Self Provers and Σ_1 Sentences

Evan Goris, *Department of Computer Science, the Graduate Center of the City University of New York, 365 Fifth avenue, New York, NY 10016, USA, E-mail: evangoris@gmail.com*

Joost J. Joosten, *Departamento de Filosofía, Lógica y Filosofía de la Ciencia, Universidad de Sevilla, c/ Camilo José Cela s/n, 41018 Sevilla, España, E-mail: jjoosten@us.es*

Abstract

This paper is the second in a series of three papers. All three papers deal with interpretability logics and related matters. In the first paper a construction method was exposed to obtain models of these logics. Using this method, we obtained some completeness results, some already known, and some new.

In this paper, we will set the construction method to work to obtain more results. First, the modal completeness of the logic **IL**M is proved using the construction method. This is not a new result, but by using our new proof we can obtain new results. Among these new results are some admissible rules for **IL**M and **GL**.

Moreover, the new proof will be used to classify all the essentially Δ_1 and also all the essentially Σ_1 formulas of **IL**M. Closely related to essentially Σ_1 sentences are the so-called *self provers*. A self-prover is a formula φ which implies its own provability, that is $\varphi \to \Box\varphi$. Each formula φ will generate a self prover $\varphi \wedge \Box\varphi$. We will use the construction method to characterize those sentences of **GL** that generate a self prover that is trivial in the sense that it is Σ_1.

Keywords: Interpretability Logics, Provability Logic, Modal Completeness, Sigma Sentences, Self-provers

1 Introduction

Mathematical interpretations occur everywhere in (meta) mathematical practice. Interpretability logics study structural behavior of interpretations. One such logic, the logic **IL**M, describes the structural behavior of interpretations over theories like Peano Arithmetic. In this paper we shall first study this logic **IL**M and then use our findings to derive new results mainly related to Σ_1 sentences of theories like Peano Arithmetic.

This paper is the second in a series of three. For more background on interpretations and their corresponding logics we refer to the first part of this paper [20]. Also, all definitions used in this paper occur with some motivation and background in [20]. For completeness, self-containenedness and for readability we shall include a short recap in this paper of those technicalities that were introduced in [20] and that are central to this paper.

2 A concise recap: central notions of this paper

In this paper we shall heavily resort to some rather technical results obtained in [20]. In particular, certain parts of proofs in [20] shall be re-used here. In this section, we shall state those parts of that paper which are necessary for results further on.

2.1 Interpretability logics

The modal sentences in this paper are mostly in the language of interpretability which is defined as follows.

$$\mathsf{Form}_{\mathbf{IL}} := \bot \mid \mathsf{Prop} \mid (\mathsf{Form}_{\mathbf{IL}} \to \mathsf{Form}_{\mathbf{IL}}) \mid (\Box \mathsf{Form}_{\mathbf{IL}}) \mid (\mathsf{Form}_{\mathbf{IL}} \rhd \mathsf{Form}_{\mathbf{IL}})$$

Here Prop is a countable set of propositional variables $p, q, r, s, t, p_0, p_1, \ldots$. We employ the usual definitions of the logical operators $\neg, \vee, \wedge$ and $\leftrightarrow$. Also shall we write $\Diamond\varphi$ for $\neg\Box\neg\varphi$. Formulas that start with a $\Box$ are called box-formulas or $\Box$-formulas. Likewise we talk of $\Diamond$-formulas.

For standard reading conventions on bracketing please refer to [20].

The basic interpretability logic is called **IL** and is captured in the following definition.

Definition 2.1
The logic **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 $\Box(A \to B) \to (\Box A \to \Box B)$
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 $\Diamond A \rhd A$

We will write $\mathbf{IL} \vdash \varphi$ for $\varphi \in \mathbf{IL}$. 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**. Gödel Löb's logic **GL** is obtained from **IL** by omitting all the J axioms and not allowing the $\rhd$ modality in the language.

The standard semantics for interpretability logics are given by the following definitions.

Definition 2.2
An **IL**-frame is a triple $\langle W, R, S\rangle$. Here W is a non-empty countable universe, R is a binary relation on W and S is a set of binary relations on W, indexed by elements of W. The R and S satisfy the following requirements.

1. R is conversely well-founded[1]

[1] A relation R on W is called conversely well-founded if every non-empty subset of W has an R-maximal element.

2. xRy & $yRz \rightarrow xRz$
3. $yS_xz \rightarrow xRy$ & xRz
4. $xRy \rightarrow yS_xy$
5. $xRyRz \rightarrow yS_xz$
6. $uS_xvS_xw \rightarrow uS_xw$

IL-frames are sometimes also called Veltman frames. We will on occasion speak of R or S_x transitions instead of relations. If we write ySz, we shall mean that yS_xz for some x. W is sometimes called the universe, or domain, of the frame and its elements are referred to as worlds or nodes. With $x{\upharpoonright}$ we shall denote the set $\{y \in W \mid xRy\}$. We will often represent S by a ternary relation in the canonical way, writing $\langle x, y, z\rangle$ for yS_xz.

Definition 2.3
An **IL**-model is a quadruple $\langle W, R, S, \Vdash\rangle$. Here $\langle W, R, S, \rangle$ is an **IL**-frame and $\Vdash$ is a subset of $W \times \mathsf{Prop}$. We write $w \Vdash p$ for $\langle w, p\rangle \in \Vdash$. As usual, $\Vdash$ is extended to a subset $\widetilde{\Vdash}$ of $W \times \mathsf{Form}_{\mathbf{IL}}$ by demanding the following.

- $w \widetilde{\Vdash} p$ iff $w \Vdash p$ for $p \in \mathsf{Prop}$
- $w \not\widetilde{\Vdash} \bot$
- $w \widetilde{\Vdash} A \rightarrow B$ iff $w \not\widetilde{\Vdash} A$ or $w \widetilde{\Vdash} B$
- $w \widetilde{\Vdash} \Box A$ iff $\forall v\ (wRv \Rightarrow v \widetilde{\Vdash} A)$
- $w \widetilde{\Vdash} A \rhd B$ iff $\forall u\ (wRu \wedge u \widetilde{\Vdash} A \Rightarrow \exists v\ (uS_wv \widetilde{\Vdash} B))$

Note that $\widetilde{\Vdash}$ is completely determined by $\Vdash$. Thus we will denote $\widetilde{\Vdash}$ also by $\Vdash$. It is an easy observation that the truth of a modal formula in a particular world in the model is completely determined by the part of the model that "can be seen"from that world. This observation is used often and therefore we explicitly restate it here.

Definition 2.4 (Generated Submodel)
Let $M = \langle W, R, S, \Vdash\rangle$ be an **IL**-model and let $m \in M$. We define $m{\upharpoonright}*$ to be the set $\{x \in W \mid x{=}m \vee mRx\}$. By $M{\upharpoonright}m$ we denote the submodel generated by m defined as follows.

$$M{\upharpoonright}m := \langle m{\upharpoonright}*, R \cap (m{\upharpoonright}*)^2, \bigcup_{x \in m{\upharpoonright}*} S_x \cap (m{\upharpoonright}*)^2, \Vdash \cap (m{\upharpoonright}* \times \mathsf{Prop})\rangle$$

Lemma 2.5 (Generated Submodel Lemma)
Let M be an **IL**-model and let $m \in M$. For all formulas φ and all $x \in m{\upharpoonright}*$ we have that

$$M{\upharpoonright}m, x \Vdash \varphi \quad \text{iff} \quad M, x \Vdash \varphi.$$

In [20] models are built by gluing sets of modal sentences together. We shall briefly recapitulate the main definitions of those sets of sentences here.

Definition 2.6
A set Γ is **IL**X-consistent iff $\Gamma \nvdash_{\mathbf{ILX}} \bot$. An **IL**X-consistent set is maximal **IL**X-consistent if for any φ, either $\varphi \in \Gamma$ or $\neg\varphi \in \Gamma$.

We will often abbreviate "maximal consistent set" by MCS and refrain from explicitly mentioning the logic **IL**X when the context allows us to do so. We define three useful relations on MCS's, the *successor* relation $\prec$, the *C-critical successor* relation $\prec_C$ and the *Box-inclusion* relation $\subseteq_\Box$.

Definition 2.7
Let Γ and Δ denote maximal **IL**X-consistent sets.

- $\Gamma \prec \Delta := \Box A \in \Gamma \Rightarrow A, \Box A \in \Delta$
- $\Gamma \prec_C \Delta := A \rhd C \in \Gamma \Rightarrow \neg A, \Box \neg A \in \Delta$
- $\Gamma \subseteq_\Box \Delta := \Box A \in \Gamma \Rightarrow \Box A \in \Delta$

It is clear that $\Gamma \prec_C \Delta \Rightarrow \Gamma \prec \Delta$. For, if $\Box A \in \Gamma$ then $\neg A \rhd \bot \in \Gamma$. Also $\bot \rhd C \in \Gamma$, whence $\neg A \rhd C \in \Gamma$. If now $\Gamma \prec_C \Delta$ then $A, \Box A \in \Delta$, whence $\Gamma \prec \Delta$. It is also clear that $\Gamma \prec_C \Delta \prec \Delta' \Rightarrow \Gamma \prec_C \Delta'$.

Lemma 2.8
Let Γ and Δ denote maximal **IL**X-consistent sets. We have $\Gamma \prec \Delta$ iff $\Gamma \prec_\bot \Delta$.

2.2 The construction method and the Main Lemma

The main purpose of [20] was to provide some background in the modal theory of provability logics. Moreover, in that paper, a construction method was developed. The construction method provided a way of gluing sets of modal sentences together as to obtain models with desired properties. The ideas involved are quite similar to the definition of canonical models with the exception that the model is constructed step by step rather than defined at once and, moreover, only that part of the model that you need is constructed and nothing more.

