Supremum in the Lattice of Interpretability

MSc Thesis *(Afstudeerscriptie)*

written by

Paula Henk
(born June 15th, 1986 in Tallinn, Estonia)

under the supervision of **Prof Dr Dick de Jongh** and **Prof Dr Albert Visser**, and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam.*

Date of the public defense:	**Members of the Thesis Committee:**
September 6, 2012	Prof Dr Dick de Jongh
	Prof Dr Benedikt Löwe
	Prof Dr Frank Veltman
	Prof Dr Rineke Verbrugge
	Prof Dr Albert Visser

Institute for Logic, Language and Computation

Abstract

This thesis is located in the field of provability and interpretability logic, where modal logic is used in the study of formal systems of arithmetic. We are interested in the notion of interpretability, which can be seen as a tool for comparing axiomatic theories. Intuitively, if a theory T interprets a theory S , T is at least as strong as S. The modal logic ILM captures exactly what Peano Arithmetic (PA) can prove about interpretability between finite extensions of itself. As it turns out, finite extensions of PA form a lattice under the relation of interpretability, i.e. any two theories have an infimum and a supremum in the interpretability ordering. The supremum in this lattice is the main subject of study in this thesis.

We will extend the logic ILM with a binary operator for the supremum, and explore the possibilities of having a modal semantics for the resulting system ILMS. For that purpose, the supremum will be studied from the arithmetical as well as from the modal perspective. We see that the exact content of the logic ILMS depends on the formula that is chosen as the arithmetical representative of the supremum. This is different from ILM, where the meaning of the modal symbols is fixed from the outset. Proceeding to the modal side, we establish an important negative result: there can be no structural characterization of ILM–models that validate the defining axiom for the supremum. This precludes the possibility of having a relational semantics for the system ILMS — at least one that would extend the usual semantics for ILM. Finally, we examine an elegant but unfortunately failed attempt to find a relational semantics for a particular representative of the supremum.

CHAPTER 1

Introduction

Wenn man aber einmal mit Logik beginnt, wo ein Gedanke von selbst aus dem vorhergehenden folgt, weiß man zum Schluß nie, wie das endet.

(Robert Musil, "Der Mann ohne Eigenschaften")

The central notion of this thesis is that of interpretability. This notion was first introduced and carefully studied by Tarski, Mostowski, and Robinson in [**TMR53**], who used it to prove the undecidability of certain theories. Interpretations can also be used to yield relative consistency proofs, and to provide languages with conceptual resources that are not *prima facie* present.

The notion of interpretability can also be used to compare formal theories. Intuitively, a theory T is at least as strong as a theory S if T interprets S; we write $T \rhd S$. Clearly, T is at least as strong as S if T proves all theorems of S. However we would also want to be able to compare theories whose languages are different. For example, we might want to compare a system in the language of arithmetic with a system in the language of set theory. In this situation, the idea of a translation arises naturally. Roughly, an interpretation of S in T is a structure-preserving[1] translation from the language of S into the language of T. The translation should have the property that if a sentence A is a theorem of S, then the translation of A is a theorem of T.

The relation of interpretability is a partial preorder on theories. The equivalence classes of the induced equivalence relation — the relation of mutual interpretability — are called degrees (of interpretability). Equipped with the notion of interpretability, we can investigate the structure of the degrees by asking questions like:

i. Is the preorder of interpretability dense? I.e. given theories T and S with $T \rhd S$, is there a theory U with $T \rhd U \rhd S$, $U \not\rhd T$, and $S \not\rhd U$?

ii. Are there incomparable theories? I.e. are there theories T and S with $T \not\rhd S$ and $S \not\rhd T$?

iii. Do any two theories have an infimum? I.e. given theories T and S, is there a theory U with $T \rhd U$, $S \rhd U$, and s.t. for any U' with $T \rhd U'$, $S \rhd U'$, we have that $U \rhd U'$?

[1] "Structure preserving" means that the translation should commute with the propositional connectives. For example, we want $\bot$ to be translated as $\bot$.

iv. Do any two theories have a supremum? I.e. given theories T and S, is there a theory U with $U \rhd T$, $U \rhd S$, and s.t. for any U' with $U' \rhd T$, $U' \rhd S$, we have that $U' \rhd U$?

Assuming that the theories we consider have a certain strength, all the above questions have an affirmative answer[2]. According to iii and iv, the degrees form a lattice. It is the supremum in this lattice that this thesis is mainly about.

1. Background

Our main object of study is the supremum in the lattice $(\mathrm{V}_{\mathsf{PA}}, \rhd)$, where V_{PA} is the set of degrees of finite extensions of Peano Arithmetic (PA). This lattice was first studied by Švejdar in [**Šve78**]. However, instead of being interested in the supremum in this lattice *per se*, we are interested in what is provable about it in PA. This thesis contributes to the research programme of provability and interpretability logic, where modal logic is used to study what formal systems of arithmetic can prove about their own metamathematical properties.

This research programme has been rather successful. Löb's Logic GL (named after Gödel and Löb) captures exactly what a sufficiently strong formal system can prove about provability in itself. Extending GL with a binary modality $\rhd$ for interpretability yields an elegant modal system IL, all of whose theorems are provable interpretability principles for any reasonable arithmetical theory. The modal system ILM is a result of adding to IL the axiom M (Montagna's principle) $(A \rhd B) \to (A \wedge \Box C) \rhd (B \wedge \Box C)$. The logic ILM captures exactly what an essentially reflexive theory — such as PA — can prove about interpretability between its finite extensions. Adding to IL the axiom P (the persistence principle) $A \rhd B \to \Box(A \rhd B)$, we get the logic ILP which captures exactly what a finitely axiomatizable theory[3] can prove about interpretability between its finite extensions.

The goal of this thesis is to study the system ILMS, which is the result of adding to ILM a binary operator $\oslash$ for the supremum, plus axiom S — the defining axiom for $\oslash$:

$$(C \rhd A) \wedge (C \rhd B) \leftrightarrow C \rhd (A \oslash B).$$

As in the case of GL and ILM, the ultimate reason for having a system like ILMS is to use it for finding out what is and what is not provable in PA about the supremum in the lattice $(\mathrm{V}_{\mathsf{PA}}, \rhd)$. For this, arithmetical soundness and completeness is needed. The question what an arithmetical interpretation of the extended modal language is, turns out to be rather delicate. We will provide a careful discussion of the various strategic choices that are possible.

The proofs of the arithmetical completeness for GL and ILM depend on these systems being modally complete w.r.t. a certain class of frames. A natural first step to make is thus to find some nicely describable class of frames, and prove modal completeness of ILMS w.r.t. this class of frames. For this, we would want to extend the modal

[2]On the other hand, it is an interesting fact that all *naturally* occurring theories in the literature have turned out to be linearly ordered under the relation of interpretability. Harvey Friedman calls this "perhaps the most intriguing, thought provoking, fundamental, and deep phenomenon in the whole of the foundations of mathematics" [**Fri07**, p 37].

[3]To be more precise, such a finitely axiomatizable theory is also required to extend $\mathrm{I}\Delta_0$+SUPEXP.

semantics for ILM to a modal semantics for ILMS. The main result of this thesis is that even this first step is not possible — at least not in the way we would expect it to be possible in the best of all worlds.

2. Overview

The reader of this thesis is assumed to be familiar with basic facts about lattices, and the fundamentals of modal logic. Also some knowledge of Gödel's Incompleteness Theorems is useful, though not necessary. Given this assumption, Chapter 2 contains the preliminaries for the rest of this thesis. It gives an introduction to the field of provability and interpretability logic, where modal logic is used in the study of formal systems of arithmetic. We describe the methodology used in this discipline, explain the main results, and provide necessary background concerning arithmetical as well as modal matters. We introduce the systems GL, IL, ILM, and ILP. The final part of Chapter 2 introduces the lattice of degrees of interpretability. We focus on the lattice $(\mathrm{V}_{\mathsf{PA}}, \rhd)$, and discuss some issues concerning the supremum in this lattice.

The next three chapters present original research on the supremum in the lattice $(\mathrm{V}_{\mathsf{PA}}, \rhd)$. The goal of this research is to lay the ground for the logic ILMS, i.e. ILM plus axiom S for the supremum operator $\owedge$.

Chapter 3 deals with the arithmetical side of the supremum. We will present some methodological considerations concerning the intended arithmetical meaning of the modal symbol $\owedge$. The intended meaning of $\owedge$ is of course an arithmetical supremum; however there are different ways for the supremum to be represented in PA. We call a representative of the supremum in PA an *implementation*. As it turns out, the choice of an implementation does matter from the perspective of the logic ILMS. This is different from GL and IL, where the arithmetical meanings of the modal symbols $\Box$ and $\rhd$ are fixed from the outset. Two possible arithmetical meanings of the modal symbol $\owedge$, i.e. two implementations, will be studied in more detail. We call them Švejdar's implementation and Visser's implementation respectively. We also establish a principle about the supremum which is true but nevertheless unprovable in PA — regardless of which implementation we choose to represent the supremum.

Chapters 4 and 5 deal with the modal side of the supremum. Chapter 4 examines the possibility of having a modal semantics for the minimal logic ILMS, extending ILM only by axiom S for $\owedge$. For this, we have to specify what we mean by having a modal semantics in the first place. In the context of modal logic, the ideal would be to find a structural condition determining the truth value of formulas of the form $A \owedge B$ in an ILM–model. Also having a structural characterization of frames satisfying axiom S would be desirable. For this, we determine certain minimal structural conditions that an ILM–model should satisfy if it is to validate axiom S. The main result of Chapter 4 is that the ILM–frames are too "narrow" in order to satisfy these conditions. Hence it is impossible to give a structural characterization of ILM–frames satisfying axiom S. As a consequence, we cannot have a nice modal semantics for ILMS. We then turn to a very weak notion of modal semantics, where the only thing we require from an ILMS–model is that it validates axiom S. We

prove completeness w.r.t. this notion of semantics for a very simple case, explaining also why the strategy used in the proof cannot be extended to the general case.

Chapter 5 contains an attempt to find a relational semantics for the system ILMS, where the intended meaning of the symbol $\owedge$ is Visser's implementation of the supremum. For this, we have to try to "forget" about the negative result of the previous chapter, according to which such a semantics is impossible. Although wrong at the end, this relational semantics is an elegant system on its own. Furthermore, the match between what is valid in the semantics and what is provable in PA about Visser's implementation goes a surprisingly long way.

3. Genesis

We end this chapter by saying a few words about the origin of this thesis. The supremum in the lattice $(\mathrm{V}_{\mathsf{PA}}, \rhd)$ was not our original object of interest. Instead, we wanted to study the supremum in the lattice $(\mathrm{F}, \rhd)$, where F is the set of degrees of finite theories. The structure $(\mathrm{F}, \rhd)$ was studied by Harvey Friedman in [**Fri07**]. He shows that $(\mathrm{F}, \rhd)$ is a distributive lattice, that it contains incomparable elements, and that it is dense.

Our initial goal was to devise a modal logic where Friedman's density argument for the lattice $(\mathrm{F}, \rhd)$ could be formalized. Since the argument is about finite theories, our desired logic would have been an extension of ILP. Examining Friedman's argument, it is clear that the language of interpretability logic alone is not sufficient for expressing it. Apart from the concept of a supremum, the argument uses witness comparisons, and the theory of a natural number. As a first step, we wanted to extend the modal system ILP with a supremum operator. However already this first step turned out to be more difficult than expected. The way Friedman constructs the supremum of given sentences A and B is not obviously suitable to be subject to a modal treatment. It is an open question whether a reasonable arithmetical counterpart for the modal symbol $\owedge$ exists for finitely axiomatizable theories. As a result of Švejdar's investigations ([**Šve78**]), such a counterpart is available for essentially reflexive theories such as PA, and hence we turned our attention to the latter instead.

CHAPTER 2

Preliminaries

In this thesis, modal logic is used to study formal systems of arithmetic. We want to extend a system of interpretability logic with a binary connective whose intended meaning is an "arithmetical supremum". In order to do that, a solid knowledge of both arithmetical and modal matters is needed. This chapter introduces both the arithmetical and the modal preliminaries, as well as the connection between the two, thus laying the ground for the rest of this thesis. Section 1 deals with provability logic, and Section 2 with its extension interpretability logic, extending which in turn is the main goal of this thesis. Section 3 introduces the lattice of interpretability. The supremum in this lattice is the main object of study in this thesis — it is the arithmetical object that we would want to capture with the tools of modal logic.

1. Provability Logic

This section introduces the modal system GL, also known as Löb's Logic or Provability Logic, and its relation to sufficiently strong theories of arithmetic.

1.1. Formal systems of arithmetic. The language of arithmetic is the language of 0, S (successor), $+$, and $\times$. We write $x < y$ as an abbreviation for $\exists z\,(x+\mathsf{S}z = y)$. We define for each $n \in \mathbb{N}$ a term $\overline{n}$ of the language of arithmetic by letting $\overline{0} = 0$, and $\overline{n+1} = \mathsf{S}\overline{n}$. Terms of the form $\overline{n}$ are called *numerals*. A formula A is said to be *bounded* if all quantifiers occurring in A are of the form $\forall x < t$ or $\exists x < t$, where t is a term not containing x.

If Γ is a set of arithmetical formulas closed under subformulas and substitution of terms, then $\mathrm{I}\Gamma$ is the theory containing basic facts about 0, S, $+$, and $\times$, plus induction restricted to Γ. The basic axioms are:

i. 0 is not in the range of the S function

ii. S is injective

iii. recursive definitions of the operations $+$ and $\times$

The weakest arithmetical theory considered in this thesis is the theory EA (elementary arithmetic) aka $\mathrm{I}\Delta_0{+}\mathrm{EXP}$, where EXP states that the exponentiation function is total, and Δ_0 is the set of bounded formulas. EA is known to be finitely axiomatizable (Theorem V.5.6 in [**HP91**]). Although Gödel's Incompleteness Theorems apply also to weaker theories than EA (e.g. Robinson arithmetic Q, or Buss' S^1_2), EA is the weakest theory that naturally occurs in the literature whose provability logic is known to be GL. Our choice is thus motivated by the modal logical perspective.

Another advantage of EA is that it turns out to be much stronger than its superficial weakness (compared to e.g. the full theory of Peano Arithmetic) might lead one to expect. In fact, Harvey Friedman has made the conjecture (in [**FoM**]) that every theorem published in the Annals of Mathematics whose statement involves only finitary mathematical objects can be proved in EA.

The set Δ_0 of bounded formulas is also denoted by Σ_0 and Π_0. We define the sets Σ_n and Π_n for all $n \in \mathbb{N}$.

DEFINITION 1. Let $m \geq 0$. A formula is Σ_{n+1} if it is of the form $\exists z_0 \ldots, \exists z_m\, A$, with A a Π_n-formula. A formula is Π_{n+1} if it is of the form $\forall z_0 \ldots, \forall\, z_m A$, with A a Σ_n-formula. A formula is Δ_n if it is both Σ_n and Π_n. ❑

We shall also say that a formula is Σ_n (Π_n) if it is equivalent to a Σ_n- (Π_n-) formula in the theory T we are considering.

DEFINITION 2. A theory T is *reflexive* if it proves the consistency of each of its finite subtheories. A theory T is *essentially reflexive* if all its finite extensions in the same language are reflexive. ❑

The theory PA (Peano arithmetic) contains the basic facts about 0, S, $+$, and $\times$, plus induction for all formulas. PA can be seen[1] as the theory $\mathsf{EA} + \bigcup_{n\in\omega} \mathrm{I}\Sigma_n$. It is known that for each $n > 0$, $\mathrm{I}\Sigma_n$ is finitely axiomatizable, and that PA is essentially reflexive. For proofs of these facts, see Theorems I.2.52 and III.2.35 in [**HP91**]. Note that as a consequence of the Second Incompleteness Theorem, no reflexive theory can be finitely axiomatized.

1.2. Incompleteness. Until the end of this subsection, let $T \supseteq \mathsf{EA}$ be a Σ_1-sound (hence also consistent) recursively axiomatizable theory in the language of arithmetic. Having the above properties, T is strong enough to be subject to Gödel's Incompleteness Theorems. By the First Incompleteness Theorem, there is a sentence G s.t. $\nvdash_T G$ and $\nvdash_T \neg G$. We say that G is *independent* from T. By the Second Incompleteness Theorem, $\nvdash_T \mathrm{Con}_T$, where Con_T is a formalized statement of T's consistency.

An essential ingredient of Gödel's proof is the idea of arithmetization, i.e. the coding of syntactical objects as natural numbers. This is done in such a way that the basic properties of the syntactical objects (e.g. being a formula, an axiom, or a proof) become computationally simple predicates of the codes. Similarly, basic functions on the syntactical objects (e.g. logical connectives, or substituting a term into a formula) become computationally simple functions on the codes. When A is a formula, we write $\ulcorner A \urcorner$ for the code (also: *gödelnumber*) of A. Thus, whereas by default T is seen as proving statements about natural numbers, it can now be interpreted as proving statements about its own syntax. The coding should be computationally simple so as to ensure that the basic facts concerning T's syntax are provable in T. For our purpose, computationally simple functions will be the Kalmar elementary functions. The Kalmar elementary functions are primitive recursive functions whose growth is bounded by iterations of the exponentiation function. See [**Ros84**] for a number of equivalent characterizations of Kalmar

[1] We will adopt this perspective in Chapter 3.

elementary functions. As usual, a relation R is said to be Kalmar elementary if its characteristic function is Kalmar elementary.

DEFINITION 3. We say that a formula A_R *numerates* a relation R in T if for all $n_0, \ldots, n_k$,

$$\mathbb{N} \vDash R(n_0, \ldots, n_k) \;\Leftrightarrow\; \vdash_T A_R(\overline{n_0}, \ldots \overline{n}_k).$$

A formula A_R is said to *binumerate* R in T if in addition,

$$\mathbb{N} \nvDash R(n_0, \ldots, n_k) \;\Rightarrow\; \vdash_T \neg A_R(\overline{n_0}, \ldots \overline{n}_k).$$

A function f is *provably recursive* if there is a formula A_f binumerating the graph of f, and furthermore $\vdash_T \forall x_0 \ldots, x_m \exists! y\, A_f(x_0, \ldots, x_m, y)$. ❑

THEOREM 4. *For every Kalmar elementary relation R, there is a Δ_0-formula that binumerates R in* EA. *The provably recursive functions of* EA *are exactly the Kalmar elementary functions.*

For a proof that Kalmar elementary functions are provably recursive in EA, see [**Avi03**]. Using Theorem 4, we can assume that T is equipped with function and relation symbols for all Kalmar elementary functions and relations. The computational simplicity of the coding ensures that the set of codes of theorems is Kalmar elementary. Hence by Theorem 4 there is a Δ_0-formula $\mathrm{Prf}_T(x, y)$ that binumerates the axioms of T in T, i.e. for all m and A:

$$m \text{ codes a } T\text{-proof of } A \;\Leftrightarrow \vdash_T \mathrm{Prf}_T(\overline{m}, \overline{\ulcorner A \urcorner}) \tag{1}$$

$$m \text{ does not code a } T\text{-proof of } A \;\Leftrightarrow \vdash_T \neg\mathrm{Prf}_T(\overline{m}, \overline{\ulcorner A \urcorner}) \tag{2}$$

A formula $\mathrm{Prf}_T(x, y)$ as above is said to be a *proof predicate* of T. The formula $\mathrm{Pr}_T(y) := \exists x\, \mathrm{Prf}_T(x, y)$ then numerates the theorems of T, i.e.

$$\varphi \text{ is provable in } T \;\Leftrightarrow \vdash_T \mathrm{Pr}_T(\overline{\ulcorner A \urcorner}). \tag{3}$$

For the direction from right to left, the assumption of Σ_1-soundness of T is needed. A formula $\mathrm{Pr}_T(y)$ satisfying (3) is said to be a *provability predicate* of T. The condition in (3) states that Pr_T is *extensionally correct*[2] w.r.t. theoremhood in T. Note that since $\mathrm{Prf}_T(x, y)$ is Δ_0, the formula $\mathrm{Pr}_T(y)$ is Σ_1.

Another key idea of Gödel's proof is that of diagonalization. Gödel constructed a formula G with $\vdash_T G \leftrightarrow \neg\mathrm{Prf}_T(\overline{\ulcorner G \urcorner})$. We say that G is a fixed point of the formula $\neg\mathrm{Prf}_T(x)$. Carnap ([**Car34**]) was the first to isolate the general Fixed Point Theorem as an independent statement from Gödel's proof. According to the general version of the Fixed Point Theorem, any formula of arithmetic has a fixed point.

