Notre Dame Journal of Formal Logic
Volume 53, Number 2, 2012

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 study provability logics with restricted realizations and show that for various natural candidates of T and restriction set Γ, 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 $\mathsf{I}\Sigma_1$. We then study interpretability logics, obtaining upper bounds for $\mathbf{IL}$(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 nontrivial result that $\mathbf{IL}(\text{PRA}) \subset \mathbf{ILM}$.

1 Introduction

The cornerstone result of provability logic is Solovay's Arithmetical Completeness Theorem, which provides an exact modal characterization of the standard provability predicate in arithmetic. One of the most outstanding problems in the area is to provide an alternative proof of this theorem, both to shed light on Solovay's original proof and to provide ideas for how to obtain completeness theorems for fragments of arithmetic too weak for Solovay's proof method to work.

This paper deals with restricted cases of Solovay's Theorem where alternative proof methods are available. One underlying motivation for the current work is to make progress toward an alternative completeness proof (see Section 7). However, the method of provability logics with restricted realizations, which we introduce

Received February 24, 2010; accepted June 29, 2011; printed May 4, 2012
2010 Mathematics Subject Classification: Primary 03F45, 03B45
Keywords: provability logic, interpretability logic, restricted substitutions, GLP, $\mathsf{I}\Sigma_1$, PRA
 10.1215/00294527-1715653

here, merits interest in its own right, as we shall explain shortly. Let us first briefly restate Solovay's Theorem and the necessary background on provability logic.

1.1 Provability logics The propositional modal logic **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 obtained by extending the basic modal logic **K** with a schematic formalization of Löb's Theorem (**L** in the following definition).

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

$$\begin{array}{lll} \mathbf{K} & : & \Box(A \to B) \to (\Box A \to \Box B); \\ \mathbf{L} & : & \Box(\Box A \to A) \to \Box A. \end{array}$$

The logic is closed under modus ponens and necessitation, and modal schema **4** is derivable: $\Box A \to \Box\Box A$.

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 $\bot$ to $0 = 1$. A realization $*$ can be extended uniformly so that we can interpret an arbitrary modal formula as an arithmetical formula by stipulating

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

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

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 [27] it is known that **GL** is sound for a large class of theories T; that is, $\mathbf{GL} \subseteq \mathbf{PL}(\mathsf{T})$. 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.

For soundness, that is, that $\mathbf{GL} \subseteq \mathbf{PL}(T)$, the theory can be as weak as $\mathsf{I}\Delta_0 + \Omega_1$ or Buss's S^1_2 (see [8] and [12]). Arithmetical completeness, that is, that $\mathbf{PL}(T) \subseteq \mathbf{GL}$, is known to hold for any sound[1] theory extending $\mathsf{I}\Delta_0 + \exp$ (see [13]).

Solovay proved that whenever $\mathbf{GL} \nvdash A$, there is a realization $*$ so that $\mathsf{PA} \nvdash A^*$. 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 $\mathbf{GL} \nvdash A$. Next, a new root is added. A primitive recursive function f on the model is then 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 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.[2] A natural question to ask is whether we can find translations with clearer arithmetical and proof theoretic content. And conversely, given a set of arithmetical sentences with understood arithmetical content, what modal logic results from restricting realizations to this particular set? These questions motivate the following definition. We shall write $* \in R_\Gamma$ to mean that the realization $*$ takes on all its values within the set of sentences Γ.

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

Notice that $\mathbf{PL}_\Gamma(T)$ need not even be closed under substitution. 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 sentences we get $\mathbf{PL}_\Gamma(T) = \mathbf{PL}(T)$. For a large class of theories, however, we can improve this to the following theorem.

Theorem 1.5 *For all those theories T for which Solovay's Theorem 1.2 can be proved using the original proof we have that*

$$\mathbf{PL}_{\mathcal{B}(\Sigma_1)}(T) \subseteq \mathbf{PL}(T).$$

Here, $\mathcal{B}(\Sigma_1)$ denotes the class of Boolean combinations of Σ_1 sentences.

Proof By close inspection of the proof of Solovay's theorem, we see that all substitutions are disjunctions of limit statements. It is clear that for elementary functions h, the statement "h has a limit" can be expressed as a Σ_2 sentence. However, as Visser has pointed out to us, the statement "h has limit i" which is actually all that is needed in Solovay's proof, can be expressed as $\mathcal{B}(\Sigma_1)$:

$$[\exists x\ h(x) = i] \ \wedge\ [\forall y, z\ ((y \leq z \wedge hy = i) \rightarrow h(z) = i)]\,.$$

Disjunctions of these sentences will of course remain in $\mathcal{B}(\Sigma_1)$. □

Lemma 1.4 can be used to establish upper bounds for a provability logic in cases where full provability logic is unknown. For example, it is a longstanding open question what the provability logic is of bounded arithmetics such as S^1_2.[3]

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 as follows:

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

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 likewise, 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 [30], 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 x only occurs directly under the scope of some Bew_T predicate, then applying the fixed point to this formula does not give us more 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 sorts of applications of the arithmetical fixed point theorem only yield formulas in $\mathcal{B}$ and so, *pace* Theorem 3.2, cannot suffice for a proof of Solovay's completeness result, Theorem 1.2.[4]