Thus, the building blocks are maximal consistent sets of modal interpretability logics. Instead of gluing these sets together outright, we shall glue variables $u, v, \ldots$ together and label these variables by the sets. We denote the labeling by ν. Thus, if we added a new element x, by $\nu(x)$ we refer to the corresponding set of modal sentences. Likewise, certain R transitions will be labeled via ν with a single formula, for example $\nu(\langle x, y\rangle) = C$.

The following two notions are central to the construction method. As they are so central to the paper we strongly advice the reader who is novice to this field to read the motivation of these notions in Section 3 of [20].

Definition 2.9
Let x be a world in some **IL**X-labeled frame $\langle W, R, S, \nu\rangle$. The *C-critical cone above* x, we write $\mathcal{C}_x^C$, is defined inductively as

- $\nu(\langle x, y\rangle) = C \Rightarrow y \in \mathcal{C}_x^C$
- $x' \in \mathcal{C}_x^C$ & $x' S_x y \Rightarrow y \in \mathcal{C}_x^C$
- $x' \in \mathcal{C}_x^C$ & $x' R y \Rightarrow y \in \mathcal{C}_x^C$

Definition 2.10
Let x be a world in some **IL**X-labeled frame $\langle W, R, S, \nu\rangle$. The *generalized C-cone above* x, we write $\mathcal{G}_x^C$, is defined inductively as

- $y \in \mathcal{C}_x^C \Rightarrow y \in \mathcal{G}_x^C$
- $x' \in \mathcal{G}_x^C$ & $x' S_w z \Rightarrow z \in \mathcal{G}_x^C$ for arbitrary w
- $x' \in \mathcal{G}_x^C$ & $x' R y \Rightarrow y \in \mathcal{G}_x^C$

The construction method in essence deals step by step with existential requirements – so-called *problems*– and with universal requirements – so-called *deficiencies* – both defined below.

DEFINITION 2.11 (Problems)
Let $\mathcal{D}$ be some set of sentences. A *$\mathcal{D}$-problem* is a pair $\langle x, \neg(A \rhd B)\rangle$ such that $\neg(A \rhd B) \in \nu(x) \cap \mathcal{D}$ and for no $y \in \mathcal{C}_x^B$ we have $A \in \nu(y)$.

DEFINITION 2.12 (Deficiencies)
Let $\mathcal{D}$ be some set of sentences and let $F = \langle W, R, S, \nu\rangle$ be an **IL**X-labeled frame. A *$\mathcal{D}$-deficiency* is a triple $\langle x, y, C \rhd D\rangle$ with xRy, $C \rhd D \in \nu(x) \cap \mathcal{D}$, and $C \in \nu(y)$, but for no z with yS_xz we have $D \in \nu(z)$.

If the set $\mathcal{D}$ is clear or fixed, we will just speak about problems and deficiencies. The labeled frames we will construct are always supposed to satisfy some minimal reasonable requirements. We summarize these in the notion of adequacy.

DEFINITION 2.13 (Adequate frames)
A frame is called *adequate* if the following conditions are satisfied.

1. $xRy \Rightarrow \nu(x) \prec \nu(y)$
2. $A \neq B \Rightarrow \mathcal{G}_x^A \cap \mathcal{G}_x^B = \varnothing$
3. $y \in \mathcal{C}_x^A \Rightarrow \nu(x) \prec_A \nu(y)$

We need three more technical definitions before we can re-state the Main Lemma.

DEFINITION 2.14
Let $\mathcal{D}$ be some set of formulas and let M be an interpretability model. We say that *a Truth-Lemma holds on M with respect to $\mathcal{D}$* if for all x in M we have that

$$\forall\, \varphi{\in}\mathcal{D}\ [x \Vdash \varphi \text{ iff. } \varphi \in x].$$

DEFINITION 2.15 (Depth)
The *depth* of a finite frame F, we will write $\mathsf{depth}(F)$ is the maximal length of sequences of the form $x_0 R \ldots R x_n$. (For convenience we define $\max(\varnothing) = 0$.)

DEFINITION 2.16 (Union of Bounded Chains)
An indexed set $\{F_i\}_{i\in\omega}$ of labeled frames is called a *chain* if for all i, $F_i \subseteq F_{i+1}$. It is called a *bounded chain* if for some number n, $\mathsf{depth}(F_i) \leq n$ for all $i \in \omega$. The *union* of a bounded chain $\{F_i\}_{i\in\omega}$ of labeled frames F_i is defined as follows.

$$\cup_{i\in\omega} F_i := \langle \cup_{i\in\omega} W_i, \cup_{i\in\omega} R_i, \cup_{i\in\omega} S_i, \cup_{i\in\omega} \nu_i\rangle$$

Finally the Main Lemma can be formulated.

LEMMA 2.17 (Main Lemma)
Let **IL**X be an interpretability logic and let $\mathcal{C}$ be a (first or higher order) frame condition such that for any **IL**-frame F we have

$$F \models \mathcal{C} \Rightarrow F \models \mathsf{X}.$$

Let $\mathcal{D}$ be a finite set of sentences. Let $\mathcal{I}$ be a set of so-called *invariants* of labeled frames so that we have the following properties.

- $F \models \mathcal{I}^{\mathcal{U}} \Rightarrow F \models \mathcal{C}$, where $\mathcal{I}^{\mathcal{U}}$ is that part of $\mathcal{I}$ that is closed under bounded unions of labeled frames.
- $\mathcal{I}$ contains the following invariant: $xRy \rightarrow \exists\, A {\in} (\nu(y) \setminus \nu(x)) \cap \{\Box\neg D \mid D$ a subformula of some $B \in \mathcal{D}\}$.
- For any adequate labeled frame F, satisfying all the invariants, we have the following.
  - Any $\mathcal{D}$-problem of F can be eliminated by extending F in a way that conserves all invariants.
  - Any $\mathcal{D}$-deficiency of F can be eliminated by extending F in a way that conserves all invariants.

In case such a set of invariants $\mathcal{I}$ exists, we have that any **IL**X-labeled adequate frame F satisfying all the invariants can be extended to some labeled adequate **IL**X-frame $\hat{F}$ on which a truth-lemma with respect to $\mathcal{D}$ holds.

Moreover, if for any finite $\mathcal{D}$ that is closed under subformulas and single negations, a corresponding set of invariants $\mathcal{I}$ can be found as above and such that moreover $\mathcal{I}$ holds on any one-point labeled frame, we have that **IL**X is a complete logic.

The following two lemmata indicate how problems and deficiencies can be dealt with.

Lemma 2.18
Let Γ be a maximal **IL**X-consistent set such that $\neg(A \rhd B) \in \Gamma$. Then there exists a maximal **IL**X-consistent set Δ such that $\Gamma \prec_B \Delta \ni A, \Box\neg A$.

Lemma 2.19
Consider $C \rhd D \in \Gamma \prec_B \Delta \ni C$. There exists Δ' with $\Gamma \prec_B \Delta' \ni D, \Box\neg D$.

2.3 Modal Completeness of **IL**

In [20] the first application of the construction method was reproving the modal completeness of **IL**. Large parts of this new completeness for **IL** can be re-used in other proofs. We mention here the ingredients of the completeness proof of **IL** that will be re-used in this paper.

The Main Lemma will basically add bits and pieces to a model until all the necessary requirements are met. By adding a bit to a labeled frame a structure will arise that is almost a new frame but not quite yet. Those structures are called *quasi frames* and are defined below.

Definition 2.20
A *quasi-frame* G is a quadruple $\langle W, R, S, \nu\rangle$. Here W is a non-empty set of worlds, and R a binary relation on W. S is a set of binary relations on W indexed by elements of W. The ν is a labeling as defined on labeled frames. Critical cones and generalized cones are defined just in the same way as in the case of labeled frames. G should posess the following properties.

1. R is conversely well-founded
2. $yS_xz \to xRy$ & xRz
3. $xRy \to \nu(x) \prec \nu(y)$
4. $A \neq B \to \mathcal{G}_x^A \cap \mathcal{G}_x^B = \varnothing$
5. $y \in \mathcal{C}_x^A \to \nu(x) \prec_A \nu(y)$

Once the Main Lemma is around, the main effort in the proof of the modal completeness of **IL** lies in showing that each quasi frame can be extended to adequate labeled frame. We restate here this fact and hint at the main ingredients of the proof.

LEMMA 2.21 (**IL**-closure)
Let $G = \langle W, R, S, \nu \rangle$ be a quasi-frame. There is an adequate **IL**-frame F extending G. That is, $F = \langle W, R', S', \nu \rangle$ with $R \subseteq R'$ and $S \subseteq S'$.

PROOF. We define an *imperfection* on a quasi-frame F_n to be a tuple γ having one of the following forms.

(*i*) $\gamma = \langle 0, a, b, c \rangle$ with $F_n \models aRbRc$ but $F_n \not\models aRc$
(*ii*) $\gamma = \langle 1, a, b \rangle$ with $F_n \models aRb$ but $F_n \not\models bS_ab$
(*iii*) $\gamma = \langle 2, a, b, c, d \rangle$ with $F_n \models bS_acS_ad$ but not $F_n \models bS_ad$
(*iv*) $\gamma = \langle 3, a, b, c \rangle$ with $F_n \models aRbRc$ but $F_n \not\models bS_ac$

Now let us start with a quasi-frame $G = \langle W, R, S, \nu \rangle$. We will define a chain of quasi-frames. Every new element in the chain will have at least one imperfection less than its predecessor. The union will have no imperfections at all. It will be our required adequate **IL**-frame. ∎

3 The Logic ILM

Let us first recall the principle M, also called Montagna's principle.

$$\mathsf{M}: \quad A \rhd B \to A \wedge \Box C \rhd B \wedge \Box C$$

The modal logic **IL**M is of importance because it is the interpretability logic of theories like Peano Arithmetic.

THEOREM 3.1 (Berarducci [3], Shavrukov [24])
If T is an essentially reflexive theory, then **IL**(T) = **IL**M.