LEMMA 1.1 (Gödel-Carnap Fixed Point Lemma). *Let $A(x_0 \ldots, x_n)$ be a formula whose free variables are exactly $x_0, \ldots, x_n$. Then there is a formula $B(x_1, \ldots, x_n)$ whose free variables are exactly $x_1, \ldots, x_n$, and s.t*

$$\vdash_T B(x_1, \ldots, x_n) \leftrightarrow A(\overline{\ulcorner B(v_1, \ldots, v_n) \urcorner}, x_1, \ldots, x_n). \tag{4}$$

We say that $\vdash B(x_1, \ldots, x_n)$ is a fixed point of $A(x_0 \ldots, x_n)$.

[2]Apart from the "natural" provability predicate, there are many strange predicates that satisfy condition (3). An example is the Rosser provability predicate. Informally, a formula is Rosser provable if it has a proof (in the usual sense), and there is no smaller (in terms of gödelnumbers) proof of its negation. See also footnote 3.

For a proof, see [**Smo85**], or [**Boo93**].

DEFINITION 5. Let T be a theory and $\mathrm{Pr}_T(y)$ a provability predicate of T. A *Gödel-sentence of* T *(w.r.t.* Pr_T*)* is a sentence G which is a fixed point of $\neg\mathrm{Pr}_T(y)$ in the sense of Lemma 1.1, i.e. $\vdash_T G \leftrightarrow \neg\mathrm{Pr}_T(\overline{\ulcorner G \urcorner})$. ❑

Using the fact that the formula $\mathrm{Pr}_T(y)$ is extensionally correct, i.e. that it satisfies (3) above, it is easy to see that G must be independent of T.

THEOREM 6 (The First Incompleteness Theorem). *Let G be a Gödel-sentence of T. Then $\nvdash_T G$ and $\nvdash_T \neg G$.*

For the natural way of defining the formula[3] $\mathrm{Pr}_T(y)$, the following holds:

1. $\vdash_T A \;\Rightarrow\; \vdash_T \mathrm{Pr}_T(\overline{\ulcorner A \urcorner})$

2. $\vdash_T (\mathrm{Pr}_T(\overline{\ulcorner A \to B \urcorner}) \wedge \mathrm{Pr}_T(\overline{\ulcorner A \urcorner})) \to \mathrm{Pr}_T(\overline{\ulcorner B \urcorner})$

3. $\vdash_T \mathrm{Pr}_T(\overline{\ulcorner A \urcorner}) \to \mathrm{Pr}_T(\overline{\ulcorner \mathrm{Pr}_T(A) \urcorner})$

The formulas A and B above are also allowed to contain free variables. Properties 1 - 3 are referred to as the *Hilbert-Bernays-Löb derivability conditions*. Note that 1 is just one direction of the property in (3), and 2 the formalized version of modus ponens. Property 3 follows from the fact that for any Σ_1-formula S (possibly containing free variables), $\vdash_T S \to \mathrm{Pr}_T(\overline{\ulcorner S \urcorner})$. This property is often referred to as *provable Σ_1-completeness*.

THEOREM 7 (The Second Incompleteness Theorem). *Let $\mathrm{Pr}_T(y)$ be a provability predicate of T satisfying the Hilbert-Bernays-Löb derivability conditions. Then*

$$\nvdash \neg\mathrm{Pr}_T(\overline{\ulcorner \bot \urcorner}),$$

i.e. T does not prove (the formalized statement of) its own consistency.

PROOF. Using the fact that $\mathrm{Pr}_T(y)$ satisfies the Hilbert-Bernays-Löb derivability conditions, it is not hard to show that $\vdash_T G \leftrightarrow \neg\mathrm{Pr}_T(\overline{\ulcorner \bot \urcorner})$. Hence using the First Incompleteness Theorem, $\neg\mathrm{Pr}_T(\overline{\ulcorner \bot \urcorner})$ must be independent from T. ❑

As seen above, a sentence asserting its own unprovability in T is independent of T. In 1952, Leon Henkin asked what can be said about sentences asserting their own provability. Henkin's question was answered by Martin Hugo Löb in [**Löb55**]. Löb showed that for any formula A, if $\vdash_T \mathrm{Pr}_T(\overline{\ulcorner A \urcorner}) \to A$, then already $\vdash_T A$. This result is referred to as *Löb's Theorem*. Also the formalized version of Löb's Theorem holds.

THEOREM 8 (Formalized Löb's Theorem).

$$\vdash_T \mathrm{Pr}_T(\overline{\ulcorner \mathrm{Pr}_T(\overline{\ulcorner A \urcorner}) \to A \urcorner}) \to \mathrm{Pr}_T(\overline{\ulcorner A \urcorner}).$$

[3]There are extensionally correct provability predicates which do not satisfy properties 2 and 3. See [**Fef60**] for a discussion of how different (extensionally correct) choices of the predicate $\mathrm{Pr}_T(y)$ affect which properties of $\mathrm{Pr}_T(y)$ are provable in T.

For a proof, see Chapter 3 of [**Boo93**]. According to Löb's Theorem, a sentence asserting its own provability is provable. In fact, T proves $\mathrm{Pr}_T(\overline{\ulcorner A \urcorner}) \to A$ only in the trivial case when it already proves A. Note that in the case the axioms of T are true in $\mathbb{N}$, we do believe $\mathrm{Pr}_T(\overline{\ulcorner A \urcorner}) \to A$ to be true (in $\mathbb{N}$) for every A.

Löb's Theorem tells us something about the (un)provability in T of simple sentences containing the provability predicate. It is natural to ask which other such sentences are provable in T — is there a way to characterize *all* such sentences? Looking at the form of the Hilbert-Bernays-Löb derivability conditions, one recognizes the rule of necessitation, the K-axiom, and the transitivity axiom familiar from modal logic. Also the formalized Löb's Theorem seems apt for being viewed as a modal principle. This motivates the idea of using propositional modal logic to study what T can prove about $\mathrm{Pr}_T(y)$. We will give an overview of the successes of this programme below.

1.3. The modal system GL. We assume basic knowledge of modal logic, in particular the system K and its modal semantics. An introduction to modal logic, as well as an accessible treatment of the material presented in this section can be found in [**Boo93**].

Definition 9. The modal logic GL (named after Gödel and Löb) is K plus the principle[4] $\Box(\Box A \to A) \to \Box A$. ❑

The principle $\Box(\Box A \to A) \to \Box A$ is also known as the *Löb axiom*. It is the modal counterpart of the formalized version of Löb's Theorem. The following lemma was first proven by Dick de Jongh. It shows that the (modal counterparts of the) Hilbert-Bernays-Löb derivability conditions are provable in GL.

Lemma 1.2. $\vdash_{\mathsf{GL}} \Box A \to \Box\Box A$

The system GL has a Kripke semantics. A relation R is said to be converse well-founded if for every set $X \neq \emptyset$ there is an element $x \in X$ s.t. there is no $y \in X$ with xRy; i.o.w. if there are no infinite ascending sequences $x_0 R x_1 R x_2 \ldots$. It turns out that GL characterizes Kripke frames whose accessibility relation R is transitive and converse well-founded. The modal completeness of GL w.r.t. transitive converse well-founded frames was proved by Krister Segerberg in [**Seg71**]. In fact, Segerberg proved that GL is even modally complete w.r.t the more restricted class of finite transitive irreflexive trees.

Theorem 10 (Modal Completeness of GL). *Let $\mathcal{K}$ be the class of frames that are transitive irreflexive finite trees. Then $\mathsf{GL} \vdash A \Leftrightarrow \forall \mathcal{F}\, [\mathcal{F} \in \mathcal{K} \Rightarrow \mathcal{F} \Vdash A]$.*

An important result concerning GL is the Fixed Point Theorem, proved independently by Dick de Jongh and Giovanni Sambin. We say that a propositional letter p is *modalized* in A if all its occurrences in A are under the scope of a $\Box$.

Theorem 11 (Fixed Point Theorem for GL). *Let p be modalized in $A(p)$. There is a formula D containing the same propositional letters as $A(p)$ but not containing p s.t.*

[4] A principle is just a rule with an empty antecedent.

i. $\vdash_{\mathsf{GL}} D \leftrightarrow A(D)$

ii. $\vdash_{\mathsf{GL}} [(p \leftrightarrow A(p)) \wedge \Box(p \leftrightarrow A(p))] \rightarrow [(p \leftrightarrow D) \wedge \Box(p \leftrightarrow D)]$

A sentence D with the above properties is called the fixed point of $A(p)$. An algorithm for calculating fixed points in GL is presented e.g. in [**Smo77**]. Applying this algorithm to the sentence $\neg\Box p$ (where p is modalized), we get as a fixed point $\neg\Box\bot$. Thus $\vdash_{\mathsf{GL}} \neg\Box\bot \leftrightarrow \neg\Box\neg\Box\bot$, and

$$\vdash_{\mathsf{GL}} [(p \leftrightarrow \neg\Box p) \wedge \Box(p \leftrightarrow \neg\Box p)] \rightarrow (p \leftrightarrow \neg\Box\bot).$$

This should remind the reader of the situation in arithmetic, where the Gödel-sentence G (i.e. the fixed point of $\neg\mathrm{Pr}_T(y)$) turned out to be equivalent to the consistency statement of T, whose modal counterpart is the sentence $\neg\Box\bot$.

1.4. GL and arithmetic. In order to make precise the idea of GL describing what an arithmetical theory T can prove about its own provability predicate, we need to translate sentences of the modal language to sentences in the language of arithmetic.

Definition 12. An *arithmetical realization* $*$ is a function from the propositional letters of the modal language to sentences in the language of arithmetic. The domain of a realization is extended to all formulas of the modal language by requiring:

i. $(\bot)^* = \bot$

ii. $(A \rightarrow B)^* = A^\star \rightarrow B^*$

iii. $(\Box A)^* := \mathrm{Pr}_T(\overline{\ulcorner A^* \urcorner})$. ❑

Using the notion of a realization, we can make precise the idea of GL being the logic of the provability predicate of a sufficiently strong arithmetical theory.

Definition 13. A modal formula A is a *provability principle* of a theory T if for all realizations $*$, $\vdash_T A^*$. The *provability logic* of a theory T, we write $\mathrm{PrL}(T)$, is a logic that generates exactly the provability principles of T. ❑

Theorem 14 (Arithmetical Soundness and Completeness of GL). *Let $T \supseteq \mathsf{EA}$ be recursively axiomatizable and Σ_1-sound. Then* $\mathrm{PrL}(T) = \mathsf{GL}$.

For the direction $\mathsf{GL} \subseteq \mathrm{PrL}(T)$ (arithmetical soundness), one has to check that the axioms and rules of GL are provable in T under all realizations. But these are just the Hilbert-Bernays-Löb derivability conditions plus the formalized version of Löb's rule introduced in Section 1.2. The proof of the other direction $\mathrm{PrL}(T) \subseteq \mathsf{GL}$ (arithmetical completeness), for $T = \mathsf{PA}$, is due to Robert Solovay ([**Sol76**]). Dick de Jongh, Marc Jumelet and Franco Montagna ([**dJJM91**]) extended the result to weaker theories, in particular to EA.

We will from now on write $\Box_T A$ or even just $\Box A$ (if T is fixed or clear from the context) instead $\mathrm{Pr}_T(\overline{\ulcorner A \urcorner})$. Similarly, we will write $\Diamond_T A$ instead of $\neg\mathrm{Pr}_T((\overline{\ulcorner \neg A \urcorner}))$. The sentence $\Diamond_T A$ is the arithmetization of the assertion that A is consistent with T. If T is sufficiently strong in the sense of Theorem 14, we will refer to the principles of GL when reasoning about what is provable in T about $\mathrm{Pr}_T(\overline{\ulcorner A \urcorner})$. Depending on the context, $\Box A$ can thus denote either an arithmetical or a modal formula. We are sure that no confusion will arise from this.

2. Interpretability Logic

2.1. The notion of interpretability. The notion of interpretability we are interested in was first introduced and carefully studied by Tarski, Mostowski, and Robinson in [**TMR53**]. It can be seen as a rigorous definition of what it means for a theory T to be at least as strong as a theory S, i.e. as a tool for comparing theories.

As an example, consider PA and Zermelo-Fraenkel set theory ZF. It is common knowledge among logicians that ZF is stronger than PA. But what does this mean exactly? How can we compare a theory about numbers to a theory about sets? In general, how can we compare theories if their languages are different? A natural solution is to translate from one language into the other. Although the language of ZF is not the language of arithmetic, ZF can still talk about the natural numbers, since it can talk about finite ordinals. We want to have a translation from formulas in the language of arithmetic to formulas in the language of set theory, such that for every formula provable in PA, its translation turns out to be provable in ZF. Of course, there have to be some restrictions. For example, we have to exclude the possibility of "cheating" by mapping every formula to a tautology. Furthermore, it is clear that we also need a domain function, as ZF can talk about much more than just its natural numbers (i.e. the finite ordinals). Thus if $\forall x\, \varphi$ is a sentence in the language in arithmetic, then the translation should restrict the universal quantifier to the finite ordinals. Otherwise ZF will not even be able to prove the simplest truths about natural numbers, such as every number different from 0 having a predecessor. For an extensive treatment of interpretability between PA and ZF, see [**KW07**].

We will now give a precise definition of interpretability. An even more precise definition can be found in [**Vis98a**]. For simplicity, we assume that our theories are formulated in a purely relational way. Although this is certainly not the case for theories in the language of arithmetic, the assumption does not restrict us in any essential way — function symbols can be replaced by relation symbols by a well-known algorithm.

Definition 15. Let S and T be first order theories. An *interpretation* j of S in T is a tuple $\langle \delta, \tau \rangle$, where δ is a formula in the language of T with one free variable, and τ is a map from relation symbols R (including identity) in the language of S to formulas R^τ in the language of T. We require the number of free variables in R^τ to be equal to the arity of R. We extend τ to a translation from formulas in the language of S to formulas in the language of T by requiring:

i. $(R(\overline{x}))^\tau = R^\tau(\overline{x})$

ii. $(A \to B)^\tau = A^\tau \to B^\tau$

iii. $\bot^\tau = \bot$

iv. $(\forall x A)^\tau = \forall x\, (\delta(x) \to A^\tau)$

Finally, we require that $\vdash_T \exists x\, \delta(x)$, and $\vdash_T A^\tau$ for all axioms A of S. ∎

We write $j : T \rhd S$ if j is an interpretation of S in T. We write $T \rhd S$ if there is an interpretation of S in T. We often also write τ for the interpretation $\langle \delta, \tau \rangle$. If $T \rhd S$ and $S \rhd T$, we write $T \equiv S$ and say that T and S are mutually interpretable.

From a semantic perspective, an interpretation $j : T \rhd S$ is a way of uniformly defining a model of S inside a given model of T. Using the example from above: given a model of ZF, the set of finite ordinals of this model can be seen as a model of PA. The semantic perspective is made precise by the following definition.

DEFINITION 16. Let τ be an interpretation of S in T. Let $\mathcal{M} \vDash T$. Then $\mathcal{M}^\tau$ is defined as:

i. $\mathrm{dom}(\mathcal{M}^\tau) := \{a \in \mathrm{dom}(\mathcal{M}) \mid \mathcal{M} \vDash \delta(a)\} / \sim$, where $a \sim b :\Leftrightarrow \mathcal{M} \vDash a =^\tau b$.

ii. If $\alpha_0, \ldots, \alpha_n \in \mathrm{dom}(\mathcal{M}^\tau)$, let $\mathcal{M}^\tau \vDash R(\alpha_0 \ldots \alpha_n) :\Leftrightarrow \mathcal{M} \vDash R^\tau(a_0 \ldots a_n)$, for some $a_0 \in \alpha_1, \ldots, a_0 \in \alpha_n$.

If $\mathcal{M}^\tau$ is obtained in this way from $\mathcal{M}$, we call it an *internal model* of $\mathcal{M}$. ❑

In [**TMR53**], interpretations are used to show undecidability of certain theories. We say that S is essentially undecidable if any consistent extension S' of S is undecidable. Suppose that S is essentially undecidable. Then if $T \rhd S$, we can conclude that also T is undecidable. Another application of interpretations are relative consistency results. Suppose that $T \rhd S$ and T is consistent. Let $\mathcal{M} \vDash T$. Using Definition 16, we can construct inside $\mathcal{M}$ a model $\mathcal{M}^\tau$ with $\mathcal{M}^\tau \vDash S$. Hence also S must be consistent.

As is usually the case with modal-logically inspired work on interpretability, we will study interpretability between finite extensions of a given base theory. Apart from technical reasons (which will be explained below), such an approach is quite natural, as one is often interested in how additional axioms influence the strength of some base theory. For example, think about adding the Continuum Hypothesis or its negation to ZFC. If T is our base theory, we shall write $A \rhd_T B$ for $T + A \rhd T + B$. If T is fixed or clear from the context, we will just write $A \rhd B$.

2.2. Formalized interpretability. Just as in GL we want to model what a sufficiently strong and Σ_1-sound theory T can prove about provability in itself, in interpretability logic we are interested in what such a T can prove about interpretability between its finite extensions. For the purposes of interpretability logic, sufficiently strong will mean containing $\mathrm{I}\Sigma_1$. We could also work in a weaker theory, however in that case two different notions of interpretability (see [**Joo04**]) would have to be distinguished. To keep things simple, we will thus require all our theories to contain[5] $\mathrm{I}\Sigma_1$.

We can think of formalized interpretability as the following sentence (see [**Joo98**]):

$$A \rhd_T B :\Leftrightarrow \exists \tau \left[\Box_T(A \to B^\tau) \wedge \forall y \left(\mathrm{Ax}_T(y) \to \Box_T(A \to y^\tau)\right)\right] \tag{5}$$

[5]The two different notions of interpretability are provably equivalent already in the weaker theory $\mathsf{EA}+\mathrm{B}\Sigma_1$, i.e. EA plus Σ_1-collection. See Chapter 7 of [**Kay91**] for information about the collection axioms.

Above, T is a theory containing $\mathrm{I}\Sigma_1$, $\Box_T$ is a provability predicate of T, and $\mathrm{Ax}_T(y)$ is a formula binumerating[6] the axioms of T in T. τ is a (code for) for the interpretation, i.e. it is (the code of) a tuple $\langle \delta, \tau \rangle$ as in Definition 15. Indeed, the above formula follows very closely the informal definition of interpretability. Note that the complexity of this formula is much higher than that of $\Box_T$ — it is Σ_3. Hence we *cannot* expect it to be extensionally correct like the provability predicate, i.e. to have for all A, B

$$A \rhd_T B \Leftrightarrow \vdash_T A \rhd_T B.$$

The sentence on the left hand side is the informal statement that $T + A$ interprets $T + B$, whereas the sentence on the right hand side is the statement formalized in T, i.e. the formula in (5). Thus we use the symbol $\rhd_T$ for both formal and informal interpretability. We are certain that the intended meaning is always clear from the context.

2.3. Arithmetical principles for interpretability. Let $T \supseteq \mathrm{I}\Sigma_1$. We will list some valid principles concerning interpretability over T as base theory. All these principles are verifiable in T (in fact, it is for the verifiability that we require $T \supseteq \mathrm{I}\Sigma_1$ – see also footnote 5 above). For a proof, see [**Vis91**], or [**Joo04**]. For principle 4 below, also [**Fef60**]. In order to get good intuitions about these principles, it is useful to look at them from the semantic perspective of Definition 16.

1. $\vdash_T A \to B \;\Rightarrow\; A \rhd_T B$. In the case that $\vdash_T A \to B$, we can use some tautology (e.g. $x = x$) as the domain formula δ, and the identity function as the translation τ. It is easy to check that this choice satisfies the conditions of Definition 15.

2. $(A \rhd_T B,\ B \rhd_T C) \;\Rightarrow\; A \rhd_T C$. For this, it suffices to check that the composition of interpretations is again an interpretation. This principle expresses that interpretability is transitive.

3. $A \rhd_T C,\ B \rhd_T C \;\Rightarrow\; A \vee B \rhd C$. This principle expresses that we can define interpretations by case distinctions. Suppose that the fact that $A \rhd_T C$ is witnessed by the interpretation $\langle \delta_1, \tau_1 \rangle$, and the fact that $B \rhd_T C$ by the interpretation $\langle \delta_2, \tau_2 \rangle$. We can then define a third interpretation $\langle \delta_3, \tau_3 \rangle$ by letting: $\delta_3(x) := (A \to \delta_1(x)) \wedge (\neg A \to \delta_2(x))$, and similarly for the translation τ_3.

4. $A \rhd_T B \;\Rightarrow$ if $T + A$ is consistent, then $T + B$ is consistent. The formalized version of this is: $A \rhd_T B \to (\Diamond_T A \to \Diamond_T B)$. This principle reflects the fact that interpretability yields relative consistency statements.

5. $\Diamond_T A \rhd_T A$. This principle reflects the fact that Henkin's completeness proof can be formalized in T. The analogy is more apparent from the semantic perspective: if a model $\mathcal{M}$ satisfies the statement of A's consistency, then $\mathcal{M}$ has an internal model of A. Although this result already appears in [**Wan51**], the first fully formalized proof is due to Feferman ([**Fef60**]). See also [**Hen11**] for an accessible exposition.