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 (Theorem 4.5). However, 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 profitably 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 frame 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 $V : \mathsf{Prop} \to \wp(W)$ 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 $\langle W, R, V\rangle, w \vDash A$, just in case $w \in V(A)$. We write $\langle W, R, V\rangle \vDash A$ if $A \in V(w)$ for all $w \in W$. Overloading notation, we also write $\langle W, R\rangle \vDash A$, if $\langle W, R, V\rangle \vDash A$ 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 $\sigma_1, \sigma_2, \sigma_3, \ldots$, 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 $s_1, s_2, s_3, \ldots$, 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 $s_1, s_2, s_3, \ldots$ 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,

$$T \vdash A \;\Leftrightarrow\; \mathcal{M} \vDash A, \tag{1}$$

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 induction[5] on C we see again that

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

As i was arbitrary, it is clear that $\mathbb{F} \vDash C$. □

As we shall see below, in many occasions we will actually have something stronger than (1). In particular, we shall often have, apart from the frame, an auxiliary modal logic $\mathbf{L}$ for which this equivalence holds:

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

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 realizations. Recall the definition of the closed fragment $\mathcal{B}$ in Subsection 1.3.

Definition 3.1 **GL.3** is the logic **GL** together with the linearity axiom,

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

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

Theorem 3.2 $\mathbf{PL}_{\mathcal{B}}(T)$= **GL.3** *for a large class*[6] *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 [11].

Thus, the linear frame $\langle \omega, > \rangle$ is universal for $\mathcal{B}$ in the sense that if a formula $A \in \mathcal{B}$ 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 $T \vdash A \Leftrightarrow \langle \omega, > \rangle \vDash A$.

Furthermore, it is known that the logic of the frame $\langle \omega, > \rangle$ is axiomatized by **GL.3**. (See, for example, Chapter 13 of [11].) Thus, $\langle \omega, > \rangle \vDash A \Leftrightarrow \mathbf{GL.3} \vdash A$ 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 $\Diamond^n \top \wedge \Box^{n+1} \bot$. Thus, by Theorem 2.1 we have our result. □

4 Substitutions from the Closed Fragment of GLP

Japaridze's Logic **GLP** [21] 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 [20] 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) $[n]([n]A \to A) \to [n]A$, for all n,
(iii) $[m]A \to [n]A$, for $m \leq n$,
(iii) $\langle m \rangle A \to [n]\langle m \rangle A$, 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., [4] and [2]). In particular, Ignatiev [20] has defined a *universal frame* for the closed fragment of **GLP**, denoted $\mathbf{GLP}_0$, which will be of use.[7]

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

$$\mathcal{D} \ := \ \bot \mid \mathcal{D} \to \mathcal{D} \mid [0]\mathcal{D} \mid [1]\mathcal{D} \mid [2]\mathcal{D} \mid \ \ldots$$

$\mathbf{GLP}_0$ is simply **GLP** restricted to the fragment $\mathcal{D}$, with no variables.

We can describe Ignatiev's universal frame for $\mathbf{GLP}_0$ as follows. Let Ω consist of the set of ω-sequences of ordinals $(\alpha_0, \alpha_1, \alpha_2, \ldots)$, 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 $\omega^{\lambda_n} + \cdots + \omega^{\lambda_1}$, 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 $(\alpha, e(\alpha), e(e(\alpha)), \ldots)$, where $\alpha_{i+1} = e(\alpha_i)$ for all i, is called a *root point* and is denoted by $\widehat{\alpha}$ when α is the first coordinate. Thus every coordinate

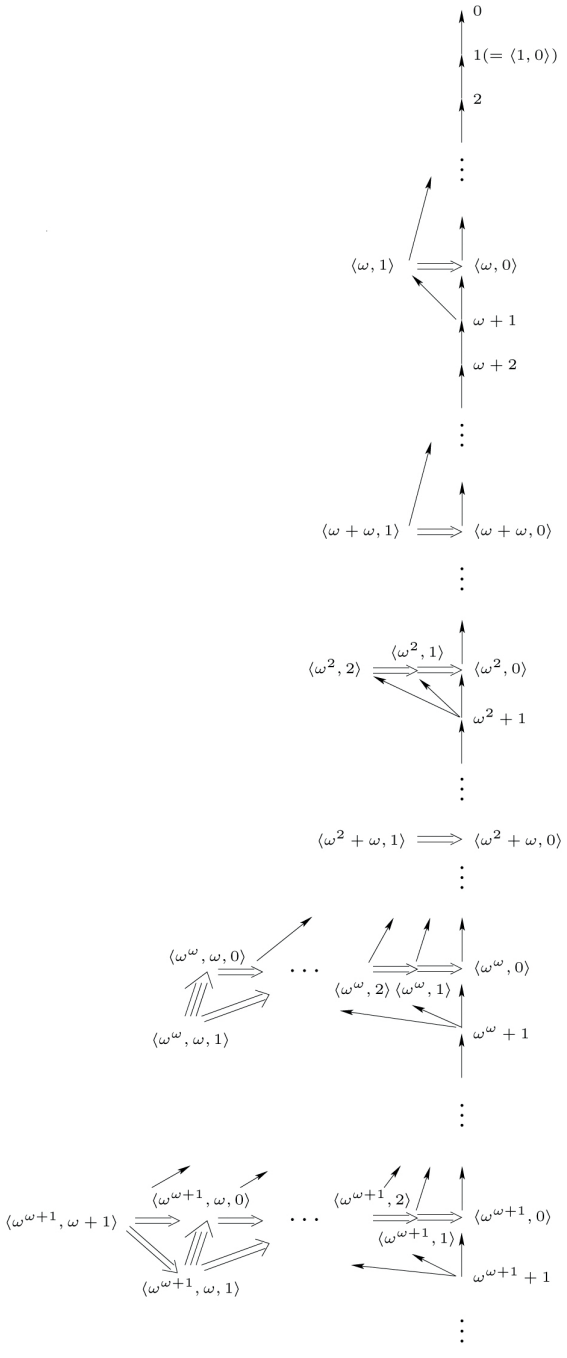

Figure 1 The universal model for $\mathbf{GLP_0}$

of $\widehat{\alpha}$ is uniquely determined by α. The following lemma is then obvious, given the definition of $\mathcal{U}$.