The modal completeness of **IL**M was proved by de Jongh and Veltman in [8]. In this section we will reprove the modal completeness of the logic **IL**M via the Main Lemma. This is done in 3.1 and 3.2. In 3.3 the new completeness proof is used to obtain some new results on admissible rules of **IL**M.

The general approach to the new completeness proof of **IL**M is not much different from the completeness proof for **IL**. The novelty consists of incorporating the **IL**M frame condition, that is, whenever yS_xzRu holds, we should also have yRu. In this case, adequacy imposes $\nu(y) \prec \nu(u)$.

Thus, whenever we introduce an S_x relation, when eliminating a deficiency, we should keep in mind that in a later stage, this S_x can activate the **IL**M frame condition. It turns out to be sufficient to demand $\nu(y) \subseteq_\Box \nu(z)$ whenever ySz. Also, we should do some additional book keeping as to keep our critical cones fit to our purposes.

3.1 Preparations

We start by defining a frame condition for **IL**M.

DEFINITION 3.2
An **IL**M-frame is a frame such that $yS_xzRu \to yRu$ holds on it. A(n adequate) labeled **IL**M-frame is a labeled **IL**M-frame on which $yS_xz \to \nu(y) \subseteq_\Box \nu(z)$ holds. We call $yS_xzRu \to yRu$ the frame condition of **IL**M.

The next lemma tells us that the frame condition of **IL**M, indeed characterizes the frames of **IL**M.

LEMMA 3.3
$F \models \forall x, y, u, v\ (yS_xuRv \to yRv) \Leftrightarrow F \models \mathbf{IL}\mathsf{M}$

We will now introduce a notion of a quasi-**IL**M-frame and a corresponding closure lemma. In order to get an **IL**M-closure lemma in analogy with Lemma 2.21 we need to introduce a technicality.

DEFINITION 3.4
The A-critical $\mathcal{M}$-cone of x, we write $\mathcal{M}_x^A$, is defined inductively as follows.

- $xR^Ay \to y \in \mathcal{M}_x^A$
- $y \in \mathcal{M}_x^A\ \&\ yRz \to z \in \mathcal{M}_x^A$
- $y \in \mathcal{M}_x^A\ \&\ yS_xz \to z \in \mathcal{M}_x^A$
- $y \in \mathcal{M}_x^A\ \&\ yS^{\mathsf{tr}}uRv \to v \in \mathcal{M}_x^A$

DEFINITION 3.5
A quasi-frame is a quasi-**IL**M-frame if[2] the following properties hold.

- $R^{\mathsf{tr}}; S^{\mathsf{tr}}$ is conversely well-founded[3]
- $yS_xz \to \nu(y) \subseteq_\Box \nu(z)$
- $y \in \mathcal{M}_x^A \Rightarrow \nu(x) \prec_A \nu(y)$

It is easy to see that $\mathcal{C}_x^A \subseteq \mathcal{M}_x^A \subseteq \mathcal{G}_x^A$. Thus we have that $A \neq B \to \mathcal{M}_x^A \cap \mathcal{M}_x^B = \varnothing$. Also, it is clear that if F is an **IL**M-frame, then $F \models \mathcal{M}_x^A = \mathcal{C}_x^A$. Actually we have that a quasi-**IL**M-frame F is an **IL**M-frame iff $F \models \mathcal{M}_x^A = \mathcal{C}_x^A$.

LEMMA 3.6 (**IL**M-closure)
Let $G = \langle W, R, S, \nu\rangle$ be a quasi-**IL**M-frame. There is an adequate **IL**M-frame F extending G. That is, $F = \langle W, R', S', \nu\rangle$ with $R \subseteq R'$ and $S \subseteq S'$.

PROOF. The proof is very similar to that of Lemma 2.21. As a matter of fact, we will use large parts of the latter proof in here. For quasi-**IL**M-frames we also define the notion of an imperfection. An *imperfection* on a quasi-**IL**M-frame F_n is a tuple γ that is either an imperfection on the quasi-frame F_n, or it is a tuple of the form

$$\gamma = \langle 4, a, b, c, d\rangle \text{ with } F_n \models bS_acRd \text{ but } F_n \not\models bRd.$$

[2] By R^{tr} we denote the transitive closure of R, inductively defined as the smallest set such that $xRy \to xR^{\mathsf{tr}}y$ and $\exists z\ (xR^{\mathsf{tr}}z \wedge zR^{\mathsf{tr}}y) \to xR^{\mathsf{tr}}y)$. Similarly we define S^{tr}. The ; is the composition operator on relations. Thus, for example, $y(R^{\mathsf{tr}}; S)z$ iff there is a u such that $yR^{\mathsf{tr}}u$ and uSz. Recall that uSv iff uS_xv for some x. In the literature one often also uses the $\circ$ notation, where $xR \circ Sy$ iff $\exists z\ xSzRy$. Note that $R^{\mathsf{tr}}; S^{\mathsf{tr}}$ is conversely well-founded iff $R^{\mathsf{tr}} \circ S^{\mathsf{tr}}$ is conversely well-founded.

[3] In the case of quasi-frames we did not need a second order frame condition. We could use the second order frame condition of **IL** via $yS_xz \to xRy\ \&\ xRz$. Such a trick seems not to be available here.

As in the closure proof for quasi-frames, we define a chain of quasi-**IL**M-frames. Each new frame in the chain will have at least one imperfection less than its predecessor. We only have to consider the new imperfections, in which case we define

$$F_{n+1} := \langle W_n, R_n \cup \{\langle b, d\rangle\}, S_n, \nu_n\rangle.$$

We now see by an easy but elaborate induction that every F_n is a quasi-**IL**M-frame. Again, this boils down to checking that at each of (i)-(v), all the eight properties from Definition 3.5 are preserved.

During the closure process, the critical cones do change. However, the critical $\mathcal{M}$-cones are invariant. Thus, it is useful to prove

$$8'.\ \ F_{n+1} \models y \in \mathcal{M}^A_x \text{ iff } F_n \models y \in \mathcal{M}^A_x.$$

Our induction is completely straightforward. As an example we shall see that $8'$ holds in Case (i): We have eliminated an imperfection concerning the transitivity of the R relation and $F_{n+1} := \langle W_n, R_n \cup \{\langle a, c\rangle\}, S_n, \nu_n\rangle$.

To see that $8'$ holds, we reason as follows. Suppose $F_{n+1} \models y \in \mathcal{M}^A_x$. Thus $\exists z_1, \ldots, z_l\ (0 \leq l)$ with[4] $F_{n+1} \models xR^A z_1(S_x \cup R \cup (S^{\mathsf{tr}}; R))z_2, \ldots, z_l(S_x \cup R \cup (S^{\mathsf{tr}}; R))y$. We transform the sequence $z_1, \ldots, z_l$ into a sequence $u_1, \ldots, u_m$ $(0 \leq m)$ in the following way. Every occurrence of aRc in $z_1, \ldots, z_l$ is replaced by $aRbRc$. In case that for some $n < l$ we have $z_n S^{\mathsf{tr}} aRc = z_{n+1}$, we replace z_n, z_{n+1} by z_n, b, c and thus $z_n(S^{\mathsf{tr}}; R)bRc$. We leave the rest of the sequence $z_1, \ldots, z_l$ unchanged. Clearly $F_n \models xR^A u_1(S_x \cup R \cup (S^{\mathsf{tr}}; R))u_2, \ldots, u_m(S_x \cup R \cup (S^{\mathsf{tr}}; R))y$, whence $F_n \models y \in \mathcal{M}^A_x$.

We shall include one more example for Case (v): We have eliminated an imperfection concerning the **IL**M frame-condition and $F_{n+1} := \langle W_n, R_n \cup \{\langle b, d\rangle\}, S_n, \nu_n\rangle$. To see the conversely well-foundedness of R, we reason as follows. Suppose for a contradiction that there is an infinite sequence such that $F_{n+1} \models x_1 R x_2 R \ldots$. We now get an infinite sequence $y_1, y_2, \ldots$ by replacing every occurrence of bRd in $x_1, x_2, \ldots$ by $bS_a cRd$ and leaving the rest unchanged. If there are infinitely many S_a-transitions in the sequence $y_1, y_2, \ldots$ (note that there are certainly infinitely many R-transitions in $y_1, y_2, \ldots$), we get a contradiction with our assumption that $R^{\mathsf{tr}}; S^{\mathsf{tr}}$ is conversely well-founded on F_n. In the other case we get a contradiction with the conversely well-foundedness of R on F_n.

Once we have seen that indeed, every F_n is a quasi-**IL**M-frame, it is not hard to see that $F := \cup_{i\in\omega} F_i$ is the required adequate **IL**M-frame. To this extend we have to check a list of properties $(a.)$-$(n.)$. The properties $(a.)$-$(l.)$ are as in the proof of Lemma 2.21.

The one exception is Property $(d.)$. To see $(d.)$, the conversely well-foundedness of R, we prove by induction on n that $F_n \models xRy$ iff $F_0 \models x(S^{\mathsf{tr,refl}}; R^{\mathsf{tr}})y$. Thus, a hypothetical infinite sequence $F \models x_0 R x_1 R x_2 R \ldots$ defines an infinite sequence $F_0 \models x_0(S^{\mathsf{tr,refl}}; R^{\mathsf{tr}})x_1(S^{\mathsf{tr,refl}}; R^{\mathsf{tr}})x_2 \ldots$, which contradicts either the conversely well-foundedness of R or of $S^{\mathsf{tr}}; R^{\mathsf{tr}}$ on F_0.

The only new properties in this list are $(m.)$: $uS_x vRw \rightarrow uRw$ and $(n.)$: $yS_x z \rightarrow \nu(y) \subseteq_\Box \nu(z)$, but they are easily seen to hold on F. ∎

[4]The union operator on relations can just be seen as the set-theoretical union. Thus, for example, $y(S_x \cup R)z$ iff $yS_x z$ or yRz.