We also have that if A is consistent, then $A \not\rhd_T \Diamond_T A$. This uses the formalized version of principle 4 above, and the Second Incompleteness Theorem. We refer to this principle as the Interpretability version of the Second Incompleteness Theorem.

[6] As explained in Section 1.2 this means: φ is an axiom of T iff $\vdash_T \mathrm{Ax}_T(\overline{\ulcorner \varphi \urcorner})$.

If our base theory is finitely axiomatizable and Σ_1-sound, another principle holds. Namely, if $A \rhd_T B$, then $\vdash_T A \rhd_T B$. To see this, note that if T is finite, then the sentence

$$\Box_T(A \to B^\tau) \wedge \forall y\,(\mathrm{Ax}_T(y) \to \Box_T(A \to y^\tau))$$

can be replaced by the Σ_1-sentence $\Box_T(A \to C^\tau)$, where C is the conjunction of all axioms of $T+B$. Hence if $A \rhd_T B$, then $\vdash_T A \rhd_T B$, because T is Σ_1-complete (since T is an extension of $\mathrm{I}\Sigma_1$, even provable Σ_1-completeness holds.). The formalized version of this principle, i.e. $A \rhd_T B \to \Box_T(A \rhd_T B)$, is the so-called persistence principle P.

2.4. Interpretability over PA. In this section, we will see what can be said about $A \rhd_T B$ if $T = \mathsf{PA}$ (in fact, the results hold for any essentially reflexive theory). Let $\Box_n$ denote the provability predicate of a finite subtheory T_n of PA with[7] $\mathsf{PA} = \bigcup_{n\in\omega} T_n$. The following characterization of interpretability over PA is referred to as the *Orey-Hàjek characterization* (see [**Ore61**], [**Háj71**],and [**Háj72**]).

THEOREM 17 (The Orey-Hájek characterization). *The following are equivalent:*

1. $A \rhd_{\mathsf{PA}} B$
2. *for all* n, $\vdash_{\mathsf{PA}} A \to \Diamond_n B$
3. *if* C *is* Π_1 *and* $\vdash_{\mathsf{PA}} B \to C$, *then* $\vdash_{\mathsf{PA}} A \to C$ *(*A *is* Π_1*-conservative over* B*)*

In fact, all this can be verified in $\mathrm{I}\Sigma_1$.

For a proof (especially of the verifiability in $\mathrm{I}\Sigma_1$), see also [**Joo04**].

We use Theorem 17 to show that the following principle is valid for interpretability over PA. If S is Σ_1, then

$$A \rhd_{\mathsf{PA}} B \Rightarrow (A \wedge S) \rhd_{\mathsf{PA}} (B \wedge S). \tag{6}$$

PROOF. Assume $A \rhd_{\mathsf{PA}} B$. By the Orey-Hájek characterization, A is Π_1-conservative over B. It suffices to show that if S is Σ_1, then $A \wedge S$ is Π_1-conservative over $B \wedge S$. So let P be Π_1 and suppose that $B \wedge S \to P$. Then $B \to (\neg S \vee P)$. Since $\neg S \vee P$ is Π_1, we have that $A \to (\neg S \vee P)$ (since A is Π_1-conservative over B), i.e. $A \wedge S \to P$, which is what we wanted to show. The proof uses the Orey-Hájek characterization and simple predicate logic, and can thus be verified in $\mathrm{I}\Sigma_1$. ❑

2.5. Interpretability logics. Like provability, interpretability can be studied by means of modal logic. This can be done by extending the basic modal language by a binary modality $\rhd$ for interpretability. Thus, apart from informal and formal interpretability, $\rhd$ will now also denote a modal symbol.

In order to make precise the idea of formal interpretability being the intended meaning of $\rhd$, we have to extend the notion of an arithmetical realization to modal formulas of the form $A \rhd B$. In the light of this, it makes sense that we have insisted on working with interpretability between sentential extensions of a given base theory. It is unclear how one should translate propositional letters to whole theories. For example, if p is mapped to PA, then what should $\neg p$ be mapped to?

[7]For example, we could take $T_n = \mathrm{I}\Sigma_n$.

Hence we will always fix some base theory T. The sentence $p \rhd q$ of the modal language will then be mapped to the arithmetical sentence $T + p^* \rhd T + q^*$, where $*$ is a realization as defined as in Section 1.4.

DEFINITION 18. Let T be a base theory. An arithmetical realization $*$ for an interpretability logic is an arithmetical realization for GL extended by the clause:

$$(A \rhd B)^* = A^* \rhd_T B^*,$$

where $\rhd_T$ on the right hand side is formalized interpretability as in (5). ❑

The formulas of interpretability logic are defined as follows:

$$\mathrm{F}_{\mathsf{IL}} ::= \bot \mid \mathrm{Prop} \mid (\mathrm{F}_{\mathsf{IL}} \to \mathrm{F}_{\mathsf{IL}}) \mid \Box \mathrm{F}_{\mathsf{IL}} \mid (\mathrm{F}_{\mathsf{IL}} \rhd \mathrm{F}_{\mathsf{IL}}) \tag{7}$$

Prop is always either a countable or finite set of proposition letters $p, q, r, p_0, \ldots$. We employ the usual definitions of the logical operators $\neg, \vee, \wedge$, and $\leftrightarrow$, and write $\Diamond A$ for $\neg\Box\neg A$. We shall omit brackets that are superfluous according to the following reading conventions. The operators $\Box$, $\rhd$, and $\neg$ bind equally strong. They bind stronger than the equally strong binding $\wedge$ and $\vee$ which in turn bind stronger than $\rhd$. The weakest (weaker than $\rhd$) binding connectives are $\to$ and $\leftrightarrow$. Hence we shall write $A \rhd B \to A \wedge \Box C \rhd B \wedge \Box C$ instead of $(A \rhd B) \to ((A \wedge \Box C) \rhd (B \wedge \Box C))$. We also write $A \equiv B$ for $A \rhd B \wedge B \rhd A$. We say that a formula is a $\Box$-formula or a $\rhd$-formula if its principal connective is $\Box$ or $\rhd$ respectively.

We will now introduce the basic interpretability logic IL.

DEFINITION 19. The logic IL is the smallest logic containing the tautologies of propositional logic, closed under modus ponens, necessitation, and the following principles:

$\mathsf{L1}$ $\Box(A \to B) \to (\Box A \to \Box B)$

$\mathsf{L2}$ $\Box A \to \Box\Box A$

$\mathsf{L3}$ $\Box(\Box A \to A) \to \Box A$

$\mathsf{J1}$ $\Box(A \to B) \to A \rhd B$

$\mathsf{J2}$ $(A \rhd B) \wedge (B \rhd C) \to (A \rhd C)$

$\mathsf{J3}$ $(A \rhd C) \wedge (B \rhd C) \to (A \vee B) \rhd C$

$\mathsf{J4}$ $A \rhd B \to (\Diamond A \to \Diamond B)$

$\mathsf{J5}$ $\Diamond A \rhd A$

Principles $\mathsf{L1} - \mathsf{L3}$ are just the principles of GL. Principles $\mathsf{J1} - \mathsf{J5}$ correspond to the valid principles 1-5 of interpretability introduced in Section 2.3. ❑

LEMMA 2.1. *The following are provable in* IL.

i. $\Box A \leftrightarrow \neg A \rhd \bot$

ii. $A \rhd (A \wedge \Box\neg A)$

iii. $\bot \rhd A$

iv. $A \vee \Diamond A \rhd A$

PROOF. For i note that $\Box A \to \Box(\neg A \to \bot)$ and use J2. For the other direction use J4. According to i we can view $\Box A$ as an abbreviation for $\neg A \rhd \bot$. For ii note that $A \rhd (A \wedge \Box\neg A) \vee (A \wedge \Diamond A)$, whence it suffices to show that $A \wedge \Diamond A \rhd A \wedge \Box\neg A$. By J5 it is sufficient to show that $A \wedge \Diamond A \rhd \Diamond(A \wedge \Box\neg A)$. But this follows by J4 since by L3 we have $\Diamond A \to \Diamond(A \wedge \Box\neg A)$. For iii use J1, and for iv J5 and J3. ❑

Just as for GL, there is a Fixed Point Theorem for IL. We say that p is *modalized* in $A(p)$ if all occurrences of p occur in the scope of a $\Box$ or a $\rhd$.

THEOREM 20 (Fixed Point Theorem for IL). *Let $A(p)$ be modalized in p. Then there is a unique (modulo provable equivalence) formula B containing the same propositional letters as $A(p)$ but not containing p s.t.*

$$\vdash_{\mathsf{IL}} B \leftrightarrow A(B).$$

For a proof, see [**dJV91**].

2.6. Interpretability logics and arithmetic. As in the case of GL, we will make precise the idea of an interpretability logic characterizing what a sufficiently strong theory can prove about interpretability between its finite extensions.

DEFINITION 21. A modal formula A is an *interpretability principle* of a theory T if for all realizations $*$, $T \vdash A^*$. The interpretability logic of a theory T, we write $\mathsf{IL}(T)$, is a logic that generates all interpretability principles of T. ❑

The logic IL is arithmetically sound w.r.t. a wide range of theories.

THEOREM 22 (Arithmetical Soundness of IL). *If $T \supseteq \mathrm{I}\Sigma_1$, then $\mathsf{IL} \subseteq \mathsf{IL}(T)$*

To prove Theorem 22, one just needs to show that the valid principles of Section 2.4 can be verified in T. As mentioned above, this can be done if $T \supseteq \mathrm{I}\Sigma_1$.

IL is not arithmetically complete w.r.t. any class of theories. In order to get arithmetical completeness, we have to add specialized principles to IL. As it turns out, we have different interpretability logics for different classes of theories.

DEFINITION 23. The logic ILP is IL plus the persistency principle P.

$$\mathsf{P}: \quad A \rhd B \to \Box(A \rhd B)$$ ❑

As we showed at the end of Section 2.3, the persistence principle is valid if our base theory T is finitely axiomatized.

DEFINITION 24. The logic ILM is IL plus Montagna's principle M.

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

Montagna's principle is the modal counterpart of interpretability principle (6) introduced in Section 2.4. We saw that it is valid if our base theory T is essentially reflexive. Since in the language of modal logic we cannot talk about Σ_1-sentences, we take as a representative $\Box$-formulas (remember that the intended interpretation of $\Box$, i.e. the provability predicate, is a Σ_1-formula).

Albert Visser [**Vis**] proved that if $T \supseteq \mathrm{I}\Sigma_1$ is Σ_1-sound and finitely axiomatizable, then $\mathsf{IL}(T) = \mathsf{ILP}$ (we say that ILP is arithmetically complete w.r.t. finitely axiomatizable theories). Alessandro Berarducci ([**Ber90**]) and Volodya Shavrukov ([**Sha88**]) proved independently that if T is essentially reflexive, then $\mathsf{IL}(T) = \mathsf{ILM}$ (we say that ILP is arithmetically complete w.r.t. essentially reflexive theories). Rineke Verbrugge ([**Ver93**]) proved that ILM is also the logic of feasible[8] interpretability over PA.

2.7. Semantics for interpretability logics. As in the case of GL, the arithmetical completeness results for ILP and ILM use that these logics are modally complete w.r.t. to a certain class of frames. A modal semantics for interpretability logics was discovered by Frank Veltman.

DEFINITION 25. An IL*–frame* is a tuple $\langle W, R, S\rangle$, where W is a non-empty countable set of nodes, R is a binary relation on W, and S a set of binary relations on W, indexed by the elements of W. The R and S relations satisfy the following requirements:

1. R is conversely well-founded
2. $xRyRz \Rightarrow xRz$
3. $yS_xz \Rightarrow xRy$ and xRz
4. $xRy \Rightarrow yS_xy$
5. $xRyRz \Rightarrow yS_xz$
6. $xS_wyS_wz \Rightarrow xS_wz$

We write ySz to mean that for some w, yS_wz. We will sometimes represent S as a ternary relation, writing $\langle x, y, z\rangle$ for yS_xz. ❑

DEFINITION 26. An IL*–model* is a quadruple $\langle W, R, S, \Vdash\rangle$, where $\langle W, R, S\rangle$ is an IL–frame, and $\Vdash$ is a forcing relation on $\langle W, R, S\rangle$ satisfying the usual clauses (with R as the accessibility relation for $\Box$), together with

$$x \Vdash A \rhd B \;\Leftrightarrow\; \forall y\,(xRy \Vdash A \Rightarrow \exists z\,(yS_xz \Vdash B))\,. \qquad ❑$$

DEFINITION 27. An IL-frame $\langle W, R, S\rangle$ is an ILP*–frame* if it satisfies the following additional condition:

$$wRuRx \wedge xS_wz \Rightarrow xS_uz. \tag{8}$$

We will refer to (8) as the ILP–frame condition. ❑

DEFINITION 28. An IL–frame $\langle W, R, S\rangle$ is an ILM*–frame* if it satisfies the following additional condition:

$$xS_wyRz \Rightarrow xRz. \tag{9}$$

We will refer to (9) as the ILM–frame condition. ❑

[8]In feasible interpretability, the complexity of the proofs associated to the interpretation is bounded by a $\mathsf{P\text{-}Time}$ computable function.

It is easy to verify that the principles P and M characterize the class of ILP–frames and the class of ILM–frames respectively.

Modal completeness results for ILP and ILM were obtained by Frank Veltman and Dick de Jongh ([**dJV90**]).

THEOREM 29. $\vdash_{\mathsf{ILP}} A \Leftrightarrow$ *for all* ILP*–frames* $\mathcal{F}, \mathcal{F} \vDash A$

THEOREM 30. $\vdash_{\mathsf{ILM}} A \Leftrightarrow$ *for all* ILM*–frames* $\mathcal{F}, \mathcal{F} \vDash A$

Joost Joosten ([**Joo98**] and [**Joo04**]) has proved modal completeness of ILM using the construction method. An overview of this proof is given in Section 3.1 of Chapter 4, and a thorough proof in the appendix.

3. Degrees of Interpretability

The relation $\rhd$ of interpretability is a preorder (i.e. it is transitive and reflexive). Consider the induced equivalence relation $\equiv$ of mutual interpretability. Its equivalence classes are called degrees (of interpretability). If T is a theory, we denote by $[T]$ the degree of T. The relation $\rhd$ induces a partial ordering among degrees: $[T] \rhd [S] \Leftrightarrow T \rhd S$. Let $\mathcal{K}$ be a class of theories, and let $\mathrm{Deg}_{\mathcal{K}}$ be the set of degrees of theories in $\mathcal{K}$. In this section we are concerned with the structure $(\mathrm{Deg}_{\mathcal{K}}, \rhd)$. For example, is it a lattice? Is it dense? Are there incomparable elements?

3.1. Lattices of Interpretability. Questions like the ones above were asked and answered by Švejdar in [**Šve78**]. This was before the field of interpretability logics existed, and thus Švejdar was interested in facts about the structure $(\mathrm{Deg}_{\mathcal{K}}, \rhd)$, not the verifiability of these facts in some formal theory.

The theories considered by Švejdar are of the form $T + A$, where T is a fixed base theory, and A is a sentence in the language of T. If T is our base theory, we write V_T for the set of degrees of finite extensions of T. If $[T + A]$ is an element of V_T, we will often just write $[A]$ instead of $[T + A]$.

It is easy to see that the structure $(\mathrm{V}_T, \rhd)$ is a lower semilattice. The degree of all sentences provable in T (we write $[\top]$) is a minimum element of $(\mathrm{V}_T, \rhd)$, and the degree of all sentences refutable in T (we write $[\bot]$) is a maximum element[9] of V_T. We say that a degree $[A]$ is consistent if $[A] \neq [\bot]$. If T is required to have a certain strength[10], then the ordering $\rhd$ on V_T is dense, and for every degree $[A] \neq [\top], [\bot]$, there are degrees incomparable with $[A]$. Švejdar shows that if T is essentially reflexive, then the structure $(\mathrm{V}_T, \rhd)$ is a distributive lattice, where no element apart from $[\top]$ and $[\bot]$ has a complement.

Lindström ([**Lin79**]) studies the structure $(\mathrm{D}_T, \rhd)$, where D_T is the set of degrees of *all* extensions of an essentially reflexive theory T. He shows that $(\mathrm{V}_T, \rhd)$ and $(\mathrm{D}_T, \rhd)$ are isomorphic. The proof uses that for any extension S of T, there is a sentence A_S s.t. $S \equiv T + A_S$.

[9]Compared to the Lindenbaum-Tarski algebra of the theory T with $[\bot]$ as minimum and $[\top]$ as maximum element, the lattice $(\mathrm{V}_T, \rhd)$ is upside down. We shall try to not let this confuse us.
[10]Švejdar requires T to be a consistent extension of Q in a finite language.

Harvey Friedman ([**Fri07**]) deals with the degrees obtained by restricting $\rhd$ to finite theories (i.o.w. to single sentences). Let F be the set of degrees of all finite theories. Friedman shows that the structure $(\mathrm{F}, \rhd)$ is a distributive lattice with $[\bot]$ (the degree of all refutable sentences) as the maximum element. The minimum element is the degree of all sentences with a finite model. The lattice $(\mathrm{F}, \rhd)$ is dense, and for every degree $[A] \neq [\top], [\bot]$, there are degrees incomparable with $[A]$.

3.2. Infimum and Supremum in the Lattice of Interpretability. Fix a theory T with $\mathsf{IL} \subseteq \mathsf{IL}(T)$. We consider the structure $(\mathrm{V}_T, \rhd_T)$, writing $[A]$ for $[T + A]$, and $\rhd$ for $\rhd_T$. We will first show that $(\mathrm{V}_T, \rhd)$ is a lower semilattice.

Definition 31. The *infimum* of $[A]$ and $[B]$ in $(\mathrm{V}_T, \rhd)$ is the degree of a theory $A \owedge B$ with the following properties:

1. $A \rhd C, B \rhd C \Rightarrow A \owedge B \rhd C$
2. $A \rhd A \owedge B,\ B \rhd A \owedge B$ ∎

It is easy to see that one can always take $A \vee B$ for $A \owedge B$. We just need to be able to define definitions by cases (for 1), and have provability as a special case of interpretability (for 2). As argued in Section 2.3, this is indeed the case for the interpretations we are working with. Since $\mathsf{IL} \subseteq \mathsf{IL}(T)$, and the principles above are principles of IL, the fact that $A \vee B$ is the infimum of A and B is actually verifiable in T, i.e. we have that

$$\vdash_T A \rhd C \wedge B \rhd C \to A \vee B \rhd C,$$

and

$$\vdash_T (A \rhd (A \vee B)) \wedge (B \rhd (A \vee B)).$$

Definition 32. The *supremum* of $[A]$ and $[B]$ in $(\mathrm{V}_T, \rhd)$ is the degree of a theory $A \obar B$ with the following properties:

1. $C \rhd A, C \rhd B \Rightarrow C \rhd A \obar B$
2. $A \obar B \rhd A,\ A \obar B \rhd B$ ∎

Note that the second requirement in the above definition is equivalent to the requirement: $C \rhd A \obar B \Rightarrow C \rhd A, C \rhd B$. Assume $A \obar B \rhd A$ and $A \obar B \rhd B$. If $C \rhd A \obar B$, then by transitivity of $\rhd$, $C \rhd A$ and $C \rhd B$. For the other direction, assume that $C \rhd A \obar B \Rightarrow C \rhd A \wedge C \rhd B$. Since $A \obar B \rhd A \obar B$ by reflexivity, it follows that $A \obar B \rhd A$ and $A \obar B \rhd B$. Hence we can define $A \obar B$ also as a sentence satisfying $C \rhd A, C \rhd B \Leftrightarrow C \rhd A \obar B$. From now on, we will use both of the definitions interchangeably.

The solution for the infimum might inspire one to try taking $A \wedge B$ for $A \owedge B$. It is clear that in this way we get an upper bound for A and B, since obviously $A \wedge B \rhd A$, and $A \wedge B \rhd B$. As we will see below, $A \wedge B$ is not necessarily a least upper bound, i.e. there are theories in lower degrees than $A \wedge B$ which still interpret both A and B. In particular, we usually want the supremum of A and $\neg A$ to be consistent. In order to find the supremum in the lattice $(\mathrm{V}_T, \rhd)$, we need to know more about the theory T. Below, we will show how to find suprema in $(\mathrm{V}_T, \rhd)$ if T is essentially reflexive or finitely axiomatizable.

3.3. Supremum in $(\mathrm{V}_{\mathsf{PA}}, \rhd)$. Instead of focusing on $(\mathrm{V}_T, \rhd)$ for any essentially reflexive T, we will be working with PA. The results also hold for any other Σ_1-sound theory satisfying full induction. We will first prove that $A \wedge B$ is not always a least upper bound of $[A]$ and $[B]$ in $(\mathrm{V}_{\mathsf{PA}}, \rhd)$. This is a consequence of the following Theorem, proven by Švejdar in [**Šve78**]. It states that if $[A]$ and $[B]$ are both consistent, then so is their supremum.