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

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 $\mathbf{GLP}_0 \nvdash A$, then there is a root point $\widehat{\alpha} \in U$ such that $\mathcal{U}, \widehat{\alpha} \nvDash A$.*

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 (cf. Theorem 3.2).

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

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

$$\mathsf{PRA} \vdash \Box(\Box A \to B) \vee \Box(\Box^+ B \to A),$$

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 $\widehat{\beta}$ and $\widehat{\gamma}$ such that $\mathcal{U}, \widehat{\beta} \vDash \Box A \wedge \neg B$, and $\mathcal{U}, \widehat{\gamma} \vDash \Box^+ B \wedge \neg A$. By Lemma 4.3, either $\widehat{\beta} R_0 \widehat{\gamma}$, $\widehat{\gamma} R_0 \widehat{\beta}$, or $\widehat{\beta} = \widehat{\gamma}$. All three lead to contradiction. $\Box$

5 Nonlinear 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 **GL.3** or just **GL**. In this section we fill in this gap by giving sufficient conditions on constants, so that we obtain logics of nonlinear **GL**-frames. We will be working with generic fragments $\mathcal{F}_n$, with some finite number n of constants:

$$\mathcal{F}_n \;:=\; s_1 \mid s_2 \mid \ldots \mid s_n \mid \bot \mid \mathcal{F}_n \to \mathcal{F}_n \mid \Box \mathcal{F}_n \,.$$

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 $\vec{\mathbf{s}}_{\mathbf{i}}$ stand for the sentence

$$\textstyle\bigwedge_{j \in J} s_{j+1} \wedge \bigwedge_{k \in K} \neg s_{k+1},$$

where J is the set of places in the binary expansion for i with value 1, and K is the complement of J in $\{0, \ldots, i-1\}$. 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 **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 $\Box^\alpha \bot$, where $\alpha < \omega + 1$ and $\Box^\omega \bot := \top$.

$\Box(\vec{s_0} \to B) \to \Box B$;
$\vdots$
$\Box(\vec{s_{2^n-1}} \to B) \to \Box B$.

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.

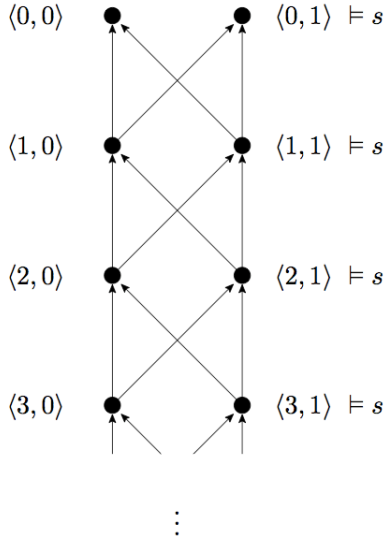

Figure 2 The model $\mathfrak{G}_1^\bullet$

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

Sketch of Proof The full proof for the case of $\mathcal{F}_1$ is established in [24]. Here we give a sketch for the general case. Soundness is routine. For completeness, we use the following two lemmas.

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

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

These lemmas are straightforwardly proven by manipulation of modal normal forms. Completeness is then clear. If $\mathbf{FGL}_n \nvdash A$, then by Lemma 5.6, $\mathbf{FGL}_n \nvdash \Box A$, and

by Lemma 5.5, $\mathbf{FGL}_n \vdash \Box A \leftrightarrow \Box^\alpha \bot$, for some $\alpha < \omega$ (in particular, $\alpha \neq \omega$). By soundness, for any point $\langle m, i\rangle \in G_n$ we know $\mathfrak{G}_n^\bullet, \langle m, i\rangle \vDash \Box A \leftrightarrow \Box^\alpha \bot$. Certainly $\mathfrak{G}_n^\bullet, \langle \alpha, 0\rangle \nvDash \Box^\alpha \bot$, so $\mathfrak{G}_n^\bullet, \langle \alpha, 0\rangle \nvDash \Box A$. That, in turn, means for some $\langle \beta, j\rangle$ with $\beta < \alpha$, we have $\mathfrak{G}_n^\bullet, \langle \beta, j\rangle \vDash \neg A$. So A is falsified on $\mathfrak{G}_n^\bullet$. □

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 Lemmas 5.5 and 5.6, arithmetical completeness of $\mathbf{L}$ depends only on arithmetical soundness of $\mathbf{L}$.

To see this, suppose $\mathbf{FGL}_n \nvDash A$. Then by Lemma 5.5, $\mathbf{FGL}_n \nvDash \Box A$. Since $\mathbf{FGL}_n \vDash \Box A \leftrightarrow \Box^\alpha \bot$, for $\alpha \neq \omega$, as long as we have soundness of $\mathbf{L}$, $T \vdash \Box A \leftrightarrow \Box^\alpha \bot$, 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 $T \nvdash \Box A$, from which it follows $T \nvdash A$.