Again do we note that the closure obtained in Lemma 3.6 is unique. Thus we can refer to the **IL**M-closure of a quasi-**IL**M-frame. All the information about the labels can be dropped in Definition 3.5 and Lemma 3.6 to obtain a lemma about regular **IL**M-frames.

Corollary 3.7
Let $\mathcal{D}$ be a finite set of sentences, closed under subformulas and single negations. Let $G = \langle W, R, S, \nu\rangle$ be a quasi-**IL**M-frame on which

$$xRy \to \exists\, A{\in}((\nu(y) \setminus \nu(x)) \cap \{\Box D \mid D \in \mathcal{D}\}) \quad (*)$$

holds. Property $(*)$ does also hold on the **IL**-closure F of G.

Proof. The proof is as the proof of Corollary 5.3 from [20]. We only need to remark on Case (v): If bS_acRd, we have $\nu(b) \subseteq_\Box \nu(c)$. Thus, $A \in ((\nu(d) \setminus \nu(c)) \cap \{\Box D \mid D \in \mathcal{D}\})$ implies $A \notin \nu(b)$. ∎

The final lemma in our preparations is a lemma that is needed to eliminate deficiencies properly.

Lemma 3.8
Let Γ and Δ be maximal **IL**M-consistent sets. Consider $C \rhd D \in \Gamma \prec_B \Delta \ni C$. There exists a maximal **IL**M-consistent set Δ' with $\Gamma \prec_B \Delta' \ni D, \Box\neg D$ and $\Delta \subseteq_\Box \Delta'$.

Proof. By compactness and by commutation of boxes and conjunctions, it is sufficient to show that for any formula $\Box E \in \Delta$ there is a Δ'' with $\Gamma \prec_B \Delta'' \ni D \wedge \Box E \wedge \Box\neg D$. As $C \rhd D$ is in the maximal **IL**M-consistent set Γ, also $C \wedge \Box E \rhd D \wedge \Box E \in \Gamma$. Clearly $C \wedge \Box E \in \Delta$, whence, by Lemma 2.19 we find a Δ'' with $\Gamma \prec_B \Delta'' \ni D \wedge \Box E \wedge \Box(\neg D \vee \neg\Box E)$. As **IL**M $\vdash \Box E \wedge \Box(\neg D \vee \neg \Box E) \to \Box\neg D$, we see that also $D \wedge \Box E \wedge \Box\neg D \in \Delta''$. ∎

3.2 Completeness

Theorem 3.9
ILM is a complete logic.

Proof. **Frame Condition** In the case of **IL**M the frame condition is easy and well known, as expressed in Lemma 3.3.

Invariants Let $\mathcal{D}$ be a finite set of sentences closed under subformulas and single negations. We define a corresponding set of invariants.

$$\mathcal{I} := \begin{cases} xRy \to \exists\, A{\in}((\nu(y) \setminus \nu(x)) \cap \{\Box D \mid D \in \mathcal{D}\}) \\ uS_xvRw \to uRw \end{cases}$$

Elimination Thus, we consider an **IL**M-labeled frame $F := \langle W, R, S, \nu\rangle$ that satisfies the invariants.

Problems Any problem $\langle a, \neg(A \rhd B)\rangle$ of F will be eliminated in two steps.

1. Using Lemma 2.18 we can find a MCS Δ with $\nu(a) \prec_B \Delta \ni A, \Box\neg A$. We fix some $b \notin W$ and define
$$G' := \langle W \cup \{b\}, R \cup \{\langle a, b\rangle\}, S, \nu \cup \{\langle b, \Delta\rangle, \langle\langle a, b\rangle, B\rangle\}\rangle.$$
We now see that G' is a quasi-**IL**M-frame. Thus, we need to check the eight points from Definitions 3.5 and 2.20. We will comment on some of these points.
To see, for example, Point 4, $C \neq D \to \mathcal{G}_x^C \cap \mathcal{G}_x^D = \varnothing$, we reason as follows. First, we notice that $\forall x, y{\in}W\ [G' \models y \in \mathcal{G}_x^C \text{ iff } F \models y \in \mathcal{G}_x^C]$ holds for any C. Suppose $G' \models \mathcal{G}_x^C \cap \mathcal{G}_x^D \neq \varnothing$. If $G' \models b \notin \mathcal{G}_x^C \cap \mathcal{G}_x^D$, then also $F \models \mathcal{G}_x^C \cap \mathcal{G}_x^D \neq \varnothing$. As F is an **IL**M-frame, it is certainly a quasi-**IL**M-frame, whence $C = D$. If now $G' \models b \in \mathcal{G}_x^C \cap \mathcal{G}_x^D$, necessarily $G' \models a \in \mathcal{G}_x^C \cap \mathcal{G}_x^D$, whence $F \models a \in \mathcal{G}_x^C \cap \mathcal{G}_x^D$ and $C = D$.
To see Requirement 8, $y \in \mathcal{M}_x^E \to \nu(x) \prec_E \nu(y)$, we reason as follows. Again, we first note that $\forall x, y{\in}W\ [G' \models y \in \mathcal{M}_x^C \text{ iff } F \models y \in \mathcal{M}_x^C]$ holds for any C. We only need to consider the new element, that is, $b \in \mathcal{M}_x^E$. If $x = a$ and $E = B$, we get the property by choice of $\nu(b)$.
For $x \neq a$, we consider two cases. Either $a \in \mathcal{M}_x^E$ or $a \notin \mathcal{M}_x^E$. In the first case, we get by the fact that F is a labeled **IL**M-frame $\nu(x) \prec_E \nu(a)$. But $\nu(a) \prec \nu(b)$, whence $\nu(x) \prec_E \nu(b)$. In the second necessarily for some $a' \in \mathcal{M}_x^E$ we have $a' S^{\mathsf{tr}} a$. But now $\nu(a') \subseteq_\Box \nu(a)$. Clearly $\nu(x) \prec_E \nu(a') \subseteq_\Box \nu(a) \prec \nu(b) \to \nu(x) \prec_E \nu(b)$.
2. With Lemma 3.6 we extend G' to an adequate labeled **IL**M-frame G. It is now obvious that both of the invariants hold on G. The first one holds due to Corollary 3.7. The other is just included in the definition of **IL**M-frames. Obviously, $\langle a, \neg(A \rhd B)\rangle$ is not a problem any more in G.

Deficiencies. Again, any deficiency $\langle a, b, C \rhd D\rangle$ in F will be eliminated in two steps.

1. We first define B to be the formula such that $b \in \mathcal{C}_a^B$. If such a B does not exist, we take B to be $\bot$. Note that if such a B does exist, it must be unique by Property 4 of Definition 2.20. By Lemma 2.8, or just by the fact that F is an **IL**M-frame, we have that $\nu(a) \prec_B \nu(b)$.
By Lemma 3.8 we can now find a Δ' such that $\nu(a) \prec_B \Delta' \ni D, \Box\neg D$ and $\nu(b) \subseteq_\Box \Delta'$. We fix some $c \notin W$ and define
$$G' := \langle W, R \cup \{\langle a, c\rangle\}, S \cup \{\langle a, b, c\rangle\}, \nu \cup \{\langle c, \Delta'\rangle\}\rangle.$$
To see that G' is indeed a quasi-**IL**M-frame, again eight properties should be checked. But all of these are fairly routine.
For Property 4 it is good to remark that, if $c \in \mathcal{G}_x^A$, then necessarily $b \in \mathcal{G}_x^A$ or $a \in \mathcal{G}_x^A$.
To see Property 8, we reason as follows. We only need to consider $c \in \mathcal{M}_x^A$. This is possible if $x = a$ and $b \in \mathcal{M}_a^A$, or if for some $y \in \mathcal{M}_x^A$ we have $y S^{\mathsf{tr}} a$, or if $a \in \mathcal{M}_x^A$. In the first case, we get that $b \in \mathcal{M}_a^A$, and thus also $b \in \mathcal{C}_a^A$ as F is an **IL**M-frame. Thus, by Property 4, we see that $A = B$. But Δ' was chosen such that $\nu(a) \prec_B \Delta'$. In the second case we see that $\nu(x) \prec_A \nu(y) \subseteq_\Box \nu(a) \prec \nu(c)$ whence $\nu(x) \prec_A \nu(c)$. In the third case we have $\nu(x) \prec_A \nu(a) \prec \nu(c)$, whence $\nu(x) \prec_A \nu(c)$.

2. Again, G' is closed off under the frame conditions with Lemma 3.6. Clearly, $\langle a, b, C \rhd D\rangle$ is not a deficiency on G.

Rounding up One of our invariants is just the **IL**M frame condition. Clearly this invariant is preserved under taking unions of bounded chains. The closure satisfies the invariants. ∎

3.3 Admissible rules

With the completeness at hand, a lot of reasoning about **IL**M gets easier. This holds in particular for derived/admissible rules of **IL**M. In the following lemma, we will use the completeness theorem to obtain models. Most of the times these models will be glued above a fresh new world to obtain new models with the desired properties.

LEMMA 3.10

(i) $\mathbf{IL}\mathsf{M} \vdash \Box A \Leftrightarrow \mathbf{IL}\mathsf{M} \vdash A$

(ii) $\mathbf{IL}\mathsf{M} \vdash \Box A \vee \Box B \Leftrightarrow \mathbf{IL}\mathsf{M} \vdash \Box A$ or $\mathbf{IL}\mathsf{M} \vdash \Box B$

(iii) $\mathbf{IL}\mathsf{M} \vdash A \rhd B \Leftrightarrow \mathbf{IL}\mathsf{M} \vdash A \to B \vee \Diamond B$.

(iv) $\mathbf{IL}\mathsf{M} \vdash A \rhd B \Leftrightarrow \mathbf{IL}\mathsf{M} \vdash \Diamond A \to \Diamond B$

(v) Let A_i be formulae such that $\mathbf{IL}\mathsf{M} \nvdash \neg A_i$. Then $\mathbf{IL}\mathsf{M} \vdash \bigwedge \Diamond A_i \to A \rhd B \Leftrightarrow \mathbf{IL}\mathsf{M} \vdash A \rhd B$.

