7. ON THE PROOF OF SOLOVAY’S THEOREM

In Sections 3 and 4 we showed that $\mathbf{PL}_{\mathcal{F}}(T) = \mathbf{GL.3}$ for a wide range of arithmetical theories T and fragments \mathcal{F} . Otherwise put, $\mathbf{PL}_{\mathcal{F}}(T)$ gives us the logic of non-branching \mathbf{GL} -frames. *Prima facie*, one might imagine the possibility of strategically adding sentences into the fragment \mathcal{F} (where \mathcal{F} is, e.g. \mathcal{B}), so as to obtain the logic of non-triple-branching \mathbf{GL} -frames, then that of non-quadruple-branching \mathbf{GL} -frames, and so on. Assuming this could be generalized it would be possible to define an infinite fragment \mathcal{H} , for which $\mathbf{PL}_{\mathcal{H}}(T) = \mathbf{GL}$. At that point, to the extent that Solovay’s Theorem is not already assumed in the determination of \mathcal{H} , we would have a new proof of the result. After all, any non-theorem of \mathbf{GL} can be falsified on some finite, and thus finitely branching, frame. So the witnessing realization would make use of some finite subset of the fragment, sufficient to falsify the formula.

What we have shown is that the first step in this process is (almost) possible, *vis-à-vis* Corollary 5.10. Adding the constant for $\mathbf{I}\Sigma_1$ and capitalizing on the well studied relationship between that theory and PRA, we are able to obtain the logic of non-triple-branching (and strongly confluent) \mathbf{GL} -frames. Two important questions remain, however, before taking the next step.

The first and most obvious question is what the further constants will be. The particular case of $\mathbf{I}\Sigma_1$ and PRA is already well studied. Going beyond that may require some significant arithmetical investigation. In Section 5.2 we isolated what arithmetical facts are sufficient to hold. So on the proposed strategy it would simply be a matter of finding a theory and a fragment that satisfy these requirements.

The second, and more curious, question is how to dispense with property C3, strong confluence. We have seen that the logic of the frame \mathbf{G}_n always contains the formula Q2, and so it will clearly remain in the limit. However Q2 is obviously

¹⁰It is not hard to see that $\Box(\Box A \rightarrow B) \vee \Box(\Box^+ B \rightarrow A)$ is equivalent to $\Box(\Box^+ A \rightarrow B) \vee \Box(\Box^+ B \rightarrow A)$ over \mathbf{GL} .

not a theorem of **GL**. Finding constants whose associated provability logics do not validate $Q2$ may prove a challenge. Understanding this situation may shed light on Solovay's original proof.

8. INTERPRETABILITY LOGICS WITH RESTRICTED SUBSTITUTIONS

Interpretations are used throughout mathematics and logic. Loosely speaking, an interpretation from a theory V into a theory U is structure preserving map that translates theorems of V to theorems of U . The notion of interpretability that we discuss below is *grosso modo* that of [33] and details can be found in, e.g. [20] or in [36].

8.1. Interpretability Logics. Interpretability can be seen as a generalization of provability. By $\alpha \triangleright_T \beta$ we denote a natural formalized version of the statement that $T + \beta$ is interpretable in $T + \alpha$.

Interpretability Logics are designed to capture the structural behavior of formalized interpretability. The language of these logics is that of provability logic together with a binary modality \triangleright , orthographically identical to the arithmetical operator, to model formalized interpretability. And indeed, *arithmetical realizations* are extended as expected by imposing that

$$(A \triangleright B)^* = A^* \triangleright B^*.$$

For a clear distinction, let $\mathbf{Form}_{\mathbf{IL}}$ denote the class of modal formulas in language of interpretability logic and $\mathbf{Form}_{\mathbf{GL}}$ the standard modal language of basic provability logic. In analogy with the definition of $\mathbf{PL}(T)$ we define $\mathbf{IL}(T)$, the interpretability logic of a theory T

$$\mathbf{IL}(T) := \{A \in \mathbf{Form}_{\mathbf{IL}} \mid \forall * T \vdash A^*\} \quad \text{and} \quad \mathbf{IL}_{\Gamma}(T) := \{A \in \mathbf{Form}_{\mathbf{IL}} \mid \forall * \in \Gamma T \vdash A^*\}.$$

By Theorem 1.2 and Footnote 2 we see that provability logics are the same for all sufficiently strong theories. This is certainly not the case for interpretability logics, which turn out to be more sensitive to differences between theories. One such example is the notion of an *essential reflexive* theory.