THEOREM 33. *Let* $[A], [B] \neq [\bot]$. *Then* $[\Diamond A \wedge \Diamond B] \rhd [A]$ *and* $[\Diamond A \wedge \Diamond B] \rhd [B]$. *Furthermore* $[\Diamond A \wedge \Diamond B] \neq [\bot]$.

PROOF. We use principles of IL to reason about $(\mathrm{V}_{\mathsf{PA}}, \rhd)$. By J5, $\Diamond A \rhd A$. Using J1 and J2, it is easy to see that $\Diamond A \wedge \Diamond B \rhd A$, and similarly $\Diamond A \wedge \Diamond B \rhd B$. Thus $\Diamond A \wedge \Diamond B$ is an upper bound for A and B. To see that it is consistent, assume for contradiction that $\vdash_{\mathsf{PA}} \Box\neg A \vee \Box\neg B$. By Σ_1-soundness, $\mathbb{N} \vDash \Box\neg A \vee \Box\neg B$. Suppose w.l.o.g. that $\mathbb{N} \vDash \Box\neg A$. By decoding, we get a PA-proof of $\neg A$, contradicting the assumption that $[A] \neq [\bot]$. ❑

Since $\mathsf{IL} \subseteq \mathsf{IL}(\mathsf{PA})$, we can use principles of IL in the proof. However, note that in this case we need more, i.e. Σ_1-soundness, hence it is not obvious that this property of $(\mathrm{V}_{\mathsf{PA}}, \rhd)$ can be verified in PA. As we will see in Chapter 3, it is in fact outside the scope of facts verifiable in PA.

COROLLARY 34. *No degree in* $(\mathrm{V}_{\mathsf{PA}}, \rhd)$ *apart from* $[\top]$ *and* $[\bot]$ *has a complement.*

PROOF. Suppose for contradiction that $[A] \neq [\top], [\bot]$ has a complement $[B]$. This means that $[A \owedge B] = [\bot]$, and $[A \vee B] = [\top]$. We cannot have $[B] = [\bot]$, otherwise $[A \vee B] = [A]$, contradicting that $[A] \neq [\top]$. But if $[B] \neq [\bot]$, then by Theorem 33 $[A \owedge B] \neq [\bot]$, contradiction. ❑

Now let A be some sentence independent of PA, e.g. the Gödel-sentence. Then both A and $\neg A$ are consistent, hence $A \owedge \neg A$ has to be consistent, thus we cannot simply take $A \wedge \neg A$ as $A \owedge \neg A$. However, also the upper bound established in Theorem 33 is too high. We obviously want that $[\top \owedge \top] = [\top]$. But according to the interpretability version of the Second Incompleteness Theorem, $\top \not\rhd \Diamond\top$, hence we cannot take $[\Diamond\top]$ as $[\top \owedge \top]$.

Thus the supremum of A and B has to be weaker than $\Diamond A \wedge \Diamond B$. A useful hint to the right direction comes from the Orey-Hájek characterization of interpretability. We know that $C \rhd A$ and $C \rhd B$ iff $C \vdash_{\mathsf{PA}} (\Diamond_n A \wedge \Diamond_n B)$ for all $n \in \mathbb{N}$. Using the Orey-Hájek characterization and the fact that $C \vdash_{\mathsf{PA}} \Diamond_n C$ for all C and n, due to essential reflexivity, we see that it would be sufficient for the sentence $A \owedge B$ to satisfy for all n,

$$\vdash_{\mathsf{PA}} \Diamond_n A \wedge \Diamond_n B \leftrightarrow \Diamond_n (A \owedge B). \tag{10}$$

With this insight, finding the supremum of given degrees $[A]$ and $[B]$ in $(\mathrm{D}_{\mathsf{PA}}, \rhd)$ becomes easy — i.o.w. it becomes easy finding an *infinite* theory in the degree of $A \owedge B$. Just let $A \owedge B := \{\Diamond_n A \wedge \Diamond_n B \mid n \in \mathbb{N}\}$. This theory is clearly not finite, hence not exactly what we are looking for (since we are interested in the structure $(\mathrm{V}_{\mathsf{PA}}, \rhd)$). Of course, by Lindström's proof that $(\mathrm{V}_{\mathsf{PA}}, \rhd)$ and $(\mathrm{D}_{\mathsf{PA}}, \rhd)$ are isomorphic, there exists some finite theory in the degree of $\{\Diamond_n A \wedge \Diamond_n B \mid n \in \mathbb{N}\}$,

and we could be done here. However Švejdar offers a more direct construction of the supremum, by "compressing" the infinite information contained in the above set into a sentence.

Let $\mathrm{Pr}_{\mathsf{PA}\restriction n}$ be the provability predicate of the theory axiomatized by axioms of PA whose gödelnumber is less than n. It is clear that for each $n \in \mathbb{N}$, $\mathrm{Pr}_{\mathsf{PA}\restriction n}$ is the provability predicate of a finite theory. As usual, we write $\mathrm{Con}_{\mathsf{PA}\restriction n}(\overline{\ulcorner \theta \urcorner})$ as an abbreviation for $\neg \mathrm{Pr}_{\mathsf{PA}\restriction n}(\overline{\ulcorner \neg \theta \urcorner})$. Given sentences A and B, Švejdar takes the formula $\forall x\,(\mathrm{Con}_{\mathsf{PA}\restriction x}(z) \to \mathrm{Con}_{\mathsf{PA}\restriction x}(\overline{\ulcorner A \urcorner}) \wedge \mathrm{Con}_{\mathsf{PA}\restriction x}(\overline{\ulcorner B \urcorner}))$ with one free variable z, and then applies the Gödel Fixed Point Lemma to obtain a sentence θ s.t.

$$\vdash_{\mathsf{PA}} \theta \leftrightarrow \forall x\,(\mathrm{Con}_{\mathsf{PA}\restriction x}(\overline{\ulcorner \theta \urcorner}) \to \mathrm{Con}_{\mathsf{PA}\restriction x}(\overline{\ulcorner A \urcorner}) \wedge \mathrm{Con}_{\mathsf{PA}\restriction x}(\overline{\ulcorner B \urcorner})).$$

By the Orey-Hájek characterization, one can then show that $\theta \rhd A$, $\theta \rhd B$, and if $C \rhd A$ and $C \rhd B$, then $C \rhd \theta$. Thus we can take θ as $A \owedge B$. We will give a proof of this in the next chapter. We will also show that this proof can be formalized in PA, hence PA verifies that V_{PA} is a lattice. We will refer to the supremum constructed by Švejdar as *Švejdar's supremum*. In the next chapter, we will show that there are other ways of uniformly choosing a representative in the degree of $[A \owedge B]$ than the one introduced by Švejdar.

3.4. Supremum in $(\mathrm{F}, \rhd)$. We finish this chapter by saying a few words about the supremum in the lattice of finite theories. We have seen that finding a representative from the degree of the supremum in $(\mathrm{V}_{\mathsf{PA}}, \rhd)$ makes heavy use of the Orey-Hájek characterization peculiar to essentially reflexive theories. Dealing with the suprema in finite theories requires a different kind of toolkit. In [**Fri07**], $A \owedge B$ is taken to be the sentence $\exists xy\,((Rx \wedge \neg Ry) \wedge (A^R \wedge B^{\neg R}))$, where R is some new relation symbol, and A^R is obtained from A by relativizing all predicates of A to R; similarly for $B^{\neg R}$. This guarantees that $A \owedge B$ is consistent whenever A and B are. The basic idea is to take the conjunction of A and B, but make their languages disjoint, to prevent any "clashes" from occurring.

In order to study the lattice of finite structures modally, i.e. using ILP, we have to work with extensions of some finite base theory which is sufficiently strong for ILP. Otherwise we cannot even be sure that ILP is the interpretability logic of all the theories comprising the lattice. Thus we would be considering a substructure of $(\mathrm{F}, \rhd)$, namely $(\mathrm{V}_T, \rhd)$, where T is a finitely axiomatized theory containing $\mathrm{I}\Sigma_1$. By the same arguments as above (using arithmetical soundness of IL), we know that T verifies that $(\mathrm{V}_T, \rhd)$ is a lower semilattice.

However, it is not clear that T also verifies that any two degrees in $(\mathrm{V}_T, \rhd)$ have a supremum. In fact, we do not even know how to express the supremum in the language of T. The problem is that Friedman's construction is not directly suitable for carrying over to a modal context. It is unclear how to implement in modal logic a procedure of making the languages disjoint. For the modal-logical perspective, it would be convenient to have a sentence $A \owedge B$ *in the language of* T which is *verifiably in* T in the degree of the supremum of A and B. Unfortunately, it is an open question whether such a sentence exists (see [**Šve78**]). Hence it is unclear how one should proceed in order to extend ILP with a supremum operator.

CHAPTER 3

Uniform Suprema in Arithmetic

This chapter deals with the arithmetical side of the supremum. The research we present is preparatory work for extending the modal system ILM with a supremum operator $\between$, and finding a modal semantics for the resulting system ILMS. Apart from ILM, the logic ILMS contains the defining equation for $\between$, i.e. the axiom S: $(C \rhd A) \wedge (C \rhd B) \leftrightarrow C \rhd A \between B$. The intended meaning of the modal symbol $\between$ is a (formalized) supremum operator in the lattice[1] $(\mathrm{V}_{\mathsf{PA}}, \rhd)$. As in the case of the modal symbols $\Box$ and $\rhd$, the precise arithmetical meaning of $\between$ will be fixed via the notion of an arithmetical realization. We would thus need to extend the notion of an arithmetical realization for ILM by a clause of the form

$$(A \between B)^* = \theta_{A^*B^*} \tag{11}$$

where $\theta_{A^*B^*}$ is verifiably in PA a supremum of A^* and B^*, i.e. for all C,

$$\vdash_{\mathsf{PA}} (C \rhd A^*) \wedge (C \rhd B^*) \leftrightarrow C \rhd \theta_{A^*B^*}. \tag{12}$$

Given sentences A^* and B^*, the procedure for finding the sentence $\theta_{A^*B^*}$ should of course be as simple as possible. The most elegant option would be to have a formula $\sigma(x, y)$ with two free variables s.t. we can simply put $(A \between B)^* = \sigma(\overline{\ulcorner A^* \urcorner}, \overline{\ulcorner B^* \urcorner})$. This is how arithmetical realizations deal with our other modal symbols $\Box$ and $\rhd$. The formula $\sigma(x, y)$ would be a *uniform* (in A and B) way of finding a verifiable supremum of A and B in the lattice $(\mathrm{V}_{\mathsf{PA}}, \rhd)$. A formula $\sigma(x, y)$ with such properties is called an implementation[2] of the supremum in PA. As we will see below, not all implementations of the supremum in PA are provably equivalent. This means that if $\sigma(x, y)$ and $\tau(x, y)$ are implementations, we do not necessarily have for all A and B that $\vdash_{\mathsf{PA}} \sigma(\overline{\ulcorner A \urcorner}, \overline{\ulcorner B \urcorner}) \leftrightarrow \tau(\overline{\ulcorner A \urcorner}, \overline{\ulcorner B \urcorner})$.

Section 1 below introduces the notion of an implementation. We discuss possible properties of implementations, as well as methodological points concerning the notion of an arithmetical realization for ILMS. Section 2 contains the preliminaries for constructing well-behaved implementations of the supremum in PA. Two such implementations — Švejdar's implementation and Visser's implementation — are then studied in more detail in sections 4 and 5. As this thesis is the first treatment of the system ILMS, the material contained in this chapter is mostly new. Exceptions are Section 2, and some (important) results in Section 4. Throughout this chapter, $\vdash$ will denote provability in PA.

[1] Remember from Section 3 of Chapter 2 that $(\mathrm{V}_{\mathsf{PA}}, \rhd)$ is the lattice of finite extensions of PA under the relation of interpretability.

[2] In principle, we could also allow as implementations computable functions which, given as input sentences A^* and B^* of arithmetic, produce a sentence $\theta_{A^*B^*}$ as in (12) above. However, since we have at our disposal formulas $\sigma(x, y)$ that do the job, we shall not bother ourselves with a more complicated notion of an implementation here.

1. Implementations of the Supremum in PA

The system ILMS would allow us to study in a modal-logical setting what is provable in PA about the lattice $(\mathrm{V}_{\mathsf{PA}}, \rhd)$. For example, we might wonder whether PA proves that $(\mathrm{V}_{\mathsf{PA}}, \rhd)$ is distributive and dense, and that the supremum of consistent degrees is always consistent. To begin with, we want PA to verify that $(\mathrm{V}_{\mathsf{PA}}, \rhd)$ is indeed a lattice, i.e. that any two elements have an infimum and a supremum.

Recall from Section 3.2 of Chapter 2 that the case of the infimum is easy. Reasoning in IL (using in particular axiom J3), we see that $A \vee B$ is verifiably in PA an infimum of A and B, i.e. for all A, B, C

$$\vdash (A \rhd C) \wedge (B \rhd C) \leftrightarrow A \vee B \rhd C. \tag{13}$$

Rember also that when it comes to the supremum, we have to be more clever. What we need is made explicit by the notion of an implementation of the supremum in PA.

Definition 35. An *implementation of the supremum in* PA is a formula $\sigma(x, y)$ s.t. for all A, B, C,

$$\vdash (C \rhd A) \wedge (C \rhd B) \leftrightarrow C \rhd \sigma(\overline{\ulcorner A \urcorner}, \overline{\ulcorner B \urcorner}). \tag{14}$$

Examples of implementations will be given later in this chapter. For now, we will just assume that a formula $\sigma(x, y)$ with the required properties can be found. We will from now on write $\sigma(A, B)$ instead of $\sigma(\overline{\ulcorner A \urcorner}, \overline{\ulcorner B \urcorner})$.

Equipped with an implementation of the supremum, we can investigate what is provable about it in PA. By arithmetical soundness of ILM, we know that apart from the defining equation (14), all its consequences in ILM have to be provable. In a way, this is not much. For any choice of A and B, (14) only fixes the degree of the sentence $\sigma(A, B)$ — its behaviour w.r.t. the more fine-grained provability properties is left undetermined. As an example, note that if $\sigma(x, y)$ is an implementation, we do not necessarily have

$$\vdash \sigma(A, B) \leftrightarrow \sigma(B, A) \tag{15}$$

for all A and B. The only thing guaranteed by (14) is that $\vdash \sigma(A, B) \equiv \sigma(B, A)$. For the same reason, we do not necessarily have

$$\vdash A \leftrightarrow \sigma(A, A). \tag{16}$$

An implementation satisfying (15) or (16) is said to be *commutative* or *idempotent* respectively. Similarly for other possible properties like distributivity[3] or associativity. Further important properties are extensionality and monotonicity.

Definition 36. A implementation $\sigma(x, y)$ is *extensional* if for all A, A', B, B',

$$\vdash \Box(A \leftrightarrow A') \wedge \Box(B \leftrightarrow B') \rightarrow \Box(\sigma(A, B) \leftrightarrow \sigma(A', B')).$$

An implementation $\sigma(x, y)$ is *monotone* if for all A, A', B, B',

$$\vdash \Box(A \rightarrow A') \wedge \Box(B \rightarrow B') \rightarrow \Box(\sigma(A, B) \rightarrow \sigma(A', B')).$$

[3] An implementation $\sigma(x, y)$ can be said to have a property such as distributivity in two ways. If we have $\vdash \sigma(A, B \vee C) \leftrightarrow \sigma(A, B) \vee \sigma(A, C)$ for all A, B and C, we say that the implementation is distributive w.r.t. provability. If $\vdash \sigma(A, B \vee C) \equiv \sigma(A, B) \vee \sigma(A, C)$, we say that the implementation is distributive w.r.t. interpretability. Similarly for other possible properties.

1.1. Closure properties of implementations. Let $\sigma(x, y)$ be an implementation. If we want to show that a formula $\tau(x, y)$ is also an implementation, it clearly suffices to show that for all A and B, $\tau(A, B)$ is (verifiably in PA) in the same degree as $\sigma(A, B)$, i.e. $\vdash \tau(A, B) \equiv \sigma(A, B)$. We will use this observation to show how new implementations can be generated from given ones.

First,we can obtain a new implementation by "iterating" an old one, i.e. letting $\tau_0(x, y) := \sigma(\sigma(x, y), y)$, $\tau_1(x, y) := \sigma(\sigma(x, y), \sigma(x, y))$, and so on. It is clear that this process of "iterating" does not change the degree of the resulting sentence — and this is all that matters for being a supremum. A second way to obtain new implementations is to define them by case distinction. Suppose that we have two implementations $\sigma(x, y)$ and $\sigma'(x, y)$. Let $A(x, y)$ be some formula. We can then define a new implementation $\tau(x, y)$ by letting

$$\tau(x, y) := (A(x, y) \to \sigma(x, y)) \wedge (\neg A(x, y) \to \sigma'(x, y)).$$

As an example, given an implementation $\sigma(x, y)$, we could define an implementation $\tau(x, y)$ s.t. for all A, $\tau(A, A) = A$, and if $B \neq A$ then $\tau(A, B) = \sigma(A, B)$. In general, such an implementation is not extensional. In fact, even the informal version of extensionality can fail, i.e. it is possible that $\vdash A \leftrightarrow B$, but $\nvdash \tau(C, A) \leftrightarrow \tau(C, B)$. To see this, note that if $\vdash (A \leftrightarrow B) \wedge A \neq B$, then by definition $\tau(A, B) = \sigma(A, B)$, but on the other hand $\tau(A, A) = A$.

We can also find new implementations by using the closure properties of the degrees. If $\sigma_1(x, y)$ and $\sigma_2(x, y)$ are implementations, then clearly for all A and B, $\vdash \sigma_1(A, B) \equiv \sigma_2(A, B)$. But then also $\vdash \sigma_1(A, B) \vee \sigma_2(A, B) \equiv \sigma_1(A, B)$, since $\sigma_1(A, B) \vee \sigma_2(A, B)$ is (verifiably in PA) the infimum of $\sigma_1(A, B)$ and $\sigma_2(A, B)$ in $(\mathrm{V}_{\mathsf{PA}}, \rhd)$. It follows that $\sigma_1(x, y) \vee \sigma_2(x, y)$ is also an implementation, i.e. implementations are closed under disjunction[4]. As a consequence, each implementation has a commutative version. If $\sigma(x, y)$ is not commutative, we can just take $\sigma'(x, y) := \sigma(x, y) \vee \sigma(y, x)$, which is clearly commutative. Two further closure properties of the degrees can be found by recalling that $\vdash_{\mathsf{IL}} A \equiv A \wedge \Box\neg A$ and $\vdash_{\mathsf{IL}} A \equiv A \vee \Diamond A$ (see Lemma 2.1 in Chapter 2). Hence if $\sigma(x, y)$ is an implementation, then so are $\tau(x, y) := \sigma(x, y) \wedge \Box\neg\sigma(x, y)$, and $\tau'(x, y) := \sigma(x, y) \vee \Diamond\, \sigma(x, y)$.

1.2. Arithmetical realizations for ILMS. The potential multitude of implementations gives us some freedom in choosing the intended arithmetical meaning of the symbol $\between$ in ILMS. While the provability properties of an implementation are in some sense irrelevant (after all, it is behaviour w.r.t. interpretability that determines whether a sentence is a supremum), the latter will inevitably play a role in our modal system ILMS. We have to know how strong the system ILMS should be. For example, should we have that $\vdash_{\mathsf{ILMS}} A \between B \leftrightarrow B \between A$, or that $\vdash_{\mathsf{ILMS}} A \between A \leftrightarrow A$?

As suggested by the discussion above, the answer to these questions depends on which implementation of the supremum we choose as the intended meaning of $\between$. For example, if we allow *any* implementation, $A \between B \leftrightarrow B \between A$ should certainly

[4]If $A \equiv A'$ then we do not necessarily have that $\vdash A \wedge A' \equiv A$. We see a concrete example in Section 5.3 below where Orey-sentences are introduced. It follows that implementations are not closed under conjunction.

not[5] be a theorem of ILMS. But we might also want to consider commutative implementations only, in which case $A \owedge B \leftrightarrow B \owedge A$ should obviously be a theorem of ILMS.