(vi) $\mathbf{IL}\mathsf{M} \vdash A \vee \Diamond A \Leftrightarrow \mathbf{IL}\mathsf{M} \vdash \Box\bot \to A$

(vii) $\mathbf{IL}\mathsf{M} \vdash \top \rhd A \Leftrightarrow \mathbf{IL}\mathsf{M} \vdash \Box\bot \to A$

PROOF. (i). $\mathbf{IL}\mathsf{M} \vdash A \Rightarrow \mathbf{IL}\mathsf{M} \vdash \Box A$ by necessitation. Now suppose $\mathbf{IL}\mathsf{M} \vdash \Box A$. We want to see $\mathbf{IL}\mathsf{M} \vdash A$. Thus, we take an arbitrary model $M = \langle W, R, S, \Vdash\rangle$ and world $m \in M$. If there is an m_0 with $M \models m_0 R m$, then $M, m_0 \Vdash \Box A$, whence $M, m \Vdash A$. If there is no such m_0, we define (we may assume $m_0 \notin W$)

$$\begin{aligned} M' := \quad & \langle W \cup \{m_0\}, R \cup \{\langle m_0, w\rangle \mid w \in W\}, \\ & S \cup \{\langle m_0, x, y\rangle \mid \langle x, y\rangle \in R \text{ or } x{=}y \in W\}, \Vdash\rangle. \end{aligned}$$

Clearly, M' is an **IL**M-model too (the **IL**M frame conditions in the new cases follows from the transitivity of R), whence $M', m_0 \Vdash \Box A$ and thus $M', m \Vdash A$. By the construction of M' and by Lemma 2.5 we also get $M, m \Vdash A$.

(ii)."$\Leftarrow$" is easy. For the other direction we assume $\mathbf{IL}\mathsf{M} \nvdash \Box A$ and $\mathbf{IL}\mathsf{M} \nvdash \Box B$ and set out to prove $\mathbf{IL}\mathsf{M} \nvdash \Box A \vee \Box B$. By our assumption and by completeness, we find $M_0, m_0 \Vdash \Diamond\neg A$ and $M_1, m_1 \Vdash \Diamond\neg B$. We define (for some $r \notin W_0 \cup W_1$)

$$\begin{aligned} M := \quad & \langle W_0 \cup W_1 \cup \{r\}, R_0 \cup R_1 \cup \{\langle r, x\rangle \mid x \in W_0 \cup W_1\}, \\ & S_0 \cup S_1 \cup \{\langle r, x, y\rangle \mid x{=}y{\in}W_0 \cup W_1 \text{ or } \langle x, y\rangle{\in}R_0 \text{ or } \langle x, y\rangle{\in}R_1\}, \Vdash\rangle. \end{aligned}$$

Now, M is an **IL**M-model and $M, r \Vdash \Diamond\neg A \wedge \Diamond\neg B$ as is easily seen by Lemma 2.5. By soundness we get $\mathbf{IL}\mathsf{M} \nvdash \Box A \vee \Box B$.

(iii)."$\Leftarrow$" goes as follows. $\vdash A \to B \vee \Diamond B \Rightarrow \vdash \Box(A \to B \vee \Diamond B) \Rightarrow \vdash A \rhd B \vee \Diamond B \Rightarrow \vdash A \rhd B$. For the other direction, suppose that $\nvdash A \to B \vee \Diamond B$. Thus, we can find a model $M = \langle W, R, S, \Vdash\rangle$ and $m \in M$ with $M, m \Vdash A \wedge \neg B \wedge \Box\neg B$. We now define (with $r \notin W$)

$$\begin{array}{rl} M' := & \langle W \cup \{r\}, R \cup \{\langle r, x\rangle \mid x{=}m \text{ or } \langle m, x\rangle \in R\}, \\ & S \cup \{\langle r, x, y\rangle \mid (x{=}y \text{ and } (\langle m, x\rangle{\in}R \text{ or } x{=}m)) \text{ or } \langle m, x\rangle, \langle x, y\rangle{\in}R\}, \Vdash\rangle. \end{array}$$

It is easy to see that M' is an **IL**M-model. By Lemma 2.5 we see that $M', x \Vdash \varphi$ iff $M, x \Vdash \varphi$ for $x \in W$. It is also not hard to see that $M', r \Vdash \neg(A \rhd B)$. For, we have $rRm \Vdash A$. By definition, $mS_r y \to (m{=}y \vee mRy)$ whence $y \nVdash B$.

(*iv*). By the J4 axiom, we get one direction for free. For the other direction we reason as follows. Suppose **IL**M $\nvdash A \rhd B$. Then we can find a model $M = \langle W, R, S, \Vdash\rangle$ and a world l such that $M, l \Vdash \neg(A \rhd B)$. As $M, l \vdash \neg(A \rhd B)$, w can find some $m \in M$ with $lRm \Vdash A \wedge \neg B \wedge \Box\neg B$. We now define (with $r \notin W$)

$$\begin{array}{rl} M' := & \langle W \cup \{r\}, R \cup \{\langle r, x\rangle \mid x{=}m \text{ or } \langle m, x\rangle \in R\}, \\ & S \cup \{\langle r, x, y\rangle \mid (x{=}y \text{ and } (\langle m, x\rangle{\in}R \text{ or } x{=}m)) \text{ or } \langle m, x\rangle, \langle x, y\rangle{\in}R\}, \Vdash\rangle. \end{array}$$

It is easy to see that M' is an **IL**M-model. Lemma 2.5 and general knowledge about **IL**M tells us that the generated submodel from l is a witness to the fact that **IL**M $\nvdash \Diamond A \to \Diamond B$.[5]

(*v*). The "$\Leftarrow$" direction is easy. For the other direction we reason as follows.[6]

We assume that $\nvdash A \rhd B$ and set out to prove $\nvdash \bigwedge \Diamond A_i \to A \rhd B$. As $\nvdash A \rhd B$, we can find $M, r \Vdash \neg(A \rhd B)$. By Lemma 2.5 we may assume that r is a root of M. For all i, we assumed $\nvdash \neg A_i$, whence we can find rooted models $M_i, r_i \Vdash A_i$. As in the other cases, we define a model $\tilde{M}$ that arises by gluing r under all the r_i. Clearly we now see that $\tilde{M}, r \Vdash \bigwedge \Diamond A_i \wedge \neg(A \rhd B)$.

(*vi*). First, suppose that **IL**M $\vdash \Box\bot \to A$. Then, from **IL**M $\vdash \Box\bot \vee \Diamond\top$, the observation that **IL**M $\vdash \Diamond\top \leftrightarrow \Diamond\Box\bot$ and our assumption, we get **IL**M $\vdash A \vee \Diamond A$.

For the other direction, we suppose that **IL**M $\nvdash \Box\bot \to A$. Thus, we have a counter model M and some $m \in M$ with $m \Vdash \Box\bot, \neg A$. Clearly, at the submodel generated from m, that is, a single point, we see that $\neg A \wedge \Box\neg A$ holds. Consequently **IL**M$\neg \vdash A \vee \Diamond A$.

(*vii*). This follows immediately from (*vi*) and (*iii*).

∎

Note that, as **IL**M is conservative over **GL**, all of the above statements not involving $\rhd$ also hold for **GL**. The same holds for derived statements. For example, from Lemma 3.10 we can combine (*iii*) and (*iv*) to obtain **IL**M $\vdash A \to B \vee \Diamond B \Leftrightarrow$ **IL**M $\vdash \Diamond A \to \Diamond B$. Consequently, the same holds true for **GL**.

3.4 Decidability

It is well known that **IL**M has the finite model property. It is not hard to re-use worlds in the presented construction method so that we would end up with a finite counter model. Actually, this is precisely what has been done in [18]. In that paper, one of the invariants was "there are no deficiencies". We have chosen not to include this

[5] This proof is similar to the proof of (*iii*). However, it is not the case that one of the two follows easily from the other.

[6] By a similar reasoning we can prove $\vdash \bigwedge \neg(C_i \rhd D_i) \to A \rhd B \Leftrightarrow \vdash A \rhd B$.

invariant in our presentation, as this omission simplifies the presentation. Moreover, for our purposes the completeness without the finite model property obtained via our construction method suffices.

Our purpose to include a new proof of the well known completeness of **IL**M is twofold. On the one hand the new proof serves well to expose the construction method. On the other hand, it is an indispensable ingredient in proving Theorem 4.5.

4 Essentially Σ_1-sentences of ILM

In this section we will answer the question which modal interpretability sentences are in theories T provably Σ_1 for any realization. We call these sentences essentially Σ_1-sentences. We shall answer the question only for T an essentially reflexive theory.

This question has been solved for provability logics by Visser in [31]. In [7], de Jongh and Pianigiani gave an alternative solution by using the logic **IL**M. Our proof shall use their proof method.

We will perform our argument fully in **IL**M. It is very tempting to think that our result would be an immediate corollary from for example [11], [17] or [16]. This would be the case, if a construction method were worked out for the logics from these respective papers. In [11] a sort of a construction method is indeed worked out. This construction method should however be a bit sharpened to suit our purposes. Moreover that sharpening would essentially reduce to the solution we present here.

4.1 Model construction

Throughout this subsection, unless mentioned otherwise, T will be an essentially reflexive recursively enumerable arithmetical theory. By Theorem 3.1 we thus know that **IL**(T) = **IL**M. Let us first say more precisely what we mean by an essentially Σ_1-sentence.

Definition 4.1
A modal sentence φ is called an essentially Σ_1-sentence with respect to a theory T, if $\forall *\ \varphi^* \in \Sigma_1(T)$. Likewise, a formula φ is essentially Δ_1 if $\forall *\ \varphi^* \in \Delta_1(T)$

If φ is an essentially Σ_1-formula for T we will also write $\varphi \in \Sigma_1(T)$. Analogously for $\Delta_1(T)$. For the rest of this section, T will always be a theory that has **IL**M as its interpretability logic thereby making explicit reference to T unnecessary as we shall see.