A theory is *reflexive* if it proves the consistency of any finite subpart of it. A theory is essentially reflexive whenever any finite extension of it is reflexive. The following theorem is due independently to A. Berarducci and V. Shavrukov. The definition of \mathbf{ILM} will follow below.

Theorem 8.1 (Berarducci [6], Shavrukov [30]). *If T is an essentially reflexive theory, then $\mathbf{IL}(T) = \mathbf{ILM}$.*

However, if a theory is finitely axiomatizable we get a different outcome where, again, \mathbf{ILP} is defined below.

Theorem 8.2 (Visser [35]). *If T is finitely axiomatizable, then $\mathbf{IL}(T) = \mathbf{ILP}$.*

A prominent problem in formalized interpretability is to determine the maximal interpretability logic that is contained in any reasonable arithmetical theory.

Definition 8.3. *The interpretability logic of all reasonable arithmetical theories, written $\mathbf{IL}(\text{All})$, is the set of formulas φ such that for all T and $*$, $T \vdash \varphi^*$. Here we let T range over all reasonable¹¹ arithmetical theories.*

¹¹The boundaries are not exactly determined and will depend a bit on the answer. It is legitimate to think of any theory extending $\mathbf{I}\Delta_0 + \text{exp}$.

Clearly, $\mathbf{IL}(\text{All})$ is in the intersection of \mathbf{ILM} and \mathbf{ILP} but apparently it possesses a very rich structure (see [25], and [12]). In this paper, it is only important to know that a certain very weak logic to be defined below is part of $\mathbf{IL}(\text{PRA})$.

Fact 8.4. $\mathbf{ILW} \subset \mathbf{IL}(\text{PRA})$

For most theories that do not fall under Theorems 8.1 and 8.2, the interpretability logic is unknown. The theory PRA is a notable example: the logic $\mathbf{IL}(\text{PRA})$ is still unknown. The most recent results for $\mathbf{IL}(\text{PRA})$ are presented in [8].

PRA is known to be the same as $\mathbf{I}\Sigma_1^R$ where $\mathbf{I}\Sigma_n^R$ is defined as $\mathbf{I}\Delta_0 + \text{exp}$ plus the Σ_n induction rule. See for example [2]. In that paper a proof can also be found for the following theorem.

Theorem 8.5. $\mathbf{I}\Sigma_n^R$ is reflexive, as is any extension of $\mathbf{I}\Sigma_n^R$ by Σ_{n+1} formulas.

The logical complexity of interpretability is Σ_3 and in [31] it is shown that it is essentially so. However, by a theorem due to Orey and Hájek we can often reduce the Σ_3 notion of interpretability to the Π_2 notion of Π_1 -conservativity. A theory V is Π_1 -conservative over U , we write $U \triangleright_{\Pi_1} V$, whenever for all Π_1 sentences π we have that $[V \vdash \pi \text{ implies } U \vdash \pi]$.

Theorem 8.6 (Orey-Hájek). *For reflexive theories U and V we have*

$$(U \triangleright V) \quad \Leftrightarrow \quad (U \triangleright_{\Pi_1} V)$$

and this equivalence is provable in EA.

One advantage of this characterization is evidently that the logical complexity of Π_1 -conservativity is lower than that of interpretability. Another advantage is that the so-called Π_1 -conservativity logic is a relatively stable notion. The Π_1 -conservativity logic of a theory T is just the set of modal formulas in $\text{Form}_{\mathbf{IL}}$ that are provable in T under any arithmetical realization where the \triangleright modality is mapped to \triangleright_{Π_1} .