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. First we exhibit a suitable constant for the case of $\mathcal{F}_1$.

5.3 A Constant for $\mathsf{I}\Sigma_1$. Recall $\mathsf{I}\Sigma_1$ is the theory 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 $\mathcal{F}_1$:

$$\mathcal{Q} \;:=\; \sigma \mid \bot \mid \mathcal{Q} \to \mathcal{Q} \mid \Box Q.$$

Our theory T will be Primitive Recursive Arithmetic (PRA), essentially just Q with function symbols for all of the primitive recursive functions and induction over Δ_0 formulas. The relationship between $\mathsf{I}\Sigma_1$ and PRA is well studied and understood ([28], [29], [3]). By Corollary 5.7, we need to show that $\mathbf{FGL}_n$ is sound with respect to PRA. It is already well known that $\mathbf{PL}(\mathsf{PRA}) = \mathbf{GL}$, so certainly all the axioms and rules of $\mathbf{GL}$ are sound. We need only observe the following also hold:

(i) $\mathsf{PRA} \vdash \Box(\sigma \to B) \to \Box B$,
(ii) $\mathsf{PRA} \vdash \Box(\neg\sigma \to B) \to \Box B$.

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 $\mathsf{I}\Sigma_1$ is Π_2-conservative over PRA. In [3] it is shown that this theorem is in fact formalizable in PRA, which gives us (i).

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

So this certainly holds for $\mathcal{B}(\Sigma_1)$ formulas consisting of Boolean combinations of formulas of the form $\Box^\alpha \bot$. As for (ii), it is shown in [24] that the negation of the sentence axiomatizing $\mathsf{I}\Sigma_1$ is Π_3-conservative over PRA. That is, we have the following lemma.

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

Thus, we can state the following corollary.

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

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. $\Box(\Box A \to (B \vee C)) \vee \Box(\Box^+ B \to (A \vee C)) \vee \Box(\Box^+ C \to (A \vee B))$;
Q2. $\Diamond(\Diamond A \wedge \Box B) \to \Box(\Diamond A \vee B)$.

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. Nontriple 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 [10]). 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, $\langle W^{\mathrm{GL.4}}, R^{\mathrm{GL.4}}, V^{\mathrm{GL.4}}\rangle$ with

(i) $W^{\mathrm{GL.4}}$ is the set of maximal **GL.4**-consistent sets;
(ii) for $\Gamma, \Delta \in W^{\mathrm{GL.4}}$, define $\Gamma R^{\mathrm{GL.4}} \Delta$ if for all $\varphi \in \Delta$ we have $\Diamond\varphi \in \Gamma$;
(iii) $V(p) = \{\Gamma : p \in \Gamma\}$, 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 [10], 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 $\mathbf{GL.4} \nvdash A$, 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 $\Delta \in W^{\mathrm{GL.4}}$ 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 ([10], Lemma 4.20) for normal modal logics, Δ is $R^{\mathrm{GL.4}}$-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 $\Gamma R^{\mathrm{GL.4}} \Delta$ 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 $\mathcal{M}$:

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

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 $\mathcal{M}'$ 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 $B \in \Delta$ if and only if $\mathfrak{M}', \Delta \vDash B$.*

Concluding the proof, since $A \notin \Gamma$, we have that $\mathfrak{M}', \Gamma \nvDash A$. □

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 $f : W \to W'$ 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.[8]

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 $\langle \mathbb{F}', V' \rangle, w' \nvDash A$, ensures the existence of a valuation V and point $w \in W$ such that $\langle \mathbb{F}, V \rangle, w \nvDash A$.*

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 $\mathbf{PL}_{\mathcal{Q}}(\mathsf{PRA}) = \mathbf{GL.4}$.

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 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*.

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, for example, 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 [18]). □

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 generalization[9] of nonbranching 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 nonbranching **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 **GL**-frames, then that of non-quadruple-branching **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 nontheorem of **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.

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 $\mathsf{I}\Sigma_1$ 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) **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 $\mathsf{I}\Sigma_1$ 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

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

8 Interpretability Logics with Restricted Substitutions

Interpretations, like proofs, are ubiquitous in mathematics and logic. Loosely speaking, an interpretation from a theory V into a theory U is a 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, for example, [22] or in [36].

8.1 Interpretability logics Interpretability can be seen as a generalization of provability. By $\alpha \rhd_T \beta$ 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 $\rhd$, orthographically identical to the arithmetical operator, to model formalized interpretability. And indeed, *arithmetical realizations* are extended as expected by imposing that

$$(A \rhd B)^* \;=\; A^* \rhd B^*.$$

For a clear distinction, let $\mathsf{Form}_{\mathbf{IL}}$ denote the class of modal formulas in language of interpretability logic and $\mathsf{Form}_{\mathbf{GL}}$ 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:

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

By Theorem 1.2 and Note 1 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 *essentially 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 Berarducci and Shavrukov. The definition of **ILM** will follow below.

Theorem 8.1 (Berarducci [7], Shavrukov [31]) *If T is an essentially reflexive and Σ_1 sound theory, then* $\mathbf{IL}(\mathrm{T}) = \mathbf{ILM}$.