Theorem 4.2
Modulo modal logical equivalence, there exist just two essentially Δ_1-formulas in the language of **IL**M. That is, $\Delta_1(T) = \{\top, \bot\}$.

Proof. Let φ be a modal formula. If $\varphi \in \Delta_1(T)$, then, by provably Σ_1-completeness, both $\forall *\ T \vdash \delta^* \to \Box\delta^*$ and $\forall *\ T \vdash \neg\delta^* \to \Box\neg\delta^*$. Consequently $\forall *\ T \vdash \Box\delta^* \vee \Box\neg\delta^*$. Thus, $\forall *\ T \vdash (\Box\delta \vee \Box\neg\delta)^*$ whence **IL**M $\vdash \Box\delta \vee \Box\neg\delta$. By Lemma 3.10 we see that **IL**M $\vdash \delta$ or **IL**M $\vdash \neg\delta$. ∎

We proved Theorem 4.2 for the interpretability logic of essentially reflexive theories. It is not hard to see that the theorem also holds for finitely axiomatizable theories. The only ingredients that we need to prove this are [**IL**P $\vdash \Box A \vee \Box B$ iff. **IL**P $\vdash \Box A$

or $\mathbf{IL}\mathsf{P} \vdash \Box B$] and [$\mathbf{IL}\mathsf{P} \vdash \Box A$ iff. $\mathbf{IL}\mathsf{P} \vdash A$]. As these two admissible rules also hold for **GL**, we see that Theorem 4.2 also holds for **GL**.

The following lemma is the only arithmetical ingredient in our classification of the essentially Σ_1 formulas in the language of $\mathbf{IL}\mathsf{M}$.

LEMMA 4.3
If $\varphi \in \Sigma_1(T)$, then, for any p and q, we have $\mathbf{IL}\mathsf{M} \vdash p \rhd q \to p \wedge \varphi \rhd q \wedge \varphi$.

Before we come to prove the main theorem of this section, we first need an additional lemma.

LEMMA 4.4
Let Δ_0 and Δ_1 be maximal $\mathbf{IL}\mathsf{M}$-consistent sets. There is a maximal $\mathbf{IL}\mathsf{M}$-consistent set Γ such that $\Gamma \prec \Delta_0, \Delta_1$.

PROOF. We show that $\Gamma' := \{\Diamond A \mid A \in \Delta_0\} \cup \{\Diamond B \mid B \in \Delta_1\}$ is consistent. Assume for a contradiction that Γ' were not consistent. Then, by compactness, for finitely many A_i and B_j,

$$\bigwedge_{A_i \in \Delta_0} \Diamond A_i \wedge \bigwedge_{B_j \in \Delta_1} \Diamond B_j \vdash \bot$$

or equivalently

$$\vdash \bigvee_{A_i \in \Delta_0} \Box\neg A_i \vee \bigvee_{B_j \in \Delta_1} \Box\neg B_j.$$

By Lemma 3.10 we see that then either $\vdash \neg A_i$ for some i, or $\vdash \neg B_j$ for some j. This contradicts the consistency of Δ_0 and Δ_1. ∎

With this lemma and by postponing the hard work to Subsectionsubs:sigmaLemma we can now prove the main theorem of this section.

THEOREM 4.5
$\varphi \in \Sigma_1(T) \Leftrightarrow \mathbf{IL}\mathsf{M} \vdash \varphi \leftrightarrow \bigvee_{i \in I} \Box C_i$ for some $\{C_i\}_{i \in I}$.

PROOF. Let φ be a formula that is not equivalent to a disjunction of $\Box$-formulas. According to Lemma 4.7 we can find MCS's Δ_0 and Δ_1 with $\varphi \in \Delta_0 \subseteq_\Box \Delta_1 \ni \neg\varphi$. By Lemma 4.4 we find a $\Gamma \prec \Delta_0, \Delta_1$. We define:

$$G := \langle \{m_0, l, r\}, \{\langle m_0, l\rangle, \langle m_0, r\rangle\}, \{\langle m_0, l, r\rangle\}, \{\langle m_0, \Gamma\rangle, \langle l, \Delta_0\rangle, \langle r, \Delta_1\rangle\}\rangle.$$

We will apply a slightly generalized version of the main lemma to this quasi-$\mathbf{IL}\mathsf{M}$-frame G. The finite set $\mathcal{D}$ of sentences is the smallest set of sentences that contains φ and that is closed under taking subformulas and single negations. The invariants are the following.

$$\mathcal{I} := \begin{cases} xRy \wedge x \neq m_0 \to \exists\, A {\in} ((\nu(y) \setminus \nu(x)) \cap \{\Box D \mid D \in \mathcal{D}\}) \\ uS_x vRw \to uRw \end{cases}$$

In the proof of Theorem 3.9 we have seen that we can eliminate both problems and deficiencies while conserving the invariants. The main lemma now gives us an $\mathbf{IL}\mathsf{M}$-model M with $M, l \Vdash \varphi$, $M, r \Vdash \neg\varphi$ and $lS_{m_0}r$. We now pick two fresh variables p and q. We define p to be true only at l and q only at r. Clearly $m_0 \Vdash \neg(p \rhd q \to p \wedge \varphi \rhd q \wedge \varphi)$, whence by Lemma 4.3 we get $\varphi \notin \Sigma_1(T)$.

∎

For finitely axiomatized theories T, our theorem does not hold, as also $A \rhd B$ is T-essentially Σ_1. The following theorem says that in this case, $A \rhd B$ is under any T-realization actually equivalent to a special Σ_1-sentence.

Theorem 4.6
Let T be a finitely axiomatized theory. For all arithmetical formulae α, β there exists a formula ρ with

$$T \vdash \alpha \rhd_T \beta \leftrightarrow \Box_T \rho.$$

Proof. The proof is a direct corollary of the so-called FGH-theorem. (See [33] for an exposition of the FGH-theorem.) We take ρ satisfying the following fixed point equation.

$$T \vdash \rho \leftrightarrow ((\alpha \rhd_T \beta) \leq \Box_T \rho)$$

By the proof of the FGH-theorem, we now see that

$$T \vdash ((\alpha \rhd_T \beta) \vee \Box_T \bot) \leftrightarrow \Box_T \rho.$$

But clearly $T \vdash ((\alpha \rhd_T \beta) \vee \Box_T \bot) \leftrightarrow \alpha \rhd_T \beta$. ∎

4.2 The Σ-lemma

We can say that the proof of Theorem 4.5 contained three main ingredients; Firstly, the main lemma; Secondly the modal completeness theorem for **IL**M via the construction method and; Thirdly the Σ-lemma. In this subsection we will prove the Σ-lemma and remark that it is in a sense optimal.

Lemma 4.7
If φ is a formula not equivalent to a disjunction of $\Box$-formulas. Then there exist maximal **IL**X-consistent sets Δ_0, Δ_1 such that $\varphi \in \Delta_0 \subseteq_\Box \Delta_1 \ni \neg\varphi$.

Proof. As we shall see, the reasoning below holds not only for **IL**X, but for any extension of **GL**. We define

$$\begin{aligned} \Box_\vee &:= \{ \bigvee_{0 \leq i < n} \Box D_i \mid n \geq 0, \text{each } D_i \text{ an } \mathbf{IL}\mathsf{X}\text{-formula}\}, \\ \Box_{\mathrm{con}} &:= \{Y \subseteq \Box_\vee \mid \{\neg\varphi\} + Y \text{ is consistent and maximally such}\}. \end{aligned}$$

Let us first observe a useful property of the sets Y in $\Box_{\mathrm{con}}$.

$$\bigvee_{i=0}^{n-1} \sigma_i \in Y \Rightarrow \exists i{<}n \; \sigma_i \in Y. \tag{4.1}$$

To see this, let $Y \in \Box_{\mathrm{con}}$ and $\bigvee_{i=0}^{n-1} \sigma_i \in Y$. Then for each $i{<}n$ we have $\sigma_i \in \Box_\vee$ and for some $i{<}n$ we must have σ_i consistent with Y (otherwise $\{\neg\varphi\} + Y$ would prove $\bigwedge_{i=0}^{n-1} \neg\sigma_i$ and be inconsistent). And thus by the maximality of Y we must have that some σ_i is in Y. This establishes (4.1).

Claim 4.8
For some $Y \in \Box_{\mathrm{con}}$ the set

$$\{\varphi\} + \{\neg\sigma \mid \sigma \in \Box_\vee - Y\}$$

is consistent.

Proof. [Proof of the claim] Suppose the claim were false. We will derive a contradiction with the assumption that φ is not equivalent to a disjunction of $\Box$-formulas. If the claim is false, then we can choose for each $Y \in \Box_{\mathrm{con}}$ a finite set $Y^{\mathrm{fin}} \subseteq \Box_{\vee} - Y$ such that

$$\{\varphi\} + \{\neg\sigma \mid \sigma \in Y^{\mathrm{fin}}\} \tag{4.2}$$

is inconsistent. Thus, certainly for each $Y \in \Box_{\mathrm{con}}$

$$\vdash \varphi \rightarrow \bigvee_{\sigma \in Y^{\mathrm{fin}}} \sigma. \tag{4.3}$$

Now we will show that:

$$\{\neg\varphi\} + \{\bigvee_{\sigma \in Y^{\mathrm{fin}}} \sigma \mid Y \in \Box_{\mathrm{con}}\} \text{ is inconsistent.} \tag{4.4}$$

For, suppose (4.4) were not the case. Then for some $S \in \Box_{\mathrm{con}}$

$$\{\bigvee_{\sigma \in Y^{\mathrm{fin}}} \sigma \mid Y \in \Box_{\mathrm{con}}\} \subseteq S.$$

In particular we have $\bigvee_{\sigma \in S^{\mathrm{fin}}} \sigma \in S$. But for all $\sigma \in S^{\mathrm{fin}}$ we have $\sigma \notin S$. Now by (4.1) we obtain a contradiction and thus we have shown (4.4).