Theorem 8.7. *For any sound theory T extending \mathbf{III}_1^- we have that the Π_1 -conservativity logic of T is \mathbf{ILM} .*

The theorem was first proven by Hájek and Montagna in [13] and [14] to hold for any sound theory containing $\mathbf{I}\Sigma_1$. Beklemishev and Visser lowered the threshold to the rather weak theory \mathbf{III}_1^- that allows only induction for parameter free formulas of complexity Π_1 . It is well known that PRA extends \mathbf{III}_1^- ([2]).

Remark 8.8. *The proof of Theorem 8.7 is rather similar to that of Solovay's original proof and again (see Footnote 4), the substitutions in the completeness proof can be taken¹² to be Σ_2 .*

The logics \mathbf{ILM} and \mathbf{ILP} have elegant syntactical presentations. We shall define them in parts. First, we define a logic \mathbf{IL} that is present to all interpretability logics studied. Next this logic \mathbf{IL} is extended by adding more axiom schemata.

(When we write formulas in $\text{Form}_{\mathbf{IL}}$ we adhere to the following binding conventions. We say that \triangleright binds stronger than \rightarrow but weaker than all other connectives. Using this convention we can save a lot of brackets.)

¹²Albert Visser (p.c.) notes that close inspection of the proof actually reveals that the substitutions can be taken to be $\Delta_2(\mathbf{III}_1^-)$. That is, a Σ_2 sentences that is probably in \mathbf{III}_1^- equivalent to a Π_2 sentence.

Definition 8.9. *The logic \mathbf{IL} is the smallest set of formulas being closed under the rules of Necessitation and of Modus Ponens, that contains all tautological formulas and all instantiations of the following axiom schemata.*

- L1 $\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B)$
- L2 $\Box A \rightarrow \Box \Box A$
- L3 $\Box(\Box A \rightarrow A) \rightarrow \Box A$
- J1 $\Box(A \rightarrow B) \rightarrow A \triangleright B$
- J2 $(A \triangleright B) \wedge (B \triangleright C) \rightarrow A \triangleright C$
- J3 $(A \triangleright C) \wedge (B \triangleright C) \rightarrow A \vee B \triangleright C$
- J4 $A \triangleright B \rightarrow (\Diamond A \rightarrow \Diamond B)$
- J5 $\Diamond A \triangleright A$

Apart from the axiom schemata enumerated in Definition 8.9 we will need consider other axiom schemata too.

- M $A \triangleright B \rightarrow A \wedge \Box C \triangleright B \wedge \Box C$
- P $A \triangleright B \rightarrow \Box(A \triangleright B)$
- W $A \triangleright B \rightarrow A \triangleright B \wedge \Box \neg A$

If X is a set of axiom schemata we will denote by \mathbf{ILX} the logic that arises by adding the axiom schemata in X to \mathbf{IL} .

8.2. The closed fragment. Because closed formulas in \mathbf{ILW} can be reduced to those of \mathbf{GL} ([15]) we can prove that $\mathbf{IL}_{\mathcal{B}}(\text{PRA})$ is again the logic of linear frames.

Definition 8.10. *The logic $\mathbf{ILW.3}$ is obtained by adding the linearity axiom schema $\Box(\Box A \rightarrow B) \vee \Box(\Box B \rightarrow A)$ to \mathbf{ILW} .*

Theorem 8.11. $\mathbf{IL}_{\mathcal{B}}(\text{PRA}) = \mathbf{ILW.3}$

Proof. We give a translation from formulas φ in $\text{Form}_{\mathbf{IL}}$ to formulas φ^{tr} in $\text{Form}_{\mathbf{GL}}$ such that

$$\begin{aligned} \mathbf{ILW.3} \vdash \varphi &\Leftrightarrow \mathbf{GL.3} \vdash \varphi^{\text{tr}} \quad (*) \\ &\text{and} \\ \mathbf{ILW.3} \vdash \varphi &\Leftrightarrow \varphi^{\text{tr}}. \quad (**) \end{aligned}$$