However, if a theory is finitely axiomatizable we get a different outcome where, again, **ILP** is defined below.

Theorem 8.2 (Visser [35]) *If T is finitely axiomatizable, Σ_1 sound, and extends $I\Delta_0 + \mathsf{supexp}$, then* $\mathbf{IL}(\mathrm{T}) = \mathbf{ILP}$.

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 **IL**(All), is the set of formulas φ such that for all T and $*$, $T \vdash \varphi^*$. Here we let T range over all reasonable[10] arithmetical theories.

Clearly, **IL**(All) is in the intersection of **ILM** and **ILP** but apparently it possesses a very rich structure (see [26], and [14]). In this paper, it is only important to know that a certain very weak logic to be defined below is part of **IL**(PRA).

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

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 **IL**(PRA) is still unknown. The most recent results for **IL**(PRA) are presented in [9].

PRA is known to be the same as $\mathrm{I}\Sigma_1^R$ where $\mathrm{I}\Sigma_n^R$ is defined as $\mathsf{I}\Delta_0 + \exp$ plus the Σ_n induction rule. See, for example, [1]. In that paper a proof can also be found for the following theorem.

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

The logical complexity of interpretability is Σ_3 and in [32] 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 $U \rhd_{\Pi_1} V$, whenever for all Π_1 sentences π we have that $[V \vdash \pi$ implies $U \vdash \pi]$.

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

$$(U \rhd V) \quad \Leftrightarrow \quad (U \rhd_{\Pi_1} V)$$

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 $\mathsf{Form}_{\mathbf{IL}}$ that are provable in T under any arithmetical realization where the $\rhd$ modality is mapped to $\rhd_{\Pi_1}$.

Theorem 8.7 *For any sound theory T extending $\mathsf{I}\Pi_1^-$ we have that the Π_1-conservativity logic of T is* **ILM**.

The theorem was first proven by Hájek and Montagna in [15] and [16] to hold for any sound theory containing $\mathsf{I}\Sigma_1$. Beklemishev and Visser in [6] improved this to any theory containing $\mathsf{I}\Pi_1^-$, which allows induction only for parameter-free formulas of complexity Π_1. It is well known that PRA extends $\mathsf{I}\Pi_1^-$ [1].

Remark 8.8 The proof of Theorem 8.7 is rather similar to that of Solovay's original proof and again (see Theorem 1.5), the substitutions in the completeness proof can be taken[11] to be Σ_2.

The logics **ILM** and **ILP** have elegant syntactical presentations. We shall define them in parts. First, we define a logic **IL** that is a sublogic of all interpretability logics of interest. Next this logic **IL** is extended by adding more axiom schemata.

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

Definition 8.9 The logic **IL** is the smallest set of formulas closed under the rules of Necessitation and of Modus Ponens, containing all tautological formulas and all instantiations of the following axiom schemata.

L1 $\Box(A \to B) \to (\Box A \to \Box B)$
L2 $\Box A \to \Box\Box A$
L3 $\Box(\Box A \to A) \to \Box A$
J1 $\Box(A \to B) \to A \rhd B$
J2 $(A \rhd B) \wedge (B \rhd C) \to A \rhd C$
J3 $(A \rhd C) \wedge (B \rhd C) \to A \vee B \rhd C$
J4 $A \rhd B \to (\Diamond A \to \Diamond B)$
J5 $\Diamond A \rhd A$

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

M $A \rhd B \to A \wedge \Box C \rhd B \wedge \Box C$
P $A \rhd B \to \Box(A \rhd B)$
W $A \rhd B \to A \rhd B \wedge \Box\neg A$

If X is a set of axiom schemata we will denote by **IL**X the logic that arises by adding the axiom schemata in X to **IL**.

8.2 The closed fragment Because closed formulas in **ILW** can be reduced to those of **GL** [17] we can prove that $\mathbf{IL}_{\mathcal{B}}(\mathrm{PRA})$ is again the logic of linear frames.

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

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

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

$$\mathbf{ILW.3} \vdash \varphi \Leftrightarrow \mathbf{GL.3} \vdash \varphi^{\mathrm{tr}} \quad (*)$$

and

$$\mathbf{ILW.3} \vdash \varphi \leftrightarrow \varphi^{\mathrm{tr}}. \quad (**)$$

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

$$\forall\, * \in \mathsf{Sub}(\mathcal{B})\ \mathrm{PRA} \vdash \varphi^* \leftrightarrow (\varphi^{\mathrm{tr}})^*$$

and consequently

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

We first see that $(***)$ holds. Certainly, by Fact 8.4, we have that $\mathbf{ILW} \subseteq \mathbf{IL}_{\mathcal{B}}(\mathrm{PRA})$. Thus it remains to show that $\mathrm{PRA} \vdash \Box(\Box A^* \to B^*) \vee \Box(\boxdot B^* \to A^*)$ for any formulas A and B in $\mathsf{Form}_{\mathbf{IL}}$ and any $* \in \mathcal{B}$. As any formula in the closed fragment of **ILW** is equivalent to a formula in the closed fragment of **GL** (see [17]), Theorem 3.2 gives us that indeed the linearity axiom holds for the closed fragment of **GL**.