Arithmetical realizations for ILMS should therefore be conceived of as arithmetical realizations for ILM that have an implementation of the supremum as a parameter. Write $*(\sigma)$ for the arithmetical realization $*$ parametrized by the implementation $\sigma(x, y)$. Before formulating arithmetical soundness and completeness for the system ILMS, one has to fix a set $\mathcal{I}$ of implementations. Arithmetical completeness and soundness can then be formulated as follows:

$$\vdash_{\mathsf{ILMS}} A \;\Leftrightarrow\; \forall \text{ realizations } *, \forall \sigma \in \mathcal{I}, \vdash_{\mathsf{PA}} A^{*(\sigma)}.$$

By choosing as $\mathcal{I}$ the set of all implementations, we get a minimal logic for the supremum. This logic would be quite weak. For example, for the reasons given above neither $A \owedge B \leftrightarrow B \owedge A$ nor $A \leftrightarrow A \owedge A$ should be a theorem of the minimal logic. The minimal logic will be studied Chapter 4. Alternatively, we could choose $\mathcal{I} = \{\sigma(x, y)\}$, aiming to find a logic that captures what is provable in PA about the specific implementation $\sigma(x, y)$. This strategy is employed in Chapter 5, where we explore the possibility of having a modal semantics for a specific implementation of the supremum. A middle way is to consider any implementation satisfying certain properties, e.g. by letting $\mathcal{I}$ to be the set of all implementations that are extensional.

2. Arithmetical Preliminaries

In order to find well-behaved implementations of the supremum in PA, we will make use of a "stratified" notion of provability. We will view provability in PA as being split up into provability in a sequence $\{T_n\}_{n\in\omega}$ of finitely axiomatized theories. We also require the T_n's to extend each other, i.e. we want that if $\vdash_{T_m} A$, then $\vdash_{T_n} A$ for all $n > m$.

In his classical paper [**Fef60**], Feferman uses this perspective on provability to construct an intensionally abnormal but extensionally correct provability predicate $\mathrm{Pr}_F(x)$ s.t.

$$\vdash \neg \mathrm{Pr}_F(\overline{\ulcorner \bot \urcorner}),$$

seemingly contradicting Gödel's Second Incompleteness Theorem[6]. The stratification applied by Feferman is

(17) $\quad T_n =$ the theory axiomatized by the axioms of PA of gödelnumber $\leq n$.

Under this choice, it is clear that the T_n's are finitely axiomatizable, increasing in strength, and $\mathsf{PA} = \bigcup_{n\in\omega} T_n$. But not much more can be said. The exact content of T_n and its relation to other theories in the sequence will depend on arbitrary details of coding. Indeed, having more control over the stratification sequence

[5]Given an implementation $\sigma(x, y)$, we can find a non-commutative version of it by defining a new implementation by case distinction. As an example, consider the implementation $\tau(x, y)$ defined as $\tau(x, y) := (x \leq y \to \sigma(x, y) \wedge \Box\neg\sigma(x, y)) \wedge (y < x \to \sigma(x, y) \vee \Diamond\sigma(x, y))$. If $\tau(x, y)$ would be commutative w.r.t. provability, we would have for all x, y that $\Diamond\sigma(x, y) \to \Box\neg\sigma(x, y)$, which is clearly not the case.

[6]The solution to the seeming paradox is that one has to be more careful when formulating the Second Incompleteness Theorem — it does not apply to provability predicates which are intensionally abnormal such as $\mathrm{Pr}_F(x)$.

will sometimes allow us to prove stronger results. This influence of the choice of stratification was first studied by Smoryński in [**Smo89**], inspired by McAloon's work on Rosser-like sentences for ZF ([**McA75**]). We will now introduce Rosser sentences, explaining how stronger results about them can be proved when using a more elaborate stratification sequence than the one in (17).

2.1. Rosser sentences. Rosser-sentences were introduced by Barkley Rosser in [**Ros36**]. Rosser-sentences are like Gödel-sentences, in that they are independent of PA. Their advantage over the latter is that the proof of their independence from PA does not need the assumption of Σ_1-soundness.

Let $\mathrm{Prf}(x, y)$ be a proof predicate of PA (see Section 1.2 of Chapter 2). Thus $\vdash \mathrm{Prf}(\overline{p}, \overline{n})$ if and only if p codes a proof of the formula coded by n. A Rosser sentence is a sentence R (guaranteed to exist by the Gödel Fixed Point Theorem) with the property

$$\vdash R \leftrightarrow \forall x \, (\mathrm{Prf}(x, \overline{\ulcorner R \urcorner}) \rightarrow \exists y < x \, \mathrm{Prf}(y, \overline{\ulcorner \neg R \urcorner})) \tag{18}$$

Given in this way, two Rosser sentences will not necessarily be provably equivalent, thus the situation is different from Gödel-sentences which can be proven to be unique. This failure of uniqueness has been studied in [**GS79**].

Smoryński ([**Smo89**]) uses the idea of stratified provability to obtain Rosser-like sentences which are unique up to provable equivalence. Informally, the original Rosser sentence says: "If there is a proof of me, there is a smaller proof of my negation" (the "smaller" here refers to the codes of the proofs). Instead, Smoryński considers a sentence saying: "If there is a proof of me, there is an *earlier* proof of my negation". The notion of "earlier" here is made precise by the concept of stratified provability. The Rosser-like sentence constructed by Smoryński says: "If I am provable in some theory T_n, then my negation is provable in a theory T_m with $m < n$."

We introduce some notation for expressing Smoryński's Rosser-like sentence. Let $\{T_n\}_{n \in \omega}$ be a sequence of finitely axiomatized theories with T_{n+1} extending T_n, and $\mathsf{PA} = \bigcup_{n \in \omega} T_n$. Write $\vdash_n$ for provability in T_n, and $\Box_n$ for the provability predicate[7] of T_n. Smoryński's Rosser-like sentence is a sentence ρ (guaranteed to exist by the Gödel Fixed Point Theorem) with the property

$$\vdash \rho \leftrightarrow \forall x \, (\Box_x \rho \rightarrow \exists y < x \, \Box_y \neg \rho). \tag{19}$$

In order to prove the uniqueness of ρ, a more elaborate choice of stratification than in (17) is needed. Smoryński uses:

$$T_n = \mathrm{I}\Sigma_{n+1}. \tag{20}$$

Equivalently, one could use

$$T_0 := \mathsf{EA}, \qquad T_{n+1} := \mathrm{I}\Sigma_{n+1}. \tag{21}$$

Recall from Section 1.1 of Chapter 2 that $\mathrm{I}\Sigma_n$ is the theory PA with induction restricted to Σ_n-formulas. The advantage of this choice is that the T_n's provably

[7]Since T_n is finitely axiomatized, there is a canonical choice for the formula $\tau_n(x)$ defining the axioms of T_n — $\tau_n(x)$ is obtained by simply listing all the axioms of T_n (c.f. [**Fef60**])

grow in strength — for each n, $\mathrm{I}\Sigma_{n+1}$ proves uniform Π_{n+2}-reflection for $\mathrm{I}\Sigma_n$. This means that for all $A(y) \in \Pi_{n+2}$,

$$\vdash_{n+1} \forall y\,(\Box_n A(y) \to A(y)).$$

For a proof of this fact, see e.g. [**Sie85**] or [**Ono87**]. Since a consistency statement for T_n is just the reflection principle for $\bot$ (i.e. $\Box_n \bot \to \bot$), it follows that T_{n+1} proves the consistency of T_n.

It is shown that under this choice of stratification, the Rosser-like sentence ρ in (19) is unique up to provable equivalence. Smoryński also shows that uniqueness can fail under a different choice of stratification sequence.

Background and information about the choice in (21) is provided in [**SA12**], where Shavrukov and Visser use it to prove uniform density for the Lindenbaum algebra of PA. In [**Sha94**], Shavrukov uses the stratification sequence in (21) to construct a joint provability logic for the standard provability predicate and the Feferman provability predicate.

Our uniform implementations of the supremum turn out to be similar to Smoryński's Rosser-like sentences. In order to formulate them in the first place, we need some notion of stratified provability. In order to prove their uniqueness and extensionality, an elaborate choice of stratification — as in (20) or (21) — seems to be needed.

2.2. Elaborate stratifications. We introduce the features which from now on any stratification $\{T_n\}_{n\in\omega}$ of PA is assumed to possess. As before, let $\vdash_n$ denote provability in T_n, and $\Box_n$ the provability predicate of T_n. First, we require the T_n's to be finitely axiomatizable and contain EA, whence we can be sure that GL is the provability logic of all T_n. Furthermore, we require

1. $\mathsf{PA} = \bigcup_{n\in\omega} T_n$
2. $T_n \subseteq T_{n+1}$,
3. $\vdash_{n+1} \forall y\,[A \in \Pi_{n+2} \to (\Box_n A(y) \to A(y))]$.

In fact, we need properties 1 - 3 to be verifiable in T_1, i.e. we want T_1 to prove:

i. $\Box A \leftrightarrow \exists x \Box_x A$

ii. $\Box_x A \wedge x < y \to \Box_y A$

iii. $\Box_{x+1}(A \in \Pi_{x+2} \to \forall y\,(\Box_x A(y) \to A(y)))$

In ii and iii, x and y are free variables, i.e. they range also over nonstandard elements. It follows from i and ii that $\vdash_1 \Diamond A \leftrightarrow \forall_x \Diamond_x A$, and $\vdash_1 \Box_0 A \leftrightarrow \forall x\, \Box_x A$. In practice, we will not bother with the complexity of the sentence[8] A for which we want reflection. In the end, we are interested in provability in PA, thus we can

[8]In the proofs below we will never use reflection for formulas containing free variables. Thus sentential reflection would suffice for our purposes, and we could also choose a stratification that grows more slowly than the one in (21).

always choose a level which is "high" enough for having reflection for A. We will refer to property ii as *monotonicity*[9], and to property iii as *reflection*.

One way to get the above properties is to choose a stratification sequence as in (21). As mentioned in Section 1.1 of Chapter 2, both EA and Σ_{n+1} (for all n) are finitely axiomatizable. It is clear that we will have $\mathsf{EA} \subseteq T_n$ for all n. That properties i and ii hold is immediate from the definitions. The proof of property iii can be traced down to the proof of Corollary 4.4. of [**Sie85**], or Exercise 10.8 of [**Kay91**].

In the rest of this chapter, we will just refer to the axioms of GL to justify steps in the proofs in PA and in T_n (thus applying the arithmetical soundness of GL). We will write $\Box$ and $\Box_x$ for the provability predicates of PA and T_x respectively. Accordingly, we also write $\Diamond A$ instead of $\neg\Box\neg A$, and $\Diamond_x A$ instead of $\neg\Box_x\neg A$. By property ii, $\Diamond_x A$ implies $\Diamond_y A$ for all $y \leq x$, and by property iii we have that $\vdash_1 \Box_{x+1}(A \in \Sigma_{x+1} \rightarrow (A \rightarrow \Diamond_x A))$.

3. A True but Unprovable Principle for the Supremum

Before we go on to study the specific implementations of the supremum, we present a result that holds for all implementations but is nevertheless not obtainable by using only the axioms of ILMS.

Recall Theorem 33 from Chapter 2. If A and B are both consistent, then so is their supremum in the lattice $(\mathrm{V}_{\mathsf{PA}}, \rhd)$. We now ask whether this fact is verifiable inside PA, i.e. whether for all A and B, and for all implementations $\sigma(x, y)$,

$$\vdash \Diamond A \wedge \Diamond B \rightarrow \Diamond\, \sigma(A, B). \tag{22}$$

The answer to this question turns out to be strongly negative: in fact there is *no* implementation for which (22) is provable. We will show this by giving a counterexample to (22) that works for any implementation. Until the end of this subsection, we will write $A \owedge B$ for the sentence $\sigma(A, B)$ (where $\sigma(x, y)$ is an implementation).

The only property of the supremum that we use is given by the Orey-Hájek characterization. Since $A \owedge B \rhd A$, $A \owedge B$ proves all Π_1-consequences of A, thus in particular it proves A if A is Π_1 itself. Hence for A, B in Π_1, $\vdash A \owedge B \rightarrow (A \wedge B)$, and by necessitation also $\vdash \Box(A \owedge B \rightarrow A \wedge B)$. Note that for Π_1-sentences A and B, this means that $A \wedge B$ and $A \owedge B$ are in the same degree (this fact was also pointed out by Švejdar in [**Šve78**]).

We will use Rosser sentences to construct a counterexample to (22). Let R be the *negation* of the Rosser sentence in (18). Thus R is s.t.

$$\vdash R \leftrightarrow \exists x\, (\mathrm{Prf}(x, \overline{\ulcorner \neg R \urcorner}) \wedge \forall y < x\, \neg \mathrm{Prf}(y, \overline{\ulcorner R \urcorner})). \tag{23}$$

Let $R^{\perp}$ (R opposite) be the sentence

$$\exists x\, (\mathrm{Prf}(x, \overline{\ulcorner R \urcorner}) \wedge \forall y \leq x\, \neg \mathrm{Prf}(y, \overline{\ulcorner \neg R \urcorner})). \tag{24}$$

It is easy to see that R and $R^{\perp}$ are both Σ_1. Note also that R and $R^{\perp}$ have the following properties:

[9] It will always be clear from the context whether *monotonicity* has the meaning specified here, or the one introduced in the previous section where it refers to a possible property of an implementation.

1. $\vdash \Box R \to R \vee R^\perp$ and $\vdash \Box R^\perp \to R \vee R^\perp$
2. $\vdash R \vee R^\perp \leftrightarrow \Box\bot$
3. $\vdash \Box R \leftrightarrow \Box\bot$ and $\vdash \Box R^\perp \leftrightarrow \Box\bot$

For 1, note that $\vdash R^\perp \to \neg R$, whence also $\vdash \Box(R^\perp \to \neg R)$. One direction of 2 follows immediately from 1. For the other direction, we use that R and $R^\perp$ are Σ_1 whence $\vdash R \to \Box R$ and $\vdash R^\perp \to \Box R^\perp$ (and thus also $\vdash R^\perp \to \Box\neg R$) by provable Σ_1-completeness. On the other hand, by the properties of R and $R^\perp$ also $R \to \Box\neg R$ and $R^\perp \to \Box R$. Combining all these facts, it is easy to see that $\vdash R \vee R^\perp \to \Box\bot$. The non-trivial directions of 3 follow from 1 and 2.

LEMMA 3.1. $\vdash \Diamond(\neg R \owedge \neg R^\perp) \to \neg\Box\Box\bot$

PROOF. By contraposition. We will show that $\vdash \Box\Box\bot \to \Box\neg(\neg R \owedge \neg R^\perp)$.

We have $\Box\Box\bot \to \Box(R \vee R^\perp)$ by property 2 above and necessitation. By Π_1-conservativity (since $\neg R$ and $\neg R^\perp$ are Π_1), we have

$$\vdash \Box(\neg R \owedge \neg R^\perp \to \neg R \wedge \neg R^\perp),$$

whence

$$\vdash \Box(R \vee R^\perp \to \neg(\neg R \owedge \neg R^\perp)).$$

Using propositional logic and L1, we are done. ❑

By property 3 of R and $R^\perp$, we have also

$$\vdash \Diamond\neg R \wedge \Diamond\neg R^\perp \leftrightarrow \neg\Box\bot. \tag{25}$$

We now have enough tools to give a counterexample to (22).

THEOREM 37. *There are A and B with $\nvdash \Diamond A \wedge \Diamond B \to \Diamond(A \owedge B)$.*

PROOF. Take $\neg R$ for A and $\neg R^\perp$ for B. Suppose for contradiction that

$$\vdash \Diamond\neg R \wedge \Diamond\neg R^\perp \to \Diamond(\neg R \owedge \neg R^\perp).$$

By Lemma 3.1 and (25), this implies that $\vdash \neg\Box\bot \to \neg\Box\Box\bot$, i.e. $\vdash \Box\Box\bot \to \Box\bot$ and so by Löb's Theorem $\vdash \Box\bot$, contradiction. ❑

Thus PA does not verify that the supremum of two consistent degrees in the lattice $(\mathrm{V}_{\mathsf{PA}}, \rhd)$ is always consistent — regardless of which implementation of the supremum we choose to represent the supremum in PA.

4. Švejdar's Implementation of the Supremum

This section deals with an implementation of the supremum in PA that is based on Švejdar's supremum. Recall (section 3.3 of Chapter 2) that Švejdar's supremum of given sentences A and B is found by applying the Gödel Fixed Point Theorem to find a sentence θ with

$$\vdash \theta \leftrightarrow \forall x\,(\mathrm{Con}_{\mathsf{PA}\restriction x}(\overline{\ulcorner\theta\urcorner}) \to \mathrm{Con}_{\mathsf{PA}\restriction x}(\overline{\ulcorner A\urcorner}) \wedge \mathrm{Con}_{\mathsf{PA}\restriction x}(\overline{\ulcorner B\urcorner})). \tag{26}$$

Remember that $\mathrm{Pr}_{\mathsf{PA}\restriction n}(x)$ is the provability predicate of the theory axiomatized by axioms of PA whose gödelnumber is less than n, and that we write $\mathrm{Con}_{\mathsf{PA}\restriction n}(\overline{\ulcorner A\urcorner})$ for

$\neg \mathrm{Pr}_{\mathsf{PA}\restriction n}(\overline{\ulcorner \neg A \urcorner})$. Note that in order to express (the right hand side of) θ, a simple stratification sequence as in (17) above is used. We will reformulate Švejdar's supremum using the more elaborate choice of a stratification sequence, as well as the notation introduced in Section 2.2. Thus instead of (26) we will write

$$\vdash \theta \leftrightarrow \forall x\,(\Diamond_x \theta \rightarrow \Diamond_x A \wedge \Diamond_x B), \tag{27}$$

thinking of $\Box_x$ as the provability predicate of T_x, where $\{T_n\}_{n\in\omega}$ is a stratification sequence as in Section 2.2.

4.1. Švejdar's supremum as an implementation. In order to extract an implementation out of Švejdar's construction of the supremum, two things have to be accomplished. First, we need Švejdar's supremum to be verifiable in PA. If θ is as above, we need that

$$\vdash (C \rhd A) \wedge (C \rhd B) \leftrightarrow C \rhd \theta. \tag{28}$$

Remember that since Švejdar's aim was to find out what is *true* about the lattice $(\mathrm{V}_{\mathsf{PA}}, \rhd)$, he did not need to worry about his supremum being verifiable in PA. Second, we need to show that there is a formula $\sigma(x, y)$ s.t. for all A and B, $\sigma(A, B)$ is the sentence θ with $\vdash \theta \leftrightarrow \forall x\,(\Diamond_x \theta \rightarrow \Diamond_x A \wedge \Diamond_x B)$. Švejdar's construction does not immediately offer us a formula $\sigma(x, y)$ with the required properties. Švejdar supremum is produced by first constructing a formula containing (the gödelnumbers of) A and B, and then applying the Fixed Point Theorem to get a fixed point of this formula. We will first show that such a formula $\sigma(x, y)$ can be found, and then that for all A and B, the sentence $\sigma(A, B)$ is *verifiably* in PA a supremum of A and B.

Using the general version of the Fixed Point Theorem given in Section 1.2 of Chapter 2, we can turn Švejdar's construction into a formula $\sigma(x, y)$ with the required properties. Let $\mathrm{Sub}(x, y, z)$ be the function representing[10] the substitution function in EA. Thus for all m, n, and A,

$$\vdash_0 \mathrm{Sub}(\overline{\ulcorner A(u, v) \urcorner}, \overline{m}, \overline{n}) = \overline{\ulcorner A(\overline{m}, \overline{n}) \urcorner},$$

and by necessitation

$$\vdash_0 \Box_0(\mathrm{Sub}(\overline{\ulcorner A(u, v) \urcorner}, \overline{m}, \overline{n}) = \overline{\ulcorner A(\overline{m}, \overline{n}) \urcorner}). \tag{29}$$

Consider the formula

$$A(w, y, z) := \forall x\,(\Diamond_x \mathrm{Sub}(w, y, z) \rightarrow \Diamond_x y \wedge \Diamond_x z). \tag{30}$$

We apply the Gödel-Carnap Fixed Point Theorem (Theorem 1.1) to find a fixed point $\sigma(y, z)$ of $A(w, y, z)$. Thus we have

$$\vdash \sigma(y, z) \leftrightarrow \forall x\,(\Diamond_x \mathrm{Sub}(\overline{\ulcorner \sigma(u, v) \urcorner}, y, z) \rightarrow \Diamond_x y \wedge \Diamond_x z). \tag{31}$$

Substituting $\overline{\ulcorner A \urcorner}$ and $\overline{\ulcorner B \urcorner}$ for the free variables, we get

$$\vdash \sigma(\overline{\ulcorner A \urcorner}, \overline{\ulcorner B \urcorner}) \leftrightarrow \forall x\,(\Diamond_x \mathrm{Sub}(\overline{\ulcorner \sigma(u, v) \urcorner}, \overline{\ulcorner A \urcorner}, \overline{\ulcorner B \urcorner}) \rightarrow \Diamond_x \overline{\ulcorner A \urcorner} \wedge \Diamond_x \overline{\ulcorner B \urcorner}), \tag{32}$$

which by (29) becomes

$$\vdash \sigma(\overline{\ulcorner A \urcorner}, \overline{\ulcorner B \urcorner}) \leftrightarrow \forall x\,(\Diamond_x \overline{\ulcorner \sigma(\overline{\ulcorner A \urcorner}, \overline{\ulcorner B \urcorner}) \urcorner} \rightarrow \Diamond_x \overline{\ulcorner A \urcorner} \wedge \Diamond_x \overline{\ulcorner B \urcorner}). \tag{33}$$

[10] Recall from Section 1.2 of Chapter 2 that we assume to have function symbols for all Kalmar elementary functions available in EA.