If we moreover know (***) : $\mathbf{ILW.3} \vdash \varphi \Rightarrow \forall * \in \mathcal{B} \text{ PRA} \vdash \varphi^*$ we would be done. For then we have by (**) and (***) that

$$\forall * \in \text{Sub}(\mathcal{B}) \text{ PRA} \vdash \varphi^* \Leftrightarrow (\varphi^{\text{tr}})^*$$

and consequently

$$\begin{aligned} \forall * \in \mathcal{B} \text{ PRA} \vdash \varphi^* &\Leftrightarrow \\ \forall * \in \mathcal{B} \text{ PRA} \vdash (\varphi^{\text{tr}})^* &\Leftrightarrow \\ \mathbf{GL.3} \vdash \varphi^{\text{tr}} &\Leftrightarrow \\ \mathbf{ILW.3} \vdash \varphi. & \end{aligned}$$

We first see that (***) holds. Certainly, by Fact 8.4, we have that $\mathbf{ILW} \subseteq \mathbf{IL}_{\mathcal{B}}(\text{PRA})$. Thus it remains to show that $\text{PRA} \vdash \Box(\Box A^* \rightarrow B^*) \vee \Box(\Box B^* \rightarrow A^*)$ for any formulas A and B in $\text{Form}_{\mathbf{IL}}$ and any $* \in \mathcal{B}$. As any formula in the closed fragment of \mathbf{ILW} is equivalent to a formula in the closed fragment of \mathbf{GL} (see [15]), Theorem 3.2 gives us that indeed the linearity axiom holds for the closed fragment of \mathbf{GL} .

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

$$(A \triangleright B)^{\text{tr}} := \Box(A^{\text{tr}} \rightarrow (B^{\text{tr}} \vee \Diamond B^{\text{tr}})).$$

We first see that we have (**). It is sufficient to show that $\mathbf{ILW.3} \vdash p \triangleright q \rightarrow \Box(p \rightarrow (q \vee \Diamond q))$. We reason in $\mathbf{ILW.3}$. An instantiation of the linearity axiom gives us $\Box(\Box\neg q \rightarrow (\neg p \vee q)) \vee \Box((\neg p \vee q) \wedge \Box(\neg p \vee q) \rightarrow \neg q)$. The first disjunct immediately yields $\Box(p \rightarrow (q \vee \Diamond q))$.

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

We now prove (*). By induction on $\mathbf{ILW.3} \vdash \varphi$ we see that $\mathbf{GL.3} \vdash \varphi^{\text{tr}}$. All the specific interpretability axioms turn out to be provable under our translation in \mathbf{GL} . The only axioms where the $\Box A \rightarrow \Box\Box A$ axiom scheme is really used is in \mathbf{J}_2 and \mathbf{J}_4 . To prove the translation of \mathbf{W} we also need \mathbf{L}_3 .

If $\mathbf{GL.3} \vdash \varphi^{\text{tr}}$ then certainly $\mathbf{ILW.3} \vdash \varphi^{\text{tr}}$ and by (**), $\mathbf{ILW.3} \vdash \varphi$. \square

We thus see that $\mathbf{ILW.3}$ is an upperbound for $\mathbf{IL}(\text{PRA})$. Using the translation from the proof of Theorem 8.11, it is not hard to see that both the principles \mathbf{P} and \mathbf{M} are provable in $\mathbf{ILW.3}$. This tells us that the upperbound is actually not very informative as we know that $\mathbf{IL}(\text{PRA}) \not\vdash \mathbf{M}$. By a straight-forward generalization of Lemma 1.4 we see that choosing larger Γ will generally yield a smaller $\mathbf{IL}_\Gamma(\text{PRA})$ and thus a sharper upperbound. Subsection 8.4 consists of reflections on just how large the Γ should be as to refute \mathbf{M} in $\mathbf{IL}_\Gamma(\text{PRA})$. First we shall include some observations on a fragment slightly larger than the closed fragment.

8.3. The closed fragment with a constant for $\mathbf{I}\Sigma_1$. If we consider the proof of Theorem 2.1, we see that it does not make any assumptions on the signature of the modal logic under considerations. In particular, the theorem still holds for interpretability logics. In the theorem below we use this to give a semantic characterization of $\mathbf{IL}_{\mathcal{F}_1}(\text{PRA})$.