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

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

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

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 we get $p \rhd q \wedge \Box\neg p$. Together with $\Box(q \to \Diamond p)$, 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 $\mathbf{ILW.3} \vdash \varphi$ we see that $\mathbf{GL.3} \vdash \varphi^{\mathrm{tr}}$. All the specific interpretability axioms turn out to be provable under our translation in **GL**. The only axioms where the $\Box A \to \Box\Box A$ axiom scheme is really used are in J_2 and J_4. To prove the translation of W we also need L_3. If $\mathbf{GL.3} \vdash \varphi^{\mathrm{tr}}$ then certainly $\mathbf{ILW.3} \vdash \varphi^{\mathrm{tr}}$ and by $(**)$, $\mathbf{ILW.3} \vdash \varphi$. □

We thus see that **ILW.3** is an upper bound for **IL**(PRA). Using the translation from the proof of Theorem 8.11, it is not hard to see that both the principles P and M are provable in **ILW.3**. This tells us that the upper bound is actually not very informative as we know that $\mathbf{IL}(\mathrm{PRA}) \nvdash \mathsf{M}$. By a straightforward generalization of Lemma 1.4 we see that choosing larger Γ will generally yield a smaller $\mathbf{IL}_\Gamma(\mathrm{PRA})$ and thus a sharper upper bound. In Subsection 8.4 we shall discuss just how large the Γ should be as to refute M in $\mathbf{IL}_\Gamma(\mathrm{PRA})$. First 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 consideration. In particular, the theorem still holds for interpretability logics. In the theorem below we use this to give a semantic characterization of $\mathbf{IL}_{\mathcal{F}_1}(\mathrm{PRA})$.

In [25] it is established that for a certain frame, that we will denote here by $\widetilde{\mathfrak{G}_1^\bullet}$, we have the following equivalence.

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

For the purpose of this paper it is not important what the frame $\widetilde{\mathfrak{G}_1^\bullet}$ looks like. We need only know that $\widetilde{\mathfrak{G}_1^\bullet}$ is just like $\mathfrak{G}_1^\bullet$ with some additional accessibility relations to model the $\rhd$ modality. This, together with the mere equivalence $(\dagger)$, is enough to obtain the following theorem.

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

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. □

In [25], a logic **PIL** is also defined such that we have

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

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

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

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

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

Lemma 8.14 *For $n \geq 1$, we have that* $\mathbf{IL}(\mathrm{I}\Sigma_n^R) \nvdash \mathsf{M}$.

Proof We define a realization $*$ such that $\mathrm{I}\Sigma_n^R \nvdash (p \rhd q \to p \wedge \Box r \rhd q \wedge \Box r)^*$. It is well known that $\mathrm{I}\Sigma_n^R \subsetneq \mathrm{I}\Sigma_n \subsetneq \mathrm{I}\Sigma_{n+1}^R$ 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 (for $n \geq 1$) $\mathbf{IL}(\mathrm{I}\Sigma_n) = \mathbf{ILP}$ and that $\mathbf{ILP} \nvdash p \rhd q \to p \wedge \Box r \rhd q \wedge \Box r$. Thus, for any $n \geq 1$, we 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

$$\mathrm{EA} \vdash \alpha_n \rhd_{\mathrm{I}\Sigma_n} \beta_n \leftrightarrow \sigma_n \wedge \alpha_n \rhd_{\mathrm{I}\Sigma_n^R} \sigma_n \wedge \beta_n$$

and

$$\mathrm{EA} \vdash \Box_{\mathrm{I}\Sigma_n}\gamma_n \leftrightarrow \Box_{\mathrm{I}\Sigma_n^R}(\sigma_n \to \gamma_n).$$

Thus, we have

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

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

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

By Theorem 1 from [9] (Theorem 12.1.1 from [24]) we know that for $\alpha, \beta \in \Sigma_2$ we have

$$\mathrm{PRA} \vdash (\alpha \rhd \beta) \to ((\alpha \wedge \Box\gamma) \rhd (\beta \wedge \Box\gamma))$$

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

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

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

Proof We use the fact that the logic of Π_1-conservativity for theories containing $\mathrm{I}\Pi_1^-$ is **ILM** as mentioned in Theorem 8.7. 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 $\mathbf{IL}_{\Sigma_2}(\mathrm{I}\Sigma_n^R) \subseteq \mathbf{ILM}$ and $\mathbf{ILM} \subseteq \mathbf{IL}_{\Delta_{n+1}}(\mathrm{I}\Sigma_n^R)$.

First, we prove by induction on the complexity of a modal formula A that for all $* \in \Delta_{n+1}$ $\mathrm{I}\Sigma_n^R \vdash A^*_{\Pi_1} \leftrightarrow A^*_{\rhd}$ 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 $\mathrm{I}\Sigma_n^R$:

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

Note that we have access to the Orey-Hájek characterization as $B^*_{\Pi_1}$ is of complexity at most Δ_{n+1} and thus $\mathrm{I}\Sigma_n^R + B^*_{\Pi_1}$ is a reflexive theory by Theorem 8.5. 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 $\mathrm{I}\Sigma_n^R \vdash A^*_{\Pi_1}$ and thus whenever $* \in \Delta_{n+1}$, $\mathrm{I}\Sigma_n^R \vdash A^*_{\rhd}$ and $\mathbf{ILM} \subseteq \mathbf{IL}_{\Delta_{n+1}}(\mathrm{I}\Sigma_n^R)$. If $\mathbf{ILM} \nvdash A$ then by Remark 8.8 for some $* \in \Sigma_2$ we have $\mathrm{I}\Sigma_n^R \nvdash A^*_{\Pi_1}$ whence $\mathrm{I}\Sigma_n^R \nvdash A^*_{\rhd}$. We may conclude that $\mathbf{IL}_{\Sigma_2}(\mathrm{I}\Sigma_n^R) \subseteq \mathbf{ILM}$. □

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

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

9 Future Research

We have seen that adding a constant for $\mathrm{I}\Sigma_1$ to PRA is sufficient to obtain a nontrivial provability logic. By a theorem of Leivant it is known that $\mathrm{I}\Sigma_1 \equiv \langle 2 \rangle_{\mathrm{EA}} \top$. An interesting fragment to consider next for PRA would be the closed fragment together with the set of constants

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

or variants thereof.