Using our sloppy notation:

$$\vdash \sigma(A,B) \leftrightarrow \forall x\,(\Diamond_x \sigma(A,B) \rightarrow \Diamond_x A \wedge \Diamond_x B). \tag{34}$$

Thus the sentence $\sigma(A,B)$ has exactly the properties of Švejdar's supremum. We will from now on write $(A \sqcap B)$ for the sentence $\sigma(A,B)$ in (34). The symbol $\sqcap$ should remind you of the universal quantifier on the right hand side of $\sigma(A,B)$. We have the convention that $\sqcap$ binds equally strong as $\wedge$ and $\vee$, thus we shall e.g. write $A \sqcap B \rightarrow A$ instead of $(A \sqcap B) \rightarrow A$.

We will now show that for any A and B, $A \sqcap B$ is verifiably in PA in the degree of the supremum of A and B.

THEOREM 38. *For all C, $\vdash (C \rhd A) \wedge (C \rhd B) \leftrightarrow C \rhd A \sqcap B$.*

PROOF. We will give an informal proof, a version of which can be found in [**Šve78**]. Using the verifiability of the Orey-Hájek characterization and the properties of our elaborate stratification sequence, it is easy to see that the proof can be verified in PA. Recall that $A \sqcap B$ is a sentence s.t.

$$\vdash A \sqcap B \leftrightarrow \forall x\,(\Diamond_x(A \sqcap B) \rightarrow \Diamond_x A \wedge \Diamond_x B).$$

We will first show that $A \sqcap B \rhd A$ and $A \sqcap B \rhd B$. By essential reflexivity we have $A \sqcap B \vdash \Diamond_n(A \sqcap B)$ for all $n \in \mathbb{N}$, whence also $A \sqcap B \vdash \Diamond_n A \wedge \Diamond_n B$ by the properties of $A \sqcap B$. By the Orey-Hájek characterization of interpretability, $A \sqcap B \rhd A$ and $A \sqcap B \rhd B$.

For the other direction, assume $C \rhd A$ and $C \rhd B$. We will show that

$$C \wedge \neg(A \sqcap B) \rhd A \sqcap B.$$

Since clearly $C \wedge (A \sqcap B) \rhd A \sqcap B$, using an interpretation defined by case distinction we get $C \wedge ((A \sqcap B) \vee \neg(A \sqcap B)) \rhd A \sqcap B$, i.e. $C \rhd A \sqcap B$. So consider $C \wedge \neg(A \sqcap B)$. $\neg(A \sqcap B)$ is the sentence

$$\exists x\,(\Diamond_x(A \sqcap B) \wedge (\Box_x \neg A \vee \Box_x \neg B)).$$

As we have assumed $C \rhd A$ and $C \rhd B$, by the Orey-Hàjek characterization we have for all $n \in \mathbb{N}$,

$$C \vdash \Diamond_n A \wedge \Diamond_n B. \tag{35}$$

Fix $n \in \mathbb{N}$. It follows from (35) that

$$C \vdash \forall x\,(\Diamond_x(A \sqcap B) \wedge (\Box_x \neg A \vee \Box_x \neg B) \rightarrow x > n),$$

whence,

$$C \wedge \neg(A \sqcap B) \vdash \exists x\,((\Diamond_x(A \sqcap B) \wedge (\Box_x \neg A \vee \Box_x \neg B)) \wedge x > n).$$

In particular,

$$C \wedge \neg(A \sqcap B) \vdash \exists x\,(\Diamond_x(A \sqcap B) \wedge x > n),$$

and so by monotonicity,

$$C \wedge \neg(A \sqcap B) \vdash \Diamond_n(A \sqcap B).$$

Thus $C \wedge \neg(A \sqcap B) \rhd A \sqcap B$ follows by the Orey-Hàjek characterization. ❑

REMARK 39. Using the properties of the elaborate stratification, the direction from left to right can also be proved by showing that for all n,

$$\vdash_{n+1} \Diamond_n A \wedge \Diamond_n B \to \Diamond_n (A \sqcap B)$$

and applying the Orey-Hájek characterization. ❑

As a consequence of Theorem 38, the formula $\sigma(y, z)$ in (31) is indeed an implementation of the supremum in the sense of Definition 35. We call this implementation *Švejdar's implementation*.

4.2. Properties of Švejdar's implementation. We will now examine some properties of Švejdar's implementation. First, it is obvious that Švejdar's implementation is commutative w.r.t. provability, i.e. that for all A and B, $\vdash A \sqcap B \leftrightarrow B \sqcap A$.

According to the next theorem, Švejdar's implementation is distributive w.r.t. interpretability.

THEOREM 40 (Distributivity of $\sqcap$). $\vdash A \sqcap (B \vee C) \equiv (A \sqcap B) \vee (A \sqcap C)$, *and* $\vdash A \vee (B \sqcap C) \equiv (A \vee B) \sqcap (A \vee C)$.

PROOF. We will present an informal proof similar[11] to the one given by Švejdar in [**Šve78**]. Observing the reasoning used in the proof, it is clear that it can be formalized in PA.

We only need to show that $\vdash A \sqcap (B \vee C) \equiv (A \sqcap B) \vee (A \sqcap C)$ as the other distributivity law follows from this one by using principles available in IL. Using the latter, we can also establish that the direction $(A \sqcap B) \vee (A \sqcap C) \rhd A \sqcap (B \vee C)$ automatically holds in any lattice. Hence the only thing we need to show is that $A \sqcap (B \vee C) \rhd (A \sqcap B) \vee (A \sqcap C)$.

Let θ be $A \sqcap (B \vee C)$, i.e. $\vdash \theta \leftrightarrow \forall x\, (\Diamond_x \theta \to \Diamond_x A \wedge \Diamond_x (B \vee C))$.

Let σ be $A \sqcap B$, i.e. $\vdash \sigma \leftrightarrow \forall x\, (\Diamond_x \sigma \to \Diamond_x A \wedge \Diamond_x B)$.

Let τ be $A \sqcap C$, i.e. $\vdash \tau \leftrightarrow \forall x\, (\Diamond_x \tau \to \Diamond_x A \wedge \Diamond_x C)$.

Let n_0 be great enough so that all the above fixed point equations are proven in T_{n_0}. We will show that for all $n \geq n_0$, $\theta \vdash_{n+1} \Diamond_n (\sigma \vee \tau)$. It follows that for all n, $\theta \vdash_{\mathsf{PA}} \Diamond_n (\sigma \vee \tau)$, whence by the Orey-Hájek characterization $\theta \rhd \sigma \vee \tau$. Fix $n \geq n_0$ and argue in T_{n+1}. Assume θ. By reflection, $\Diamond_n \theta$, and hence by properties of θ,

$$\Diamond_n A \wedge \Diamond_n (B \vee C). \tag{36}$$

Suppose for contradiction that $\Box_n \neg(\sigma \vee \theta)$. By propositional logic and distributivity of $\Box$ over $\wedge$, we get $\Box_n \neg\sigma$ and $\Box_n \neg\tau$. Since we are reasoning in T_{n+1}, we get $\neg\sigma$ and $\neg\tau$ by reflection. By the properties of σ and τ, we have:

$$\exists x\, (\Diamond_x \sigma \wedge (\Box_x \neg A \vee \Box_x \neg B)), \tag{37}$$

$$\exists x\, (\Diamond_x \tau \wedge (\Box_x \neg A \vee \Box_x \neg C)). \tag{38}$$

Since we had assumed $\Box_n \neg\tau$ and $\Box_n \neg\sigma$, the witnesses for the existential sentences in (37) and (38) have to be smaller than n (using monotonicity). So there are some

[11]The proof below differs slightly from Švejdar's, as we will make use of the properties of our elaborate stratification sequence, and the latter was not used by Švejdar.

$k, l < n$ with $\Diamond_k \sigma \wedge (\Box_k \neg A \vee \Box_k \neg B)$ and $\Diamond_l \tau \wedge (\Box_l \neg A \vee \Box_l \neg C)$. Suppose w.l.o.g. that $k > l$. By monotonicity, it follows that $(\Box_k \neg A \vee \Box_k \neg B)$ and $(\Box_k \neg A \vee \Box_k \neg C)$. By propositional logic and the distributivity of $\Box$ over $\wedge$,

$$\Box_k \neg A \vee \Box_k (\neg B \wedge \neg C). \tag{39}$$

By monotonicity, this implies $\Box_n \neg A \vee \Box_n (\neg B \wedge \neg C)$, contradicting (36). ❑

Open Question 41. Do we have $\vdash A \sqcap (B \vee C) \leftrightarrow (A \sqcap B) \vee (A \sqcap C)$? I.e. is Švejdar's implementation distributive w.r.t. provability? ❑

According to the next theorem, Švejdar's implementation is extensional.

Theorem 42 (Extensionality of $\sqcap$). *For all A and B,*

$$\vdash \Box(A \leftrightarrow A') \wedge \Box(B \leftrightarrow B') \rightarrow \Box(A \sqcap B \leftrightarrow A' \sqcap B').$$

In order to prove Theorem 42, we will first prove a lemma.

Lemma 4.1. *Let $n > 0$ and $\vdash_n \theta \leftrightarrow \forall x\, (\Diamond_x \theta \rightarrow \Diamond_x A \wedge \Diamond_x B)$. Then*

$$\vdash_{n+1} \theta \leftrightarrow ((\Diamond_n A \wedge \Diamond_n B) \wedge \forall x > n\, (\Diamond_x \theta \rightarrow \Diamond_x A \wedge \Diamond_x B)).$$

Proof. Reason in T_{n+1}. Assume θ. Then $\forall x\, (\Diamond_x \theta \rightarrow \Diamond_x A \wedge \Diamond_x B)$, and thus in particular $\forall x > n\, (\Diamond_x \theta \rightarrow \Diamond_x A \wedge \Diamond_x B)$. We also have $\theta \rightarrow \Diamond_n \theta$ (by reflection). Thus by the properties of θ, $\Diamond_n A$ and $\Diamond_n B$. For the other direction, note that by monotonicity, $\Diamond_n A \wedge \Diamond_n B \rightarrow \forall x \leq n\, (\Diamond_x A \wedge \Diamond_x B)$. The result follows by combining this with the other conjunct $\forall x > n\, (\Diamond_x \theta \rightarrow \Diamond_x A \wedge \Diamond_x B)$. ❑

We will now prove Theorem 42. As before, we give an informal proof which is easily seen to be verifiable in PA. Unlike Theorems 38 and 40, the result stated by Theorem 42 is new[12], and its proof makes essential use of our elaborate stratification sequence.

Proof. Let n be s.t.

i. $\vdash_n A \leftrightarrow A'$, $\vdash_n B \leftrightarrow B'$

ii. $\vdash_n \theta \leftrightarrow \forall x\, (\Diamond_x \theta \rightarrow (\Diamond_x A \wedge \Diamond_x B))$, $\vdash_n \sigma \leftrightarrow \forall x\, (\Diamond_x \sigma \rightarrow (\Diamond_x A' \wedge \Diamond_x B'))$

We will show that $\vdash_{n+1} \theta \leftrightarrow \sigma$.

Note first that since $\vdash_n A \leftrightarrow A'$, by necessitation (here we use that GL is the provability logic of T_n) $\vdash_n \Box_n (A \leftrightarrow A')$, by monotonicity $\vdash_{n+1} \forall x \geq n\, \Box_x (A \leftrightarrow A')$, and thus

$$\vdash_{n+1} \forall x \geq n\, (\Diamond_x A \leftrightarrow \Diamond_x A'). \tag{40}$$

Similarly, we get

$$\vdash_{n+1} \forall x \geq n\, (\Diamond_x B \leftrightarrow \Diamond_x B'). \tag{41}$$

[12]Keep in mind that properties like extensionality were not investigated by Švejdar, as he was only interested in the properties of the supremum w.r.t. interpretability. Our goal of extending the logic ILM with a supremum operator obliges us to be acquainted also with the *provability* properties of a supremum.

We will now show that

$$\vdash_{n+1} \Box_{n+1}(\theta \leftrightarrow \sigma) \rightarrow (\theta \leftrightarrow \sigma).$$

Then $\vdash_{n+1} \theta \leftrightarrow \sigma$ follows by Löb's Theorem for T_{n+1}. Reason in T_{n+1}. Assume $\Box_{n+1}(\theta \leftrightarrow \sigma)$ and θ. By Lemma 4.1, it suffices to show

$$(\Diamond_n A' \wedge \Diamond_n B') \wedge \forall x > n\, (\Diamond_x \sigma \rightarrow \Diamond_x A' \wedge \Diamond_x B').$$

Also by Lemma 4.1, we get $(\Diamond_n A \wedge \Diamond_n B) \wedge \forall x > n\, (\Diamond_x \theta \rightarrow \Diamond_x A \wedge \Diamond_x B)$ from θ. From the first conjunct, we get $(\Diamond_n A' \wedge \Diamond_n B')$ using (40) and (41). So let $x > n$ and $\Diamond_x \sigma$. By the assumption $\Box_{n+1}(\theta \leftrightarrow \sigma)$, we get $\Diamond_x \theta$, whence $\Diamond_x A \wedge \Diamond_x B$. Since $x \geq n$, $\Diamond_x A' \wedge \Diamond_x B'$ follows by (40) and (41). The other direction is similar. ❑

The next theorem is an immediate consequence of the proof of Theorem 42 — only the fixed point properties of $A \sqcap B$ and $A' \sqcap B'$, not their construction by the Fixed Point Theorem.

THEOREM 43 (Uniqueness of $\sqcap$). *Suppose that* $\vdash \theta \leftrightarrow \forall x\, (\Diamond_x \theta \rightarrow \Diamond_x A \wedge \Diamond_x B)$ *and* $\vdash \sigma \leftrightarrow \forall x\, (\Diamond_x \sigma \rightarrow \Diamond_x A \wedge \Diamond_x B)$. *Then* $\vdash \theta \leftrightarrow \sigma$.

According to Theorem 43, the solution to the fixed point equation

$$\forall x\, (\Diamond_x Y \rightarrow \Diamond_x A \wedge \Diamond_x B)$$

is unique up to provable equivalence.

OPEN QUESTION 44. Is Švejdar's implementation monotone? ❑

We will now prove a theorem concerning the suprema of $\Box$-formulas under Švejdar's implementation. As we will see, all such suprema are provably equivalent to $\top$. In a way, this is not surprising. Note that according to Lemma 2.1 of Chapter 2, we have that $\top \rhd \Box\bot$. Since for all A, $\Box\bot \rightarrow \Box A$, $\Box\bot \rhd \Box A$, and hence by J2 also $\top \rhd \Box A$. Of course, having that $\top \rightarrow \Box A$ is still stronger than the latter.

THEOREM 45. *For all* A *and* B, $\vdash \Box A \sqcap \Box B$.

To prove Theorem 45, we will first prove a lemma.

LEMMA 4.2. $\vdash \forall x\, (\Box_x \neg\Box C \leftrightarrow \Box_x \bot)$.

PROOF. The direction from right to left is clear. For the other direction:

$$\begin{aligned}
&\vdash \Box_x \neg\Box\, C \\
&\rightarrow \Box_x \neg\Box_x C && \text{(monotonicity)} \\
&\rightarrow \Box_x (\Box_x C \rightarrow \bot) \\
&\rightarrow \Box_x (\Box_x \bot \rightarrow \bot) && \text{(since } \Box_x \bot \rightarrow \Box_x C\text{)} \\
&\rightarrow \Box_x \bot && \text{(Löb's Theorem)}
\end{aligned}$$

❑

We will now give a proof of Theorem 45.

PROOF. By contraposition, $\Box A \sqcap \Box B$ is the sentence θ s.t.

$$\vdash \theta \leftrightarrow \forall x\,((\Box_x \neg \Box A \vee \Box_x \neg \Box B) \rightarrow \Box_x \neg \theta)$$

By Lemma 4.2 and propositional logic,

$$\vdash (\Box_x \neg \Box A \vee \Box_x \neg \Box B) \leftrightarrow \Box_x \bot.$$

It follows that

$$\vdash \theta \leftrightarrow \forall x\,(\Box_x \bot \rightarrow \Box_x \neg \theta),$$

Since $\vdash \forall x\,(\Box_x \bot \rightarrow \Box_x \neg \theta)$, it follows that $\vdash \theta \leftrightarrow \top$, ❑

We will now list some facts concerning Švejdar's implementation.

FACT 46. $\vdash \top \sqcap \top \leftrightarrow \top$. *This follows by Theorem 43 and the observation that* $\top$ *is a solution to the equation* $\forall x\,(\Diamond_x Y \rightarrow \Diamond_x \top)$.

The following fact establishes a weak versions of monotonicity for $\sqcap$.

FACT 47. *If* $\vdash_0 A \rightarrow C$, *then* $\vdash A \sqcap A \leftrightarrow A \sqcap C$.

PROOF. Since $\vdash_0 A \rightarrow C$, also $\vdash_0 \Box_0(A \rightarrow C)$ by necessitation for T_0.

Let θ be $A \sqcap A$, i.e. $\vdash \theta \leftrightarrow \forall x\,(\Diamond_x \theta \rightarrow \Diamond_x A)$.

Let σ be $A \sqcap C$, i.e. $\vdash \sigma \leftrightarrow \forall x\,(\Diamond_x \sigma \rightarrow \Diamond_x A \wedge \Diamond_x C)$.

Let n be great enough so that the above fixed point equations are proven in T_n. We will show that $\vdash_n \Box_n(\theta \leftrightarrow \sigma) \rightarrow (\theta \leftrightarrow \sigma)$. Then $\vdash_n \theta \leftrightarrow \sigma$ follows by applying Löb's Theorem for T_n. Argue in T_n. Assume $\Box_n(\theta \leftrightarrow \sigma) \wedge \theta$. We want to show that σ. So assume that $\Diamond_x \sigma$. If $x < n$, then the assumption θ implies $\Diamond_x \theta$, using reflection. If $x \geq n$, then we get $\Diamond_x \theta$ by the assumption $\Box_n(\theta \leftrightarrow \sigma)$ and monotonicity. Thus in both cases, $\Diamond_x \theta$. Using the assumption θ and the fixed point version of θ, we get $\Diamond_x A$. We want to show that also $\Diamond_x C$. But if $\Box_x \neg C$, then also $\Box_x \neg A$ by monotonicity and the assumption that $\Box_0(A \rightarrow C)$. Hence we have shown that $\Diamond_x A \wedge \Diamond_x C$, and we can conclude σ. The other direction is similar but easier, since we do not need to use the assumption $\vdash_0 A \rightarrow C$. ❑

OPEN QUESTION 48. Does Švejdar's implementation have an explicit form? ❑

A possible strategy for finding an explicit form for Švejdar's implementation is to ignore the first order structure of the equation $\forall x\,(\Diamond_x Y \rightarrow \Diamond_x A \wedge \Diamond_x B)$, and apply the algorithm for calculating explicit fixed points in GL. Ignoring the universal quantifier, $A \sqcap B$ has the form: $p \leftrightarrow (\Diamond p \rightarrow q)$. Applying the algorithm yields $p \leftrightarrow (\Diamond \top \rightarrow q)$. So a conjecture for the explicit from of $A \sqcap B$ would be the sentence $\forall x\,(\Diamond_x \top \rightarrow \Diamond_x A \wedge \Diamond_x B)$. But we haven't been able to prove or refute that this sentence is (verifiably) in the same degree as $A \sqcap B$.

4.3. A puzzling result. This subsection discusses a property of fixed points of the form $\forall x\,(\Diamond_x Y \to \Diamond_x A)$, which stands in stark contrast to the nice behaviour of fixed points expressible in the language of GL.

Note that in Fact 46, $\top$, which is equivalent to a fixed point of the equation $\forall x\,(\Diamond_x Y \to \Diamond_x \top)$, is itself a a fixed point of this equation. However this is not the case in general, i.e. it is possible that i, ii and iii below all hold at the same time.

i. $\vdash \theta \leftrightarrow \forall x\,(\Diamond_x \theta \to \Diamond_x A)$

ii. $\vdash \sigma \leftrightarrow \theta$

iii. $\nvdash \sigma \leftrightarrow \forall x\,(\Diamond_x \sigma \to \Diamond_x A)$

I.o.w. it is possible that although σ is equivalent to a fixed point of the equation $\forall x\,(\Diamond_x Y \to \Diamond_x A)$, it is not itself a fixed point of that equation. This is illustrated by the following example.

Example 49. Note first that if $\vdash \theta \leftrightarrow \forall x\,(\Diamond_x \theta \to \Diamond_x \bot)$, then $\vdash \bot \leftrightarrow \theta$.