In [24] it is established that for a certain frame, that we will denote here by $\widetilde{\mathfrak{G}}_1^\bullet$, we have the following equivalence.

$$\forall A \in \mathcal{F}_1 [\widetilde{\mathfrak{G}}_1^\bullet \models A \Leftrightarrow \text{PRA} \vdash A] \quad (\dagger)$$

For the purpose of this paper it is not material to know what exactly the frame $\widetilde{\mathfrak{G}}_1^\bullet$ looks like and we shall refrain from giving a formal definition. It is only important to know that $\widetilde{\mathfrak{G}}_1^\bullet$ is just \mathfrak{G}_1^\bullet with some additional accessibility relations to model the \triangleright modality. This, together with the mere equivalence (\dagger) suffices to obtain the following theorem.

Theorem 8.12. $\mathbf{IL}_{\mathcal{F}_1}(\text{PRA}) = \mathcal{L}(\widetilde{\mathfrak{G}}_1^\bullet)$

Proof. Image-finiteness and definability of separate points is clear as interpretability logic is an extension of provability logic. Thus, by Theorem 2.1 we obtain the result. \square

In [24], also a logic \mathbf{PIL} is given such that we actually have

$$\forall A \in \mathcal{F}_1 [\widetilde{\mathfrak{G}}_1^\bullet \models A \Leftrightarrow \text{PRA} \vdash A \Leftrightarrow \mathbf{PIL} \vdash A].$$

This suggests that the following conjecture should not be too hard to prove. In this conjecture, **ILM.4** denotes the logic that arises by joining **ILM** and **GL.4**.

Conjecture 8.13. $\mathcal{L}(\widetilde{\mathfrak{G}}_1^\circ) = \mathbf{ILM.4}$

The inclusion $\mathcal{L}(\widetilde{\mathfrak{G}}_1^\circ) \supseteq \mathbf{ILM.4}$ is actually very easy and follows from a direct verification of the validity of the axioms on $\widetilde{\mathfrak{G}}_1^\circ$. The other direction is harder but not too interesting as we still have $\mathbf{M} \in \mathbf{IL}_{\mathcal{F}_1}(\mathbf{PRA})$.

8.4. Fragments for refuting \mathbf{M} in $\mathbf{IL}_\Gamma(\mathbf{PRA})$. In [36] it is shown that $\mathbf{IL}(\mathbf{PRA}) \not\vdash A \triangleright \diamond B \rightarrow \square(A \triangleright \diamond B)$. It is easy to see that $\mathbf{ILM} \vdash A \triangleright \diamond B \rightarrow \square(A \triangleright \diamond B)$. This implies that \mathbf{M} is certainly not derivable in $\mathbf{IL}(\mathbf{PRA})$. We can also find explicit realizations that violate \mathbf{M} , as the following lemma tells us.

Lemma 8.14. *For $n \geq 1$, we have that $\mathbf{IL}(\mathbf{I}\Sigma_n^R) \not\vdash \mathbf{M}$.*

Proof. We define a realization $*$ such that $\mathbf{I}\Sigma_n^R \not\vdash (p \triangleright q \rightarrow p \wedge \square r \triangleright q \wedge \square r)^*$.

It is well-known that $\mathbf{I}\Sigma_n^R \subsetneq \mathbf{I}\Sigma_n \subsetneq \mathbf{I}\Sigma_{n+1}^R$ and that, for every $n \geq 1$, $\mathbf{I}\Sigma_n$ is finitely axiomatized. Let σ_n be the single sentence axiomatizing $\mathbf{I}\Sigma_n$. It is also known that (for $n \geq 1$) $\mathbf{IL}(\mathbf{I}\Sigma_n) = \mathbf{ILP}$ and that $\mathbf{ILP} \not\vdash p \triangleright q \rightarrow p \wedge \square r \triangleright q \wedge \square r$. Thus, for any $n \geq 1$ we can find α_n, β_n and γ_n such that

$$\mathbf{I}\Sigma_n \not\vdash \alpha_n \triangleright \beta_n \rightarrow \alpha_n \wedge \square \gamma_n \triangleright \beta_n \wedge \square \gamma_n.$$