Notes

1. Σ_1-soundness is sufficient.

2. There is a paper by de Jongh, Jumelet and Montagna [13] 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 main ideas are not essentially different from those used in Solovay's original proof.

3. This question has been studied in depth in [8]. It also has important connections to matters in computational complexity. For example, it is shown in [12] that if S^1_2 proves Π^b_1-completeness with parameters (Π^b_1 is the set of formulas $(\forall x \leq t)\, \theta$ with θ sharply bounded), then NP = coNP.

4. 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.

5. In 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 $\langle \mathbb{F}, V\rangle, j \vDash C \;\Leftrightarrow\; \mathcal{M}, j \vDash C^*$.

6. See Note 1 on conditions on theories. The current proof of this theorem invokes Solovay's completeness result, Theorem 1.2, in full. However, in [23] 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 $\mathsf{I}\Delta_0 + \Omega_1$.

7. This frame is studied in detail in [5] and [19].

8. See, e.g., [10], Definition 3.13, where p-morphisms go under the name *bounded morphism*.

9. It is not hard to see that $\Box(\Box A \to B) \vee \Box(\Box^+ B \to A)$ is equivalent to $\Box(\Box^+ A \to B) \vee \Box(\Box^+ B \to A)$ over **GL**.

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

11. Albert Visser (personal conversation) notes that close inspection of the proof actually reveals that the substitutions can be taken to be $\Delta_2(\mathsf{I}\Pi_1^-)$. That is, a Σ_2 sentence that is probably in $\mathsf{I}\Pi_1^-$ equivalent to a Π_2 sentence.

References

[1] Beklemishev, L. D., "Reflection schemes and provability algebras in formal arithmetic," *Uspekhi Matematicheskikh Nauk*, vol. 60 (2005), pp. 3–78. vol. 60, pp. 197–268. Zbl 1097.03054. MR 2152943.

[2] Beklemishev, L. D., G. Bezhanishvili, and T. Icard, *On Topological Semantics of* **GLP**, edited by R. Schindler, Ontos Verlag, Frankfurt am Main, 2010. Zbl 1194.03006.

[3] Beklemishev, L. D., "Bimodal logics for extensions of arithmetical theories," *The Journal of Symbolic Logic*, vol. 61 (1996), pp. 91–124. Zbl 0858.03024. MR 1380679.

[4] Beklemishev, L. D., "Kripke semantics for provability logic GLP," *Annals of Pure and Applied Logic*, vol. 161 (2010), pp. 756–74. Zbl 1223.03046. MR 2601030.

[5] Beklemishev, L. D., J. J. Joosten, and M. Vervoort, "A finitary treatment of the closed fragment of Japaridze's provability logic," *Journal of Logic and Computation*, vol. 15 (2005), pp. 447–63. Zbl 1080.03038. MR 2157727.

[6] Beklemishev, L. D., and A. Visser, "On the limit existence principles in elementary arithmetic and Σ^0_n-consequences of theories," *Annals of Pure and Applied Logic*, vol. 136 (2005), pp. 56–74. Zbl 1087.03037. MR 2162847.

[7] Berarducci, A., "The interpretability logic of Peano arithmetic," *The Journal of Symbolic Logic*, vol. 55 (1990), pp. 1059–89. Zbl 0725.03037. MR 1071315.

[8] Berarducci, A., and R. Verbrugge, "On the provability logic of bounded arithmetic," *Annals of Pure and Applied Logic*, vol. 61 (1993), pp. 75–93. Provability, Interpretability and Arithmetic Symposium (Utrecht, 1991). Zbl 0803.03037. MR 1218656.

[9] Bílková, M., D. de Jongh, and J. J. Joosten, "Interpretability in PRA," *Annals of Pure and Applied Logic*, vol. 161 (2009), pp. 128–38. Zbl 1184.03011. MR 2552733.

[10] Blackburn, P., M. de Rijke, and Y. Venema, *Modal Logic*, vol. 53 of *Cambridge Tracts in Theoretical Computer Science*, Cambridge University Press, Cambridge, 2001. Zbl 0988.03006. MR 1837791.

[11] Boolos, G., *The Logic of Provability*, Cambridge University Press, Cambridge, 1993. Zbl 0891.03004. MR 1260008.

[12] Buss, S. R., *Bounded Arithmetic*, vol. 3 of *Studies in Proof Theory. Lecture Notes*, Bibliopolis, Naples, 1986. Ph.D. thesis. Zbl 0649.03042. MR 880863.

[13] de Jongh, D., M. Jumelet, and F. Montagna, "On the proof of Solovay's theorem," *Studia Logica*, vol. 50 (1991), pp. 51–69. Zbl 0744.03057. MR 1152780.