To see this, note that $\vdash \Box_0 \top$, whence by monotonicity $\vdash \forall x\,\Box_x \top$, i.e. $\vdash \forall x \neg \Diamond_x \bot$. It follows by propositional logic that if $\vdash \theta \leftrightarrow \forall x\,(\Diamond_x \theta \to \Diamond_x \bot)$, then $\vdash \theta \leftrightarrow \forall x\,\neg \Diamond_x \theta$, i.e. $\vdash \theta \leftrightarrow \forall x\,\Box_x \neg\theta$, and $\vdash \theta \leftrightarrow \Box_0 \neg\theta$ by monotonicity. Since $\vdash \Box_0 \neg\theta \to \neg\theta$ by reflection, it follows that $\vdash \theta \leftrightarrow \bot$.

But clearly, $\nvdash \bot \leftrightarrow \forall x\,(\Diamond_x \bot \to \Diamond_x \bot)$ (because the right hand side is equivalent to $\top$). Hence while $\bot$ is provably equivalent to any fixed point of the equation $\forall x\,(\Diamond_x Y \to \Diamond_x \bot)$, it can itself never be a fixed point of this equation. ❑

Thus the fixed point equation $\forall x\,(\Diamond_x Y \to \Diamond_x A)$ differs from the ones that can be expressed in the language of GL. In GL, it is a consequence of the Fixed Point Theorem (see Section 1.3 of Chapter 2) that any sentence equivalent to a fixed point of an equation is itself a fixed point of this equation (see also [**Boo93**]).

Let $\vdash_n \theta \leftrightarrow \forall x\,(\Diamond_x \theta \to \Diamond_x A \wedge \Diamond_x B)$ and $\vdash_n \sigma \leftrightarrow \theta$. While — as shown above — we do not in general have that $\vdash \sigma \leftrightarrow \forall x\,(\Diamond_x \sigma \to \Diamond_x A \wedge \Diamond_x B)$, we can prove that σ is a fixed point of another formula, namely the version of θ given in Lemma 4.1. According to Lemma 4.1, we have that

$$\vdash_{n+1} \theta \leftrightarrow (\Diamond_n A \wedge \Diamond_n B) \wedge \forall x > n\,(\Diamond_x \theta \to \Diamond_x A \wedge \Diamond_x B). \tag{42}$$

What we can show is that σ is a fixed point of the equation in (42), i.e. that

$$\vdash_{n+1} \sigma \leftrightarrow (\Diamond_n A \wedge \Diamond_n B) \wedge \forall x > n\,(\Diamond_x \sigma \to \Diamond_x A \wedge \Diamond_x B). \tag{43}$$

Proof. Since $\vdash_n \sigma \leftrightarrow \theta$, we have that $\vdash_n \Box_n(\sigma \leftrightarrow \theta)$ by necessitation for T_n, and thus $\vdash_n \forall x \geq n\,\Box_x(\sigma \leftrightarrow \theta)$ by monotonicity, whence also

$$\vdash_{n+1} \forall x \geq n\,(\Diamond_x \sigma \leftrightarrow \Diamond_x \theta). \tag{44}$$

Using (44), we see that

$$\vdash_{n+1} \forall x \geq n\,[(\Diamond_x \theta \to \Diamond_x A \wedge \Diamond_x B) \leftrightarrow (\Diamond_x \sigma \to \Diamond_x A \wedge \Diamond_x B)] \tag{45}$$

It is now easy to see that (43) holds:

$$\begin{aligned}
\vdash_{n+1} \sigma \leftrightarrow \theta & \qquad \text{(assumption)} \\
\leftrightarrow (\Diamond_n A \wedge \Diamond_n B) \wedge \forall x > n\, (\Diamond_x \theta \to \Diamond_x A \wedge \Diamond_x B) & \qquad (42) \\
\leftrightarrow (\Diamond_n A \wedge \Diamond_n B) \wedge \forall x > n\, (\Diamond_x \sigma \to \Diamond_x A \wedge \Diamond_x B) & \qquad (45)
\end{aligned}$$

❑

Using (43), it is clear that $\vdash \sigma \to \forall x\, (\Diamond_x \sigma \to \Diamond_x A \wedge \Diamond_x B)$, however the other direction does not hold in general.

5. Visser's Implementation of the Supremum

This section explores an alternative way of implementing the supremum in PA, discovered by Albert Visser. The discovery was inspired by work on a joint paper ([**SA12**]) with Volodya Shavrukov; some of the crucial ingredients are already present [**Sha94**]. As we will see, Visser's implementation is to a certain extent dual to Švejdar's implementation.

Given sentences A and B, *Visser's supremum* is found by using the Gödel Fixed Point Theorem to obtain a sentence θ with the property that

$$\vdash \theta \leftrightarrow \forall x\, (\Box_x \theta \to \Box_x \neg A \vee \Box_x \neg B). \tag{46}$$

If θ is as above, the sentence $\neg\theta$ is verifiably in PA in the degree of the supremum of A and B. This will be proven below. Note that whereas $A \sqcap B$ was universal sentence, under Visser's implementation the supremum of A and B is an existential sentence.

5.1. Visser's supremum as an implementation. Just like Švejdar's supremum, Visser's supremum can be turned into an implementation by using the general version of the Gödel-Carnap Fixed Point Theorem. The procedure for doing that is exactly like the one presented in Section 4.1. Hence we can assume that we have a formula $\theta(y, z)$ s.t. for all A and B,

$$\vdash \theta(A, B) \leftrightarrow \forall x\, (\Box_x \theta(A, B) \to \Box_x \neg A \vee \Box_x \neg B).$$

We will show that for all A and B, the sentence $\neg\theta(A, B)$ is verifiably in PA in the degree of the supremum of A and B, and hence $\neg\theta(y, z)$ is an implementation of the supremum in PA, in the sense of Definition 35. We will write $(A \curlywedge B)$ for the sentence $\neg\theta(A, B)$, where $\theta(A, B)$ is as above. We shall omit parentheses according to the convention that $\curlywedge$ binds equally strong as $\vee$ and $\wedge$. To prove that $A \curlywedge B$ is a supremum of A and B, we will use Lemma 5.1 and Theorem 50 below.

LEMMA 5.1. *Let* $\vdash_n \theta \leftrightarrow \forall x\, (\Box_x \theta \to \Box_x \neg A \vee \Box_x \neg B)$. *Then*

$$\vdash_n \forall x \geq n\, \Box_x (\forall y < x\, (\Box_y \theta \to \Box_y \neg A \vee \Box_y \neg B)).$$

PROOF. By necessitation,

$$\vdash_n \Box_n (\theta \leftrightarrow \forall x\, (\Box_x \theta \to \Box_x \neg A \vee \Box_x \neg B)), \tag{47}$$

whence we can use θ and $\forall x\,(\Box_x\theta \to \Box_x\neg A \vee \Box_x\neg B)$ interchangeably in $\Box_x$ if $x \geq n$.

$$\begin{aligned}
\vdash_n x \geq n &\to \Box_x\forall y < x\,(\Box_y\theta \to \theta) && \text{(reflection)}\\
&\to \Box_x\forall y < x\,(\Box_y\theta \to \forall z(\Box_z\theta \to \Box_z\neg A \vee \Box_z\neg B)) && \text{(fixed point version of } \theta)\\
&\to \Box_x\forall y < x\,(\Box_y\theta \to \Box_y\neg A \vee \Box_y\neg B) && \qquad\qquad \text{❑}
\end{aligned}$$

THEOREM 50. *Let* $\vdash_n \theta \leftrightarrow \forall x\,(\Box_x\theta \to \Box_x\neg A \vee \Box_x\neg B)$. *Then*

$$\vdash_n \forall x \geq n\,(\Box_x\theta \leftrightarrow \Box_x(\Box_x\neg A \vee \Box_x\neg B)).$$

PROOF.

$$\begin{aligned}
\vdash_n x \geq n &\to (\Box_x\theta \to \Box_x\forall y\,(\Box_y\theta \to \Box_y\neg A \vee \Box_y\neg B)) && \text{(fixed point version of } \theta)\\
&\to \Box_x(\Box_x\theta \to \Box_x\neg A \vee \Box_x\neg B) && \text{(instantiating } \forall)\\
&\to (\Box_x\Box_x\theta \to \Box_x(\Box_x\neg A \vee \Box_x\neg B)) && (\mathsf{K}\text{-axiom})\\
&\to \Box_x(\Box_x\neg A \vee \Box_x\neg B) && (\text{since } \Box_x\theta \to \Box_x\Box_x\theta)\\
&\to \Box_x\forall y \geq x\,(\Box_y\neg A \vee \Box_y\neg B) && \text{(monotonicity)}\\
&\to \Box_x\forall y \geq x\,(\Box_y\theta \to \Box_y\neg A \vee \Box_y\neg B)\\
&\to \Box_x\forall y\,(\Box_y\theta \to \Box_y\neg A \vee \Box_y\neg B) && \text{(Lemma 5.1)}\\
&\to \Box_x\theta && \text{(fixed point version of } \theta)
\end{aligned}$$

We have shown that $\vdash_n \forall x \geq n\,(\Box_x\theta \leftrightarrow \Box_x(\Box_x\neg A \vee \Box_x\neg B))$. ❑

THEOREM 51. *For all* C, $\vdash (C \rhd A) \wedge (C \rhd B) \leftrightarrow C \rhd A \curlywedge B$.

PROOF. We will give an informal proof. Using the properties of our elaborate stratification sequence and the verifiability of the Orey-Hájek characterization in PA, it is straightforward to see that the proof can be verified in PA. Write θ for $\neg(A \curlywedge B)$. Then θ has the property that

$$\vdash \theta \leftrightarrow \forall x\,(\Box_x\theta \to \Box_x\neg A \vee \Box_x\neg B) \tag{48}$$

Let n be great enough so that (48) is proven in T_n. By Theorem 50, we have

$$\vdash_n \forall x \geq n\,(\Box_x\theta \leftrightarrow \Box_x(\Box_x\neg A \vee \Box_x\neg B)),$$

whence by monotonicity also

$$\vdash_{n+1} \forall x \geq n\,(\Box_x\theta \leftrightarrow \Box_x(\Box_x\neg A \vee \Box_x\neg B)).$$

In particular,

$$\vdash_{n+1} \Box_n\theta \leftrightarrow \Box_n(\Box_n\neg A \vee \Box_n\neg B).$$

We note that

$$\vdash_{n+1} \Box_n(\Box_n\neg A \vee \Box_n\neg B) \leftrightarrow (\Box_n\neg A \vee \Box_n\neg B).$$

The direction from right to left uses provable Σ_1-completeness; the other direction follows by reflection. Thus we have $\vdash_{n+1} \Box_n\theta \leftrightarrow (\Box_n\neg A \vee \Box_n\neg B)$, whence by contraposition (remember that θ was the sentence $\neg(A \curlywedge B)$)

$$\vdash_{n+1} \Diamond_n(A \curlywedge B) \leftrightarrow (\Diamond_n A \wedge \Diamond_n B). \tag{49}$$

Using (49) and the Orey-Hájek characterization, it is easy to see that

$$(C \rhd A) \wedge (C \rhd B) \Leftrightarrow C \rhd A \curlywedge B. \qquad \text{❑}$$

5.2. Properties of Visser's implementation. Using Theorem 50, we see that a fixed point of the equation $\forall x\,(\Box_x Y \to \Box_x\neg A \vee \Box_x\neg B)$ is provably equivalent to a "stratified" reflection principle for $(\Box_x\neg A \vee \Box_x\neg B)$.

THEOREM 52 (Explicit version of $\curlywedge$). *Let* $\vdash_n \theta \leftrightarrow \forall x\,(\Box_x\theta \to \Box_x\neg A \vee \Box_x\neg B)$. *Then* $\vdash_n \theta \leftrightarrow \forall x\,(\Box_x(\Box_x\neg A \vee \Box_x\neg B) \to \Box_x\neg A \vee \Box_x\neg B)$.

PROOF. Assume $\vdash_n \theta \leftrightarrow \forall x\,(\Box_x\theta \to \Box_x\neg A \vee \Box_x\neg B)$. By Theorem 50,

$$\vdash_n \theta \leftrightarrow \forall x \geq n\,(\Box_x(\Box_x\neg A \vee \Box_x\neg B) \to \Box_x\neg A \vee \Box_x\neg B). \tag{50}$$

On the other hand, we have by reflection

$$\vdash_n \theta \leftrightarrow \forall x < n\,(\Box_x(\Box_x\neg A \vee \Box_x\neg B) \to \Box_x\neg A \vee \Box_x\neg B). \tag{51}$$

Combining (50) and (51), we see that

$$\vdash_n \theta \leftrightarrow \forall x\,(\Box_x(\Box_x\neg A \vee \Box_x\neg B) \to \Box_x\neg A \vee \Box_x\neg B). \qquad \text{❑}$$

According to Theorem 53, the explicit version provided by Theorem 52 satisfies an analogue of Theorem 50. According to Corollary 54, it is also a fixed point of the equation $\forall x\,(\Box_x Y \to \Box_x\neg A \vee \Box_x\neg B)$.

THEOREM 53. *Let* $\sigma := \forall x\,(\Box_x(\Box_x\neg A \vee \Box_x\neg B) \to \Box_x\neg A \vee \Box_x\neg B)$. *Then* $\vdash_1 \forall x\,(\Box_x\sigma \leftrightarrow \Box_x(\Box_x\neg A \vee \Box_x\neg B))$.

PROOF.

$$\vdash_1 \Box_x\sigma \to \Box_x\forall y\,(\Box_y(\Box_y\neg A \vee \Box_y\neg B) \to \Box_y\neg A \vee \Box_y\neg B) \tag{52}$$
$$\to \Box_x(\Box_x(\Box_x\neg A \vee \Box_x\neg B) \to \Box_x\neg A \vee \Box_x\neg B) \tag{53}$$
$$\to \Box_x(\Box_x\neg A \vee \Box_x\neg B) \tag{54}$$
$$\to \Box_x\forall y \geq x\,(\Box_y\neg A \vee \Box_y\neg B) \tag{55}$$
$$\to \Box_x\forall y \geq x\,(\Box_y(\Box_y\neg A \vee \Box_y\neg B) \to \Box_y\neg A \vee \Box_y\neg B) \wedge \tag{56}$$
$$\Box_x\forall y < x\,(\Box_y(\Box_y\neg A \vee \Box_y\neg B) \to \Box_y\neg A \vee \Box_y\neg B) \tag{57}$$
$$\to \Box_x\forall y\,(\Box_y(\Box_y\neg A \vee \Box_y\neg B) \to \Box_y\neg A \vee \Box_y\neg B) \tag{58}$$
$$\to \Box_x\sigma \tag{59}$$

Step (54) uses formalized Löb's Theorem, step (55) monotonicity, step (56) the previous step together with propositional logic, and step (57) reflection. ❑

COROLLARY 54. *Let* $\sigma := \forall x\,(\Box_x(\Box_x\neg A \vee \Box_x B) \to \Box_x\neg A \vee \Box_x\neg B)$. *Then* $\vdash \sigma \leftrightarrow \forall x\,(\Box_x\sigma \to \Box_x\neg A \vee \Box_x\neg B)$.

PROOF. Immediate from Theorem 53. ❑

Using Corollary 54 and the proof[13] of Theorem 51, we see that we have a direct proof of the fact that the sentence

$$\exists x\,(\Box_x(\Box_x A \vee \Box_x B) \wedge (\Diamond_x A \wedge \Diamond_x B)) \tag{60}$$

[13]To prove that a sentence τ is a supremum of A and B, we only used that the negation of τ is a solution to the fixed point equation $\forall x\,(\Box_x Y \to \Box_x\neg A \vee \Box_x\neg B)$.

is a supremum of A and B. From now on, if we write $A \curlywedge B$, we mean either the fixed point version of $A \curlywedge B$, i.e. the sentence τ with

$$\vdash \tau \leftrightarrow \exists x\,(\Box_x \neg\tau \wedge (\Diamond_x A \wedge \Diamond_x B)) \tag{61}$$

or the explicit form, i.e. the sentence in (60). Note that by Corollary 54, the sentence in (60) is also a sentence τ as in (61).

The appearance of the explicit form of Visser's implementation is in some sense not surprising. Ignoring the first order structure of the fixed point version of $\neg(A \curlywedge B)$ we see that it has the form $p \leftrightarrow (\Box p \to q)$. Applying the algorithm for calculating explicit fixed points in GL, we get that $p \leftrightarrow (\Box q \to q)$, in agreement with the propositional structure of the actual explicit form provided by Theorem 52.

Hence the equation $\forall x\,(\Box_x Y \to \Box_x \neg A \vee \Box_x \neg B)$ is "well-behaved" in the sense that it has an explicit fixed point in accordance to the calculations done in GL. One could expect that also other properties of the tame fixed point equations expressible in the language of GL can be transferred. In particular, one might wonder whether the puzzling result of Section 4.3 has an analogue for Visser's implementation. I.e. do we have that if $\vdash \theta \leftrightarrow \forall x\,(\Box_x \theta \to \Box_x \neg A \vee \Box_x \neg B)$ and $\vdash \sigma \leftrightarrow \theta$, then also $\vdash \sigma \leftrightarrow \forall x\,(\Box_x \sigma \to \Box_x \neg A \vee \Box_x \neg B)$? According to Corollary 54, this is indeed the case for $\sigma = \forall x\,(\Box_x(\Box_x \neg A \vee \Box_x B) \to \Box_x \neg A \vee \Box_x \neg B)$. However, the proof of Theorem 53 (which is used to establish Corollary 54) uses the special properties of σ, i.e. that it has the form of a particular reflection principle. For the general case, we can only show something weaker[14].

THEOREM 55. *Let* $\vdash_n \theta \leftrightarrow \forall x\,(\Box_x \theta \to \Box_x \neg A)$ *and* $\vdash_n \sigma \leftrightarrow \theta$. *Then*

$$\vdash_n \sigma \leftrightarrow \forall x \geq n\,(\Box_x \sigma \to \Box_x \neg A).$$

PROOF. Since $\vdash_n \sigma \leftrightarrow \theta$, also $\vdash_n \Box_n(\sigma \leftrightarrow \theta)$ by necessitation for T_n. We reason in T_n. Assume σ. Then also θ by assumption. Let $x \geq n$ and $\Box_x \sigma$. Then also $\Box_x \theta$, and by the assumption θ it follows that $\Box_x \neg A$. For the other direction, assume that $\forall x \geq n\,(\Box_x \sigma \to \Box_x \neg A)$. Since we have $\vdash_n \sigma \leftrightarrow \theta$, in order to show that σ it suffices to show that θ, i.e. $\forall x\,(\Box_x \theta \to \Box_x \neg A)$. So assume $\Box_x \theta$. If $x < n$ we get θ by reflection, and hence $\Box_x \neg A$ by properties of θ. If $x \geq n$ we get $\Box_x \sigma$, whence $\Box_x \neg A$ follows from the assumption that $\forall x \geq n\,(\Box_x \sigma \to \Box_x \neg A)$. ❑

We will now prove that Visser's implementation is extensional.

THEOREM 56 (Extensionality of $\curlywedge$). *For all* A *and* B,

$$\vdash \Box(A \leftrightarrow A') \wedge \Box(B \leftrightarrow B') \to \Box(A \curlywedge B \leftrightarrow A' \curlywedge B').$$

PROOF. We will give an informal proof. Using the properties of our elaborate stratification sequence and the verifiability of the Orey-Hájek characterization, formalizing the proof is a straightforward procedure. Let $\vdash_n A \leftrightarrow A'$ and $\vdash_n B \leftrightarrow B'$. We will show that $\vdash_n A \curlywedge B \leftrightarrow A' \curlywedge B'$ by using the explicit version of Visser's implementation. So let

[14] In GL, we have that $\vdash_{\mathsf{GL}} \Box(p \leftrightarrow B(p)) \leftrightarrow \Box(p \leftrightarrow H)$, where H is the explicit fixed point provided by the Fixed Point Theorem for GL. Viewing $B(Y)$ as the equation $\forall x\,(\Box_x Y \to \Box_x \neg A)$, we do have the direction from left to right (since we have an explicit fixed point). It is the direction from right to left which we cannot prove in general.

i. $\theta := \forall x\,(\Box_x(\Box_x\neg A \vee \Box_x\neg B) \to \Box_x\neg A \vee \Box_x\neg B)$

ii. $\sigma := \forall x\,(\Box_x(\Box_x\neg A' \vee \Box_x\neg B') \to \Box_x\neg A' \vee \Box_x\neg B')$,

We will show that $\vdash_n \theta \leftrightarrow \sigma$, whence the desired result follows by contraposition.