Note that

$$\mathbf{EA} \vdash \alpha_n \triangleright_{\mathbf{I}\Sigma_n} \beta_n \leftrightarrow \sigma_n \wedge \alpha_n \triangleright_{\mathbf{I}\Sigma_n^R} \sigma_n \wedge \beta_n$$

and

$$\mathbf{EA} \vdash \square_{\mathbf{I}\Sigma_n} \gamma_n \leftrightarrow \square_{\mathbf{I}\Sigma_n^R} (\sigma_n \rightarrow \gamma_n).$$

Thus, we have

$$\mathbf{I}\Sigma_n^R \not\vdash \sigma_n \wedge \alpha_n \triangleright \sigma_n \wedge \beta_n \rightarrow \sigma_n \wedge \alpha_n \wedge \square (\sigma_n \rightarrow \gamma_n) \triangleright \sigma_n \wedge \beta_n \wedge \square (\sigma_n \rightarrow \gamma_n)$$

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

We see that the realizations used in the proof of Lemma 8.14 get higher and higher complexities. The complexity is certainly higher than Σ_2 .

By Theorem 1 from [8] (Theorem 12.1.1 from [23]) we know that for $\alpha, \beta \in \Sigma_2$ we have

$$\mathbf{PRA} \vdash (\alpha \triangleright \beta) \rightarrow ((\alpha \wedge \square \gamma) \triangleright (\beta \wedge \square \gamma))$$

for any sentence γ . This translates to $\mathbf{IL}_{\Sigma_2}(\mathbf{PRA}) \vdash \mathbf{M}$ and indicates that an arithmetical completeness proof for $\mathbf{IL}(\mathbf{PRA})$ can not work with only Σ_2 -realizations.

For $\mathbf{I}\Sigma_n^R$, $n \geq 2$ we know that $\mathbf{IL}(\mathbf{I}\Sigma_n^R) \subset \mathbf{ILM}$. This follows from the next lemma.

Lemma 8.15. $\mathbf{IL}_{\Sigma_2}(\mathbf{I}\Sigma_n^R) = \mathbf{IL}_{\Delta_{n+1}}(\mathbf{I}\Sigma_n^R) = \mathbf{ILM}$ whenever $n \geq 2$.

Proof. We shall use that the logic of Π_1 -conservativity for theories containing \mathbf{III}_1^- is \mathbf{ILM} as mentioned in Theorem 8.7.

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

First, we prove by induction on the complexity of a modal formula A that for all $* \in \Delta_{n+1}$ $\mathbf{I}\Sigma_n^R \vdash A_{\Pi_1}^* \leftrightarrow A_{\triangleright}^*$ 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 $A = (B \triangleright C)$. We reason in $\mathbf{I}\Sigma_n^R$:

$$\begin{array}{ll}
 (B \triangleright C)_{\triangleright}^* & \leftrightarrow \text{def.} \\
 \mathbb{I}\Sigma_n^R + B_{\triangleright}^* \triangleright \mathbb{I}\Sigma_n^R + C_{\triangleright}^* & \leftrightarrow \text{i.h.} \\
 \mathbb{I}\Sigma_n^R + B_{\Pi_1}^* \triangleright \mathbb{I}\Sigma_n^R + C_{\Pi_1}^* & \leftrightarrow \text{Orey-Hájek} \\
 \mathbb{I}\Sigma_n^R + B_{\Pi_1}^* \triangleright_{\Pi_1} \mathbb{I}\Sigma_n^R + C_{\Pi_1}^* & \leftrightarrow \text{def.} \\
 (B \triangleright C)_{\Pi_1}^* &
 \end{array}$$

Note that we have access to the Orey-Hájek characterization as $B_{\Pi_1}^*$ is at most of complexity Δ_{n+1} and thus $\mathbb{I}\Sigma_n^R + B_{\Pi_1}^*$ is a reflexive theory by Theorem 8.5. Also note that $(B \triangleright C)_{\Pi_1}^*$ is a Π_2 -sentence and thus certainly Δ_{n+1} whenever $n \geq 2$.