So we can select some finite $\Box_{\mathrm{con}}^{\mathrm{fin}} \subseteq \Box_{\mathrm{con}}$ such that

$$\vdash (\bigwedge_{Y \in \Box_{\mathrm{con}}^{\mathrm{fin}}} \bigvee_{\sigma \in Y^{\mathrm{fin}}} \sigma) \rightarrow \varphi. \tag{4.5}$$

By (4.3) we also have

$$\vdash \varphi \rightarrow \bigwedge_{Y \in \Box_{\mathrm{con}}^{\mathrm{fin}}} \bigvee_{\sigma \in Y^{\mathrm{fin}}} \sigma. \tag{4.6}$$

Combining (4.5) with (4.6) we get

$$\vdash \varphi \leftrightarrow \bigwedge_{Y \in \Box_{\mathrm{con}}^{\mathrm{fin}}} \bigvee_{\sigma \in Y^{\mathrm{fin}}} \sigma.$$

Bringing the right hand side of this equivalence in disjunctive normal form and distributing the $\Box$ over $\wedge$ we arrive at a contradiction with the assumption on φ. ∎

So, we have for some $Y \in \Box_{\mathrm{con}}$ that both the sets

$$\{\varphi\} + \{\neg\sigma \mid \sigma \in \Box_{\vee} - Y\} \tag{4.7}$$

$$\{\neg\varphi\} + Y \tag{4.8}$$

are consistent. The lemma follows by taking Δ_0 and Δ_1 extending (4.7) and (4.8) respectively. ∎

We have thus obtained $\varphi \in \Delta_0 \subseteq_{\Box} \Delta_1 \ni \neg\varphi$ for some maximal **IL**X-consistent sets Δ_0 and Δ_1. The relation $\subseteq_{\Box}$ between Δ_0 and Δ_1 is actually the best we can get among the relations on MCS's that we consider in this paper. We shall see that $\Delta_0 \prec \Delta_1$ is not possible to get in general.

It is obvious that that $p \wedge \Box p$ is not equivalent to a disjunction of $\Box$-formulas. Clearly $p \wedge \Box p \in \Delta_0 \prec \Delta_1 \ni \neg p \vee \Diamond\neg p$ is impossible. In a sense, this reflects the fact that there exist non trivial self-provers, as was shown by Kent ([21]), Guaspari ([12]) and Beklemishev ([2]). Thus, provable Σ_1-completeness, that is $T \vdash \sigma \to \Box\sigma$ for $\sigma \in \Sigma_1(T)$, can not substitute Lemma 4.3.

5 Self provers and Σ_1-sentences

A self prover is a sentence φ that implies its own provability. That is, a sentence for which $\vdash \varphi \to \Box\varphi$, or equivalently, $\vdash \varphi \leftrightarrow \varphi \wedge \Box\varphi$. Self provers have been studied intensively amongst others by Kent ([21]), Guaspari ([12]), de Jongh and Pianigiani ([7]). It is easy to see that any $\Sigma_1(T)$-sentence is indeed a self prover. We shall call such a self prover a *trivial self prover*.

In [12], Guaspari has shown that there are many non-trivial self provers around. The most prominent example is probably $p \wedge \Box p$. But actually, any formula φ will generate a self prover $\varphi \wedge \Box\varphi$, as clearly $\varphi \wedge \Box\varphi \to \Box(\varphi \wedge \Box\varphi)$.

Definition 5.1
A formula φ is called a trivial self prover generator, we shall write t.s.g., if $\varphi \wedge \Box\varphi$ is a trivial self prover. That is, if $\varphi \wedge \Box\varphi \in \Sigma_1(T)$.

Obviously, a trivial self prover is also a t.s.g. But there also exist other t.s.g.'s. The most prominent example is probably $\Box\Box p \to \Box p$. A natural question is to ask for an easy characterization of t.s.g.'s. In this section we will give such a characterization for **GL**. All results presented here are new results. In the rest of this section, $\vdash$ will stand for derivability in **GL**. We shall often write Σ instead of Σ_1.

We say that a formula ψ is Σ in **GL**, and write $\Sigma(\psi)$, if for any theory T which has **GL** as its provability logic, we have that $\forall *\ \psi^* \in \Sigma_1(T)$.

Theorem 5.2
We have that $\Sigma(\varphi \wedge \Box\varphi)$ in **GL** if and only if the following condition is satisfied.

For all formulae A_l, φ_l and C_m satisfying 1, 2 and 3 we have that $\vdash \varphi \wedge \Box\varphi \leftrightarrow \bigvee_m \Box C_m$. Here 1-3 are the following conditions.

1. $\vdash \varphi \leftrightarrow \bigvee_l(\varphi_l \wedge \Box A_l) \vee \bigvee_m \Box C_m$
2. $\nvdash \Box A_l \to \varphi$ for all l
3. φ_l is a non-empty conjunction of literals and $\Diamond$-formulas.

Proof. The $\Leftarrow$ direction is the easiest part. We can always find an equivalent of φ that satisfies 1, 2 and 3. Thus, by assumption, $\varphi \wedge \Box\varphi$ can be written as the disjunction of $\Box$-formulas and hence $\Sigma(\varphi \wedge \Box\varphi)$.

For the $\Rightarrow$ direction we reason as follows. Suppose we can find φ_l, A_l and C_m such that 1, 2 and 3 hold, but

$$\nvdash \varphi \wedge \Box\varphi \leftrightarrow \bigvee_m \Box C_m. \quad (*)$$

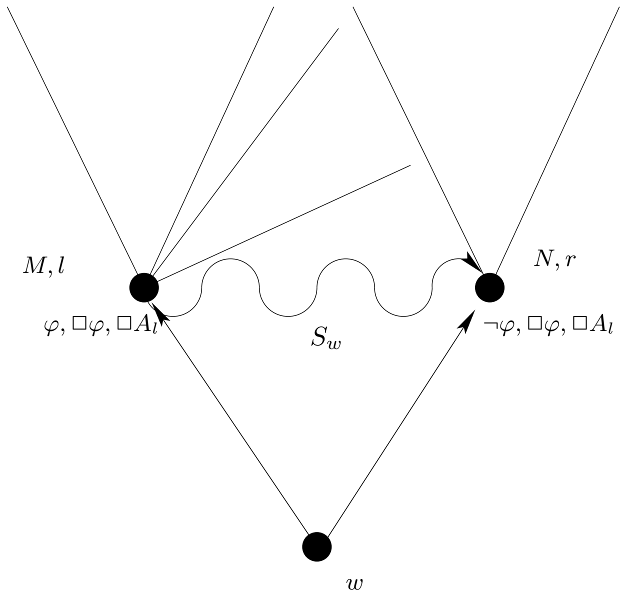

FIG. 1. T.s.g.'s

We can take now $T = \mathrm{PA}$ and reason as follows. As clearly $\vdash \bigvee\!\!\!\bigvee_m \Box C_m \to \varphi \wedge \Box\varphi$, our assumption $(*)$ reduces to $\nvdash \varphi \wedge \Box\varphi \to \bigvee\!\!\!\bigvee_m \Box C_m$. Consequently $\bigvee\!\!\!\bigvee_l(\varphi_l \wedge \Box A_l)$ can not be empty, and for some l and some rooted **GL**-model M, r with root r, we have $M, l \Vdash \Box A_l \wedge \varphi_l$.

We shall now see that $\nvdash \neg\varphi \wedge \Box\varphi \to \Diamond\neg A_l$. For, suppose for a contradiction that

$$\vdash \neg\varphi \wedge \Box\varphi \to \Diamond\neg A_l.$$

Then also $\vdash \Box A_l \to (\Box\varphi \to \varphi)$, whence $\vdash \Box A_l \to \Box(\Box\varphi \to \varphi) \to \Box\varphi$. And by $\Box A_l \to (\Box\varphi \to \varphi)$ again, we get $\vdash \Box A_l \to \varphi$ which contradicts 2. We must conclude that indeed $\nvdash \neg\varphi \wedge \Box\varphi \to \Diamond\neg A_l$, and thus we have a rooted tree model N, r for **GL** with $N, r \Vdash \neg\varphi, \Box\varphi, \Box A_l$.

We can now "glue" a world w below l and r, set $lS_w r$ and consider the smallest **IL**M-model extending this. We have depicted this construction in Figure 1. Let us also give a precise definition. If $M := \langle W_0, R_0, \Vdash_0\rangle$ and $N := \langle W_1, R_1, \Vdash_1\rangle$, then we define

$$\begin{aligned} L := \quad & \langle W_0 \cup W_1, R_0 \cup R_1 \cup \{\langle w, x\rangle \mid x \in W_0 \cup W_1\} \cup \{\langle l, y\rangle \mid N \models rRy\}, \\ & \{\langle w, l, r\rangle\} \cup \{\langle x, y, z\rangle \mid L \models xRyR^*z\}, \Vdash_0 \cup \Vdash_1\rangle. \end{aligned}$$

We observe that, by Lemma 2.5 $L, r \Vdash \Box\varphi \wedge \Box A_l \wedge \neg\varphi$ and $L \models rRx \Rightarrow L, x \Vdash \varphi \wedge A_l$. Also, if $L \models lRx$, then $L, x \Vdash \varphi \wedge A_i$, whence $L, l \Vdash \Box\varphi \wedge \Box A_l$. As $M, l \Vdash \varphi_l$ and φ_l only contains literals and and diamond-formulas, we see that $L, l \Vdash \varphi_l$, whence $L, l \Vdash \varphi \wedge \Box\varphi$. As $L, r \Vdash \neg\varphi \wedge \Box\varphi$ we see that $L, w \Vdash \neg\Box(\varphi \wedge \Box\varphi)$.