Since $\vdash_n A \leftrightarrow A'$, we have $\vdash_n \Box_n(A \leftrightarrow A')$ and $\vdash_n \Box_n\Box_n(A \leftrightarrow A')$ by necessitation. Similarly for B and B'. By monotonicity, we get:

$$\vdash_n \forall x \geq n\, \Box_x(A \leftrightarrow A') \text{ and } \vdash_n \forall x \geq n\, \Box_x(B \leftrightarrow B') \tag{62}$$

$$\vdash_n \forall x \geq n\, \Box_x\Box_x(A \leftrightarrow A') \text{ and } \vdash_n \forall x \geq n\, \Box_x\Box_x(B \leftrightarrow A') \tag{63}$$

Reason in T_n. Assume θ and $\Box_x(\Box_x\neg A' \vee \Box_x\neg B')$. If $x < n$, then by reflection, we have that $(\Box_x\neg A' \vee \Box_x\neg B')$. If $x \geq n$, then using (63), $\Box_x(\Box_x\neg A' \vee \Box_x\neg B')$ implies $\Box_x(\Box_x\neg A \vee \Box_x\neg B)$. Since we assumed θ, this implies $(\Box_x\neg A \vee \Box_x\neg B)$, and thus $(\Box_x\neg A' \vee \Box_x\neg B')$ by (62). We can conclude $\theta \to \sigma$. The other direction is similar. ❑

5.3. Properties of $\top \curlywedge \top$. This section deals with the properties of the sentence $\top \curlywedge \top$. We will first show that $\top \curlywedge \top$ is an Orey-sentence.

A sentence O is said to be an Orey-sentence (for PA) if both $\top \rhd O$ and $\top \rhd \neg O$. Thus if O is an Orey-sentence, then neither O nor its negation add interpretability strength to PA. It is clear that if O is an Orey-sentence, then so is $\neg O$. Note also that an Orey-sentence must be independent from PA.

The first Orey-sentences for PA were obtained by Orey ([**Ore61**]). Švejdar ([**Šve78**]) constructs an Orey-sentence for PA by using the Gödel Fixed Point Theorem to find a sentence θ s.t.

$$\vdash \theta \leftrightarrow \forall x\,(\Diamond_x\theta \to \Diamond_x\neg\theta). \tag{64}$$

Using the Orey-Hájek characterization and the properties of our elaborate stratification sequence[15], it is easy to see that $\theta \rhd \neg\theta$ and $\neg\theta \rhd \theta$. Reasoning in IL, it follows that also $\top \rhd \theta$ and $\top \rhd \neg\theta$.

We will now show that $\top \curlywedge \top$ is an Orey-sentence. Consider the fixed-point version of $\neg(\top \curlywedge \top)$, i.e. the sentence σ with $\vdash \sigma \leftrightarrow \forall x\,(\Box_x\sigma \to \Box_x\bot)$. Note that

$$\vdash \forall x\,(\Box_x\sigma \to \Box_x\bot) \leftrightarrow \forall x\,(\Box_x\sigma \to \Box_x\neg\sigma)$$

whence by contraposition,

$$\vdash \forall x\,(\Box_x\sigma \to \Box_x\bot) \leftrightarrow \forall x\,(\Diamond_x\sigma \to \Diamond_x\neg\sigma).$$

The sentence on the right-hand side is Švejdar's Orey-sentence. Hence also $\neg(\top \curlywedge \top)$ and $\top \curlywedge \top$ are Orey-sentences.

In [**SA12**], it is shown that $\top \curlywedge \top$ is an Orey-sentence by showing that it is a Gödel-sentence for the Feferman provability predicate[16], and that any Gödel-sentence for the Feferman provability predicate is an Orey-sentence.

---

[15] As before, Švejdar uses the stratification sequence where T_n is the subtheory of PA axiomatized by axioms of gödelnumbers $< n$.

[16] The Feferman provability predicate is informally defined as: $\mathrm{Pr}_F(A)$ if there exists some x s.t. T_x is consistent and $\vdash_x A$.

We will now show that the suprema of all $\Box$-formulas under Visser's implementation are equivalent to $\top \curlywedge \top$, and hence also Orey-sentences.

THEOREM 57. *For all A and B, $\vdash \Box A \curlywedge \Box B \leftrightarrow \top \curlywedge \top$.*

PROOF. We use the explicit form, and consider the sentence $\neg(\Box A \curlywedge \Box B)$, i.e.

$$\forall x\,(\Box_x(\Box_x\neg\Box A \vee \Box_x\neg\Box B) \to \Box_x\neg\Box A \vee \Box_x\neg\Box B). \tag{65}$$

By Lemma 4.2 we have for all C,

$$\vdash \Box_x\neg\Box C \leftrightarrow \Box_x\bot,$$

hence

$$\vdash (\Box_x\neg\Box A \vee \Box_x\neg\Box B) \leftrightarrow \Box_x\bot,$$

and thus also

$$\vdash \Box_x(\Box_x\neg\Box A \vee \Box_x\neg\Box B) \leftrightarrow \Box_x\Box_x\bot.$$

It follows that the sentence in (65) is equivalent to

$$\forall x\,(\Box_x\Box_x\bot \to \Box_x\bot),$$

which is the explicit version of $\neg(\top \curlywedge \top)$. ❑

5.4. Other properties of $\curlywedge$. We will now examine some further properties of Visser's implementation. First, it is clear that Visser's implementation is commutative w.r.t. provability, i.e. that for all A and B, $\vdash A \curlywedge B \leftrightarrow B \curlywedge A$.

As for distributivity w.r.t. interpretability, it would — as in the case of Švejdar's implementation (see Section 4.2) — suffice to show that

$$\vdash A \curlywedge (B \vee C) \rhd (A \curlywedge B) \vee (A \curlywedge C).$$

According to the next theorem, we even have this direction of distributivity w.r.t. provability.

THEOREM 58. $\vdash A \curlywedge (B \vee C) \to (A \curlywedge B) \vee (A \curlywedge C)$.

PROOF. We show this by contraposition, using the explicit form of Visser's implementation. Assume $\neg(A \curlywedge B)$ and $\neg(A \curlywedge C)$, i.e.

$$\forall x\,(\Box_x(\Box_x\neg A \vee \Box_x\neg B) \to (\Box_x\neg A \vee \Box_x\neg B))$$

and

$$\forall x\,(\Box_x(\Box_x\neg A \vee \Box_x\neg C) \to (\Box_x\neg A \vee \Box_x\neg C)).$$

We want to show that

$$\forall x\,(\Box_x(\Box_x\neg A \vee \Box_x(\neg B \wedge \neg C)) \to (\Box_x\neg A \vee \Box_x(\neg B \wedge \neg C))).$$

$\vdash \Box_x(\Box_x\neg A \vee \Box_x(\neg B \wedge \neg C))$
$\to \Box_x(\Box_x\neg A \vee (\Box_x\neg B \wedge \Box_x\neg C))$ (distributivity of $\Box$ over $\wedge$)
$\to \Box_x((\Box_x\neg A \vee \Box_x\neg B) \wedge (\Box_x\neg A \vee \Box_x\neg C))$ (distributivity of $\vee$ over $\wedge$)
$\to \Box_x(\Box_x\neg A \vee \Box_x\neg B) \wedge \Box_x(\Box_x\neg A \vee \Box_x\neg C)$ (distributivity of $\Box$ over $\wedge$)
$\to (\Box_x\neg A \vee \Box_x\neg B) \wedge (\Box_x\neg A \vee \Box_x\neg C)$ (by assumption)
$\to (\Box_x\neg A \vee (\Box_x\neg B \wedge \Box_x\neg C)$ (distributivity of $\vee$ over $\wedge$)
$\to (\Box_x\neg A \vee \Box_x(\neg B \wedge \neg C)$ (distributivity of $\Box$ over $\wedge$) ❑

COROLLARY 59. $\vdash A \curlywedge (B \vee C) \rhd (A \curlywedge B) \vee (A \curlywedge C)$

PROOF. Immediate from Theorem 58 by the principle J1 of IL. ❑

It is natural to ask whether also the other directions of distributivity are *provable* for Visser's implementation, e.g. do we have that

$$\vdash (A \curlywedge B) \vee (A \curlywedge C) \to A \curlywedge (B \vee C)?$$

The only thing we could prove is the following fact, which can be seen as an approximation of the other non-trivial direction of distributivity. Hence, somewhat curiously exactly the trivial directions of distributivity do not seem to hold w.r.t. provability.

FACT 60. $\vdash (A \vee B) \curlywedge (A \vee C) \to (A \curlywedge A) \vee (B \curlywedge C)$.

PROOF. We show this by contraposition, using the explicit form of Visser's implementation. Assume $\neg(A \curlywedge A)$ and $\neg(B \curlywedge C)$, i.e.

$$\forall x\, (\Box_x \Box_x \neg A \to \Box_x \neg A)$$

and

$$\forall x\, (\Box_x (\Box_x \neg B \vee \Box_x \neg C) \to (\Box_x \neg B \vee \Box_x \neg C)).$$

We want to show that

$$\forall x\, (\Box_x (\Box_x (\neg A \wedge \neg B) \vee \Box_x (\neg A \wedge \neg C)) \to (\Box_x (\neg A \wedge \neg B) \vee \Box_x (\neg A \wedge \neg C))).$$

$$\begin{aligned}
&\vdash \Box_x (\Box_x (\neg A \wedge \neg B) \vee \Box_x (\neg A \wedge \neg C)) \\
&\to \Box_x ((\Box_x \neg A \wedge \Box_x \neg B) \vee (\Box_x \neg A \wedge \Box_x \neg C)) && \text{(distributivity of } \Box \text{ over } \wedge) \\
&\to \Box_x (\Box_x \neg A \wedge (\Box_x \neg B \vee \Box_x \neg C)) && \text{(distributivity of } \vee \text{ over } \wedge \\
&\to \Box_x \Box_x \neg A \wedge \Box_x (\Box_x \neg B \vee \Box_x \neg C) && \text{(distributivity of } \Box \text{ over } \wedge) \\
&\to \Box_x \neg A \wedge (\Box_x \neg B \vee \Box_x \neg C) && \text{(by assumption)} \\
&\to (\Box_x \neg A \wedge \Box_x \neg B) \vee (\Box_x \neg A \wedge \Box_x \neg C) && \text{(distributivity of } \vee \text{ over } \wedge) \\
&\to \Box_x (\neg A \wedge \neg B) \vee \Box_x (\neg A \wedge \neg C) && \text{(distributivity of } \Box \text{ over } \wedge)
\end{aligned}$$ ❑

The next theorem allows us to show that Visser's implementation is not monotone. It will also play an important role in our attempt of developing a modal semantics for Visser's implementation (see Chapter 5).

THEOREM 61.

$$\nvdash ((\top \curlywedge \top) \curlywedge (\top \curlywedge \top)) \to (\top \curlywedge \top) \vee \Diamond (\top \curlywedge \top)$$

PROOF. Suppose for contradiction that

$$\vdash ((\top \curlywedge \top) \curlywedge (\top \curlywedge \top)) \to (\top \curlywedge \top) \vee \Diamond (\top \curlywedge \top),$$

whence by contraposition,

$$\vdash \neg(\top \curlywedge \top) \wedge \Box \neg (\top \curlywedge \top) \to \neg((\top \curlywedge \top) \curlywedge (\top \curlywedge \top)). \tag{66}$$

Throughout this proof, we use the explicit form of Visser's implementation. Note the following:

i. Since $\vdash \top \rhd (\top \curlywedge \top)$ and $\vdash (\top \curlywedge \top) \rhd \top$, we have that $\vdash \Diamond\top \leftrightarrow \Diamond(\top \curlywedge \top)$ (using IL), thus $\vdash \Box\neg(\top \curlywedge \top) \leftrightarrow \Box\bot$.

ii. By Theorem 53 we have that $\vdash \forall x\,(\Box_x\neg(\top \curlywedge \top) \leftrightarrow \Box_x\Box_x\bot)$.

iii. $\neg((\top \curlywedge \top) \curlywedge (\top \curlywedge \top))$ is the sentence

$$\forall x\,(\Box_x\Box_x\neg(\top \curlywedge \top) \to \Box_x\neg(\top \curlywedge \top)).$$

By ii this is equivalent to $\forall x\,(\Box_x\Box_x\Box_x\bot \to \Box_x\Box_x\bot)$.

It follows from the above that if (66) holds, then

$$\vdash \forall x\,(\Box_x\Box_x\bot \to \Box_x\bot) \wedge \Box\bot \to \forall x\,(\Box_x\Box_x\Box_x\bot \to \Box_x\Box_x\bot). \tag{67}$$

Suppose that (67) is proven in T_n. Using reflection, we have in T_n that (67) implies

$$\forall x > n\,(\Box_x\Box_x\bot \to \Box_x\bot) \wedge \Box\bot \to ((\Box_n\Box_n\bot \to \Box_n\bot) \to (\Box_n\Box_n\Box_n\bot \to \Box_n\Box_n\bot)).$$

Note that $\Box_n\Box_n\Box_n\bot$ implies $\Box_x\Box_n\Box_n\bot$ for $x > n$ by monotonicity, whence by applying reflection twice, $\Box_x\bot$. Hence the antecedent $\forall x > n\,(\Box_x\Box_x\bot \to \Box_x\bot) \wedge \Box\bot$ follows from $\Box_n\Box_n\Box_n\bot$, so we have

$$\vdash_n \Box_n\Box_n\Box_n\bot \to ((\Box_n\Box_n\bot \to \Box_n\bot) \to (\Box_n\Box_n\Box_n\bot \to \Box_n\Box_n\bot)).$$

By propositional logic,

$$\vdash_n \Box_n\Box_n\Box_n\bot \to (\Box_n\Box_n\bot \vee \Box_n\bot).$$

Since $\vdash_n \Box_n\bot \to \Box_n\Box_n\bot$, it follows by propositional logic that

$$\vdash_n \Box_n\Box_n\Box_n\bot \to \Box_n\Box_n\bot.$$

By Löb's Theorem for T_n, we get

$$\vdash_n \Box_n\Box_n\bot,$$

which is a contradiction. ❑

Corollary 62. *$\curlywedge$ is not monotone, i.e. it is possible that*

i. $\vdash A \to B$*, but*

ii. $\nvdash C \curlywedge A \to C \curlywedge B$

Proof. Suppose for contradiction that $\curlywedge$ is monotone. It follows that for any A, $\vdash A \curlywedge A \to \top \curlywedge \top$. In particular,

$$\vdash (\top \curlywedge \top) \curlywedge (\top \curlywedge \top) \to \top \curlywedge \top,$$

i.e.

$$\vdash \neg(\top \curlywedge \top) \to \neg((\top \curlywedge \top) \curlywedge (\top \curlywedge \top)),$$

contradicting Theorem 61. ❑

According to Fact 63, we have a weak version of monotonicity for Visser's implementation.

Fact 63. *If* $\vdash_0 A \to C$*, then* $\vdash A \curlywedge A \leftrightarrow A \curlywedge C$.

PROOF. We will use the explicit form and show by contraposition that

$$\vdash \forall x\,(\Box_x(\Box_x\neg A \vee \Box_x\neg C) \to (\Box_x\neg A \vee \Box_x\neg C))$$
$$\leftrightarrow \forall x\,(\Box_x\Box_x\neg A \to \Box_x\neg A).$$

The result follows immediately once we note that since $\vdash_0 \Box_0(\neg C \to \neg A)$, we have

$$\vdash (\Box_x\neg A \vee \Box_x\neg C) \leftrightarrow \Box_x\neg A,$$

whence also

$$\vdash \Box_x(\Box_x\neg A \vee \Box_x\neg C) \leftrightarrow \Box_x\Box_x\neg A.$$

❑

6. Table of Properties of ⊓ and ⅄

We finish this chapter by giving a comparative summary of the properties of our two implementations. "Distributivity (≡)" and "Distributivity (↔)" stand for distributivity w.r.t. interpretability and provability respectively. "Weak monotonicity" stands for the property that if $\vdash_0 A \to C$, then $\vdash A \between A \leftrightarrow A \between C$.

TABLE 1. Properties of the two implementations.

	Švejdar's ⊓	Visser's ⅄
Commutativity (↔)	✓	✓
Distributivity (≡)	✓	✓
Distributivity (↔)	**?**	partly
Extensionality	✓	✓
Monotonicity	**?**	✗
Weak Monotonicity	✓	✓
$\top \between \top$	equivalent to $\top$	Orey-sentence
$\Box A \between \Box B$	equivalent to $\top \sqcap \top$	equivalent to $\top$ ⅄ $\top$
Explicit form	**?**	✓

CHAPTER 4

Semantics for ILMS

Let ILMS be the system ILM plus the axiom S: $(C \rhd A) \wedge (C \rhd B) \leftrightarrow C \rhd A \owedge B$. Since ILMS is an extension of ILM (as a theory), it is natural to ask whether it can be seen as such also from the semantic point of view. In other words, can we extend the semantics for ILM to a semantics for ILMS? Answering this question is the main goal of this chapter. Section 1 introduces the system ILMS. In Section 2, we will try to find a structural characterization of ILM–models validating axiom S. The possibility of such a characterization is excluded in Section 4, thus defeating all hope of finding a relational semantics for ILMS. Section 3 explores the possibility of having a weaker notion of semantics for ILMS. In Section 5, we obtain modal completeness of ILMS w.r.t. this notion of semantics, in the simple case where the underlying frames have depth 2. In Section 6 we obtain an arithmetical completeness result for a very restricted set of formulas.

1. The Logic ILMS

This section introduces the system ILMS – the minimal logic for the supremum operator. The formulas of ILMS are defined as follows:

$$\mathrm{F}_{\mathsf{ILMS}} ::= \bot \mid \mathrm{Prop} \mid (\mathrm{F}_{\mathsf{IL}} \to \mathrm{F}_{\mathsf{IL}}) \mid \Box \mathrm{F}_{\mathsf{IL}} \mid (\mathrm{F}_{\mathsf{IL}} \rhd \mathrm{F}_{\mathsf{IL}}) \mid (\mathrm{F}_{\mathsf{IL}} \owedge \mathrm{F}_{\mathsf{IL}}) \tag{68}$$

We have the convention that $\owedge$ binds equally strong as $\wedge$ and $\vee$. Thus we shall write $A \owedge B \rhd A$ instead of $(A \owedge B) \rhd A$, and $A \owedge B \to A$ instead of $(A \owedge B) \to A$. We say at that a formula of ILMS is a $\owedge$–*formula* if $\owedge$ is its main connective.

Definition 64. The logic ILMS is ILM plus the axiom S:

$$(C \rhd A) \wedge (C \rhd B) \leftrightarrow C \rhd A \owedge B$$ ❑

Lemma 1.1. *The following are provable in* ILMS*:*

1. $\Box\neg(\bot \owedge A)$
2. $\Diamond(A \owedge B) \to \Diamond A \wedge \Diamond B$
3. $\Diamond A \leftrightarrow \Diamond(A \owedge A)$
4. $\Diamond(A \owedge A) \leftrightarrow \Diamond(A \owedge (A \vee B))$
5. $\Diamond(\Diamond A \wedge \Diamond B) \to \Diamond(A \owedge B)$

Proof. For 1 we have $\bot \owedge \bot \rhd A$ by axiom S, whence by J4 $\Diamond(\bot \owedge A) \to \Diamond\bot$. Since $\Diamond\bot \to \bot$, it follows that $\Diamond(\bot \owedge A) \to \bot$, i.e. $\Box\neg(\bot \owedge A)$. Items 2, 3 and 4 follow similarly by using axioms S and J4. For 5 we use that by J5, $\Diamond A \rhd A$. By

J1, $\Diamond A \wedge \Diamond B \rhd \Diamond A$, whence by J2 $\Diamond A \wedge \Diamond B \rhd A$. Similarly, $\Diamond A \wedge \Diamond B \rhd B$. By axiom S, $\Diamond A \wedge \Diamond B \rhd A \owedge B$, and 5 follows by J4. ❑

Recall that according to Theorem 33 the sentence $\Diamond A \wedge \Diamond B \to \Diamond(A \owedge B)$ is true, and according to Theorem 37 it is not provable in ILMS. In the context of GL, we get a logic that proves all true sentences by simply taking all theorems of GL, adding the reflection principle $\Box A \to A$ (or by contraposition, $A \to \Diamond A$) and allowing modus ponens as the only rule of inference. The resulting system GLS is the provability logic of sentences that are true (in the standard model)[1]. One might wonder whether the same strategy gives all true sentences for the system ILMS. Item 5 in Lemma 1.1 suggests that this could indeed be the case. By reflection principle, we have $\Diamond A \wedge \Diamond B \to \Diamond(\Diamond A \wedge \Diamond B)$, whence using 5, we get $\Diamond A \wedge \Diamond B \to \Diamond(A \owedge B)$.

2. Axiom S and Structural Properties of Models

In this section, we will try to find a structural characterization of ILM–frames or ILM–models validating axiom S. As the possibility of such a characterization will be ruled out in Section 4 below, the reader should not get too attached to the material presented in this section. Instead, it should be seen as preliminary to the negative result of Section 4.2, and as an illustration for how the natural strategies for finding a modal semantics break down in the case of ILMS.

2.1. Minimal conditions. When encountering a new axiom in the