If now $\mathbf{ILM} \vdash A$ then $\mathbb{I}\Sigma_n^R \vdash A_{\Pi_1}^*$ and thus whenever $* \in \Delta_{n+1}$, $\mathbb{I}\Sigma_n^R \vdash A_{\triangleright}^*$ and $\mathbf{ILM} \subseteq \mathbf{IL}_{\Delta_{n+1}}(\mathbb{I}\Sigma_n^R)$.

If $\mathbf{ILM} \not\vdash A$ then by Remark 8.8 for some $* \in \Sigma_2$ we have $\mathbb{I}\Sigma_n^R \not\vdash A_{\Pi_1}^*$ whence $\mathbb{I}\Sigma_n^R \not\vdash A_{\triangleright}^*$. We may conclude that $\mathbf{IL}_{\Sigma_2}(\mathbb{I}\Sigma_n^R) \subseteq \mathbf{ILM}$. \square

Theorem 8.16. $\mathbf{IL}(\text{PRA}) \subset \mathbf{ILM}$

Proof. Although the proof of Lemma 8.15 does not give us that $\mathbf{IL}_{\Sigma_2}(\mathbb{I}\Sigma_1^R) = \mathbf{ILM}$, it does give us that $\mathbf{IL}_{\Sigma_2}(\mathbb{I}\Sigma_1^R) \subseteq \mathbf{ILM}$. By earlier observations we saw that $\mathbf{IL}(\text{PRA}) \neq \mathbf{ILM}$. \square

9. FUTURE RESEARCH

We have seen that adding a constant for $\mathbb{I}\Sigma_1$ to PRA is sufficient to obtain a non-trivial provability logic. By a Theorem of Leivant it is known that $\mathbb{I}\Sigma_1 \equiv < 2 >_{\text{EA}} \top$. An interesting fragment to consider next for PRA would be the closed fragment together with the set of constants

$$\{(\langle 1 \rangle_{\text{EA}} < 2 \rangle_{\text{EA}})^n \top \mid n \in \omega\}$$

or variants thereof.

10. ACKNOWLEDGEMENTS

We would like to thank Lev Beklemishev, Dick de Jongh and Albert Visser for fruitful comments and discussions.

REFERENCES

- [1] L.D. Beklemishev. Bimodal logics for extensions of arithmetical theories. *Journal of Symbolic Logic*, 61(1):91-124.
- [2] L.D. Beklemishev. Reflection principles and provability algebras in formal arithmetic. *Russian Mathematical Surveys* 60(2): 197-268 2005.
- [3] L.D. Beklemishev. Kripke semantics for provability logic **GLP**. *Annals of Pure and Applied Logic*, In press.
- [4] L.D. Beklemishev, G. Bezhanishvili, and T. Icard. On topological semantics of **GLP**. Forthcoming in R. Schindler, editor, *Ways of Proof Theory*, Ontos Verlag, 2010.
- [5] L.D. Beklemishev, J.J. Joosten, and M. Vervoort. A Finitary Treatment of the Closed Fragment of Japaridze's Provability Logic. *Journal of Logic and Computation*, 15(4):447-463, 2005.
- [6] A. Berarducci. The interpretability logic of Peano arithmetic. *Journal of Symbolic Logic*, 55:1059-1089, 1990.
- [7] A. Berarducci, and R. Verbrugge. On the provability logic of bounded arithmetic. *Annals of Pure and Applied Logic*, 61(1-2): 75-93, 1993.
- [8] M. Bilkova, D. de Jongh, J. J. Joosten. Interpretability in PRA. *Annals of Pure and Applied Logic*, 161: 128138, 2009.