As in the proof of Theorem 4.5, we can take some fresh p and q and define p to hold only at l and q to hold only at r. Now, clearly $w \nVdash p \rhd q \to p \wedge (\varphi \wedge \Box\varphi) \rhd q \wedge (\varphi \wedge \Box\varphi)$, whence, by Lemma 4.3 we conclude $\neg\Sigma(\varphi \wedge \Box\varphi)$. ∎

The above reasoning showed that $\Sigma(\varphi \wedge \Box\varphi)$ is not a sufficient condition for $\Sigma(\varphi)$ to hold. We shall see that even $\Sigma(\varphi \wedge \Box\varphi) \;\wedge\; \Sigma(\varphi \wedge \Box\neg\varphi)$ is not a sufficient condition for $\Sigma(\varphi)$ to hold.

Thus, to conclude this section, we remain in **GL** and shall settle the question for which φ we have that

$$\Sigma(\varphi \wedge \Box\varphi) \ \&\ \Sigma(\varphi \wedge \Box\neg\varphi) \Rightarrow \Sigma(\varphi). \quad (\dagger)$$

We shall see that this question is non-trivial and that it can be reduced to the characterization of t.s.g.'s. Again the easiest non-trivial example satisfying (†) will be $\Box\Box p \to \Box p$.

LEMMA 5.3

$$\text{For some (possibly empty) } \bigvee\!\!\!\bigvee_i \Box C_i \text{ we have } \vdash \varphi \wedge \Box\neg\varphi \leftrightarrow \bigvee\!\!\!\bigvee_i \Box C_i$$
$$\text{iff.}$$
$$\vdash \Box\bot \to \varphi \quad \text{or} \quad \vdash \neg\varphi$$

PROOF. For non-empty $\bigvee\!\!\!\bigvee_i \Box C_i$ we have the following.

$$\begin{array}{ll} \vdash \varphi \wedge \Box\neg\varphi \leftrightarrow \bigvee\!\!\!\bigvee_i \Box C_i & \Rightarrow \\ \vdash \Diamond(\varphi \wedge \Box\neg\varphi) \leftrightarrow \Diamond(\bigvee\!\!\!\bigvee_i \Box C_i) & \Rightarrow \\ \vdash \Diamond\varphi \leftrightarrow \Diamond\top & \Rightarrow \\ \vdash \Box\bot \to \varphi & \end{array}$$

Here, the final step in the proof comes from Lemma 3.10.

On the other hand, if $\vdash \Box\bot \to \varphi$, we see that $\vdash \neg\varphi \to \Diamond\top$ and thus $\Box\neg\varphi \to \Box\bot$, whence $\vdash \varphi \wedge \Box\neg\varphi \leftrightarrow \Box\bot$.

In case of the empty disjunction we get $\vdash \varphi \wedge \Box\neg\varphi \leftrightarrow \bot$. Then also $\vdash \Box\neg\varphi \to \neg\varphi$ and by Löb $\vdash \neg\varphi$. And conversely, if $\vdash \neg\varphi$, then $\vdash \varphi \wedge \Box\neg\varphi \leftrightarrow \bot$, and $\bot$ is just the empty disjunction.

The proof actually gives some additional information. If $\Sigma(\varphi \wedge \Box\neg\varphi)$ then either ($\vdash \neg\varphi$ and $\vdash (\varphi \wedge \Box\neg\varphi) \leftrightarrow \bot$), or ($\vdash \Box\bot \to \varphi$ and $\vdash (\varphi \wedge \Box\neg\varphi) \leftrightarrow \Box\bot$). ∎

LEMMA 5.4

$$\Sigma(\varphi \wedge \Box\varphi) \wedge \Sigma(\varphi \wedge \Box\neg\varphi) \Rightarrow \Sigma(\varphi)$$
$$\text{iff.}$$
$$\Sigma(\varphi \wedge \Box\varphi) \Rightarrow \Sigma(\varphi) \text{ or } \vdash \varphi \to \Diamond\top$$

PROOF. $\Uparrow$. Clearly, if $\Sigma(\varphi \wedge \Box\varphi) \Rightarrow \Sigma(\varphi)$, also $\Sigma(\varphi \wedge \Box\varphi) \wedge \Sigma(\varphi \wedge \Box\neg\varphi) \Rightarrow \Sigma(\varphi)$. Thus, suppose $\vdash \varphi \to \Diamond\top$, or put differently $\vdash \Box\bot \to \neg\varphi$. If now $\vdash \neg\varphi$, then clearly $\Sigma(\varphi)$, whence $\Sigma(\varphi \wedge \Box\varphi) \wedge \Sigma(\varphi \wedge \Box\neg\varphi) \Rightarrow \Sigma(\varphi)$, so, we may assume that $\nvdash \neg\varphi$. It is clear that now $\neg\Sigma(\varphi \wedge \Box\neg\varphi)$. For, suppose $\Sigma(\varphi \wedge \Box\neg\varphi)$, then by Lemma 5.3 we see $\vdash \Box\bot \to \varphi$, whence $\vdash \Diamond\top$. Quod non. Thus, $\vdash \Box\bot \to \neg\varphi \Rightarrow \neg\Sigma(\varphi \wedge \Box\neg\varphi)$ and thus certainly $\Sigma(\varphi \wedge \Box\varphi) \wedge \Sigma(\varphi \wedge \Box\neg\varphi) \Rightarrow \Sigma(\varphi)$.

$\Downarrow$. Suppose $\Sigma(\varphi \wedge \Box\varphi) \wedge \neg\Sigma(\varphi)$ and $\nvdash \Box\bot \to \neg\varphi$. To obtain our result, we only have to prove $\Sigma(\varphi \wedge \Box\neg\varphi)$.

As $\nvdash \Box\bot \to \neg\varphi$, also $\nvdash \neg\varphi \vee \Diamond\neg\varphi$. Thus, under the assumption that $\Sigma(\varphi \wedge \Box\varphi)$, we can find (a non-empty collection of) C_i with $\vdash \varphi \wedge \Box\varphi \leftrightarrow \bigvee\!\!\!\bigvee_i \Box C_i$. In this case, clearly $\vdash \Box\bot \to \bigvee\!\!\!\bigvee_i \Box C_i \to \varphi$, whence, by Lemma 5.3 we conclude $\Sigma(\varphi \wedge \Box\neg\varphi)$. ∎

References

[1] C. Areces, D. de Jongh, and E. Hoogland. The interpolation theorem for IL and ILP. In *Proceedings of AiML98. Advances in Modal Logic*, Uppsala. Sweden, October 1998. Uppsala University.

[2] L.D. Beklemishev. On the complexity of arithmetic interpretations of modal formulae. *Archive for Mathematical Logic*, 32:229–238, 1993.

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

[4] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. Camebridge University Press, 2001.

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

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

[7] D. de Jongh and D. Pianigiani. Solution of a problem of David Guaspari. *Studia Logica*, 1996.

[8] D.H.J. de Jongh and F. Veltman. Provability logics for relative interpretability. In *[22]*, pages 31–42, 1990.

[9] D.H.J. de Jongh and F. Veltman. Modal completeness of ILW. In J. Gerbrandy, M. Marx, M. Rijke, and Y. Venema, editors, *Essays dedicated to Johan van Benthem on the occasion of his 50th birthday*. Amsterdam University Press, Amsterdam, 1999.

[10] D.H.J. de Jongh and A. Visser. Explicit fixed points in interpretability logic. *Studia Logica*, 50:39–50, 1991.

[11] E. Goris. Extending ILM with an operator for Σ_1–ness. Illc prepublication series, University of Amsterdam, 2003.

[12] D. Guaspari. Sentences implying their own provability. *Journal of Symbolic Logic*, 48:777–789, 1983.

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

[14] R. Hirsch and I. Hodkinson. *Relation Algebras by Games*, volume 147 of *Studies in Logic*. Elsevier, North-Holland, 2002.

[15] I. Hodkinson, S. Mikulás, and Y. Venema. Axiomatizing complex algebras by games. *Algebra Universalis*, 46:455–478, 2001.

[16] K.N. Ignatiev. The provability logic of Σ_1-interpolability. *Annals of Pure and Applied Logic*, 64:1–25, 1993.

[17] G.K. Japaridze. The logic of the arithmetical hiearchy. *Annals of Pure and Applied Logic*, 66:89–112, 1994.

[18] J.J. Joosten. Towards the interpretability logic of all reasonable arithmetical theories. Master's thesis, University of Amsterdam, 1998.

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

[20] E. Goris. Modal Matters for Interpretability Logic. *Logic Journal of the Interest Group in Pure and Applied Logics*, 16: 371 - 412, August 2008.

[21] C.F Kent. The relation of A to $Prov\ulcorner!A\urcorner$ in the Lindenbaum sentence algebra. *Journal of Symbolic Logic*, 38:359–367, 1973.

[22] P.P. Petkov, editor. *Mathematical logic, Proceedings of the Heyting 1988 summer school in Varna, Bulgaria*. Plenum Press, Boston, 1990.

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

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

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

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

[27] V. Švejdar. Some independence results in interpretability logic. *Studia Logica*, 50:29–38, 1991.

[28] A. Visser. Preliminary notes on interpretability logic. Technical Report LGPS 29, Department of Philosophy, Utrecht University, 1988.

[29] A. Visser. Interpretability logic. In *[22]*, pages 175–209, 1990.

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

[31] A. Visser. A course on bimodal provability logic. *Annals of Pure and Applied Logic*, pages 109–142, 1995.

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

[33] A. Visser. Faith & Falsity: a study of faithful interpretations and false Σ_1^0-sentences. Logic Group Preprint Series 216, Department of Philosophy, Utrecht University, Heidelberglaan 8, 3584 CS Utrecht, October 2002.

[34] M Vuković. Some correspondence of principles in interpretability logic. *Glasink Matematički*, 31(51):193–200, 1996.

Received 23-1-2006