[14] Goris, E., and J. J. Joosten, "A new principle in the interpretability logic of all reasonable arithmetical theories," *Logic Journal of the IGPL*, vol. 19 (2011), pp. 14–17. Zbl 1228.03040. MR 2770562.

[15] Hájek, P., and F. Montagna, "The logic of Π_1-conservativity," *Archive for Mathematical Logic*, vol. 30 (1990), pp. 113–23. Zbl 0713.03007. MR 1075648.

[16] Hájek, P., and F. Montagna, "The logic of Π_1-conservativity continued," *Archive for Mathematical Logic*, vol. 32 (1992), pp. 57–63. Zbl 0790.03018. MR 1186467.

[17] Hájek, P., and V. Švejdar, "A note on the normal form of closed formulas of interpretability logic," *Studia Logica*, vol. 50 (1991), pp. 25–28. Zbl 0728.03015. MR 1152777.

[18] Icard, T., "Towards an alternative proof of Solovay's Arithmetical Completeness Theorem," *Proceedings of the 12th ESSLLI Student Session*, Dublin, 2007.

[19] Icard, T., "A topological study of the closed fragment of GLP," *Journal of Logic and Computation*, vol. 21 (2011), pp. 683–96. Zbl pre05961423. MR 2823435.

[20] Ignatiev, K. N., "On strong provability predicates and the associated modal logics," *The Journal of Symbolic Logic*, vol. 58 (1993), pp. 249–90. Zbl 0795.03082. MR 1217189.

[21] Japaridze, G., *Modal Logical Means of Investigation of Provability*, Ph.D. thesis, Moscow State University, Moscow, 1986. In Russian.

[22] Japaridze, G., and D. de Jongh, "The logic of provability," pp. 475–546 in *Handbook of Proof Theory*, edited by S. R. Buss, vol. 137 of *Studies in Logic and the Foundations of Mathematics*, Elsevier, Amsterdam, 1998. Zbl 0915.03019. MR 1640331.

[23] Joosten, J. J., "Formalized interpretability in Primitive Recursive Arithmetic," *Proceedings of the ESSLLI Student Session*, Vienna, 2003.

[24] Joosten, J., *Intepretability Formalized*, Ph.D. thesis, Department of Philosophy, University of Utrecht, Utrecht, 2004.

[25] Joosten, J. J., "The closed fragment of the interpretability logic of PRA with a constant for $I\Sigma_1$," *Notre Dame Journal of Formal Logic*, vol. 46 (2005), pp. 127–46. Zbl 1077.03034. MR 2150947.

[26] Joosten, J. J., and A. Visser, "The interpretability logic of all reasonable arithmetical theories. The new conjecture," *Erkenntnis*, vol. 53 (2000), pp. 3–26. Zbl 0974.03049. MR 1799970.

[27] Löb, M. H., "Solution of a problem of Leon Henkin," *The Journal of Symbolic Logic*, vol. 20 (1955), pp. 115–18. Zbl 0067.00202. MR 0070596.

[28] Mints, G., "Quantifier-free and one-quantifier systems," *Journal of Soviet Mathematics*, vol. 1 (1972), pp. 71–84. Zbl 0222.02022.

[29] Parsons, C., "On n-quantifier induction," *The Journal of Symbolic Logic*, vol. 37 (1972), pp. 466–82. Zbl 0264.02027. MR 0325365.

[30] Sambin, G., "An effective fixed-point theorem in intuitionistic diagonalizable algebras. The algebraization of the theories which express Theor, IX," *Studia Logica*, vol. 35 (1976), pp. 345–61. Zbl 0357.02028. MR 0460116.

[31] Shavrukov, V., *The Logic of Relative Interpretability over Peano Arithmetic* (in Russian), Steklov Mathematical Institute, Moscow, 1988. Technical Report No. 5.

[32] Shavrukov, V. Y., "Interpreting reflexive theories in finitely many axioms," *Fundamenta Mathematicae*, vol. 152 (1997), pp. 99–116. Zbl 0874.03066. MR 1441229.

[33] Tarski, A., *Undecidable Theories*, Studies in Logic and the Foundations of Mathematics. North-Holland Publishing Company, Amsterdam, 1953. In collaboration with A. Mostowski and R. M. Robinson. Zbl 0053.00401. MR 0058532.

[34] Visser, A., "A propositional logic with explicit fixed points," *Studia Logica*, vol. 40 (1981), pp. 155–75. Zbl 0469.03012. MR 648575.

[35] Visser, A., "Interpretability logic," pp. 175–209 in *Mathematical Logic. Proceedings of the Heyting 1988 Summer School in Varna, Bulgaria.*, edited by P. P. Petkov, Plenum Press, Boston, 1990. Zbl 0793.03064. MR 1083994.

[36] Visser, A., "An overview of interpretability logic," pp. 307–59 in *Advances in Modal Logic, Vol. 1 (Berlin, 1996)*, edited by M. Kracht, M. de Rijke, and H. Wansing, vol. 87 of *CSLI Lecture Notes*, CSLI Publications, Stanford, 1998. Zbl 0915.03020. MR 1688529.

Acknowledgments

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

Department of Philosophy
Stanford University
Bldg 90
Stanford CA 94305
USA
icard@stanford.edu

Dept. Lògica, Història i Filosofia de la Ciència,
Universitat de Barcelona,
Montalegre, 6
08001 Barcelona, Catalonia
SPAIN
jjoosten@ub.edu