- [9] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*, Cambridge University Press, Cambridge, 2001.
- [10] G. Boolos. *The Logic of Provability*. Cambridge University Press, ISBN 0-521-43342-8, 1993.
- [11] S. Buss, *Bounded Arithmetic*, PhD Thesis. Bibliopolis, Napoli, 1986.
- [12] E. Goris, and J. J. Joosten. A new principle in the interpretability logic of all reasonable arithmetical theories. *Logic Journal of the IGPL*, accepted for publication, 2010.
- [13] P. Hájek, and F. Montagna. The logic of Π_1 -conservativity. *Archiv für Mathematische Logik und Grundlagenforschung*, 30: 113-123, 1990.
- [14] P. Hájek, and F. Montagna. The logic of Π_1 -conservativity continued. *Archiv für Mathematische Logik und Grundlagenforschung*, 32: 57-63, 1992.
- [15] P. Hajek, and V. Švejdar. A note on the normal form of closed formulas of interpretability logic. *Studia Logica*, 50: 25-38, 1991.
- [16] T. Icard. A topological study of the closed fragment of **GLP**. *Journal of Logic and Computation*, In press.
- [17] T. Icard. Towards an Alternative Proof of Solovay's Arithmetical Completeness Theorem. In *Proceedings of the 12th ESSLLI Student Session*, Dublin, 2007.
- [18] K. Ignatiev. On Strong Provability predicates and the associated modal logics. *Journal of Symbolic Logic*, 58:249-290, 1993.
- [19] G. Japaridze. *Modal Logical Means of Investigation of Provability*. Ph.D. thesis (in Russian), Moscow State University, 1986.
- [20] J. Japaridze, and D. de Jongh. The Logic of Provability. In S. R. Buss, editor, *Handbook of Proof Theory*. Studies in Logic and the Foundations of Mathematics, 137: 475-546. Elsevier, Amsterdam, 1998.
- [21] D. de Jongh, M. Jumelet, and F. Montagna. On the Proof of Solovay's Theorem. *Studia Logica* 50(1): 51-69, 1991.
- [22] J. J. Joosten. Formalized Interpretability in Primitive Recursive Arithmetic. *Proceedings of the ESSLLI student session*, Vienna, 2003.
- [23] J.J. Joosten. *Intepretability Formalized*, Ph.D. thesis, Department of Philosophy, University of Utrecht, 2004.
- [24] J.J. Joosten. The closed fragment of the intepretability logic of PRA with a constant for IS_1 . *Notre Dame Journal of Formal Logic*, 46(2):127-146, 2005.
- [25] J.J. Joosten and A. Visser. The interpretability logic of all reasonable arithmetical theories. *Erkenntnis*, 53(1-2):3-26, 2000.
- [26] M. Löb. Solution of a problem of Leon Henkin. *Journal of Symbolic Logic*, 20:115-118, 1955.
- [27] G. Mints. Quantifier-free and one-quantifier systems. *Journal of Soviet Mathematics*, 1:71-84, 1972.
- [28] C. Parsons. On n -quantifier induction. *Journal of Symbolic Logic*, 37:466-482, 1972.
- [29] G. Sambin. An effective fixed point theorem in intuitionistic diagonalizable algebras. *Studia Logica* 35: 345 361, 1975.
- [30] V. Shavrukov. The logic of relative interpretability over Peano arithmetic (in Russian). Technical Report Report No.5, Steklov Mathematical Institute, Moscow, 1988.
- [31] V. Y. Shavrukov. Interpreting reflexive theories in finitely many axioms. *Fundamenta Mathematicae*, 152:99-116,1997.
- [32] R. M. Solovay. Provability interpretations of modal logic. *Israel Journal of Mathematics*, 28: 33-71, 1976.
- [33] A. Tarski, A. Mostowski, and R.M. Robinson. *Undecidable Theories*. North Holland Publishing Company, 1953.
- [34] A. Visser. A propositional logic with explicit fixed points. *Studia Logica*, 40:155-175, 1981.
- [35] A. Visser. Interpretability logic. In P.P. Petkov, editor. *Mathematical logic, Proceedings of the Heyting 1988 summer school in Varna, Bulgaria*. Plenum Press, Boston: 175-209, 1990.
- [36] A. Visser, An overview of interpretability logic. In M. Kracht, M. de Rijke and H. Wansing, editors, *Advances in Modal Logic '96*: 307-359. CSLI Publications, Stanford, CA, 1997.

STANFORD UNIVERSITY
E-mail address: `icard@stanford.edu`

UNIVERSIDAD DE SEVILLA
E-mail address: `jjoosten@us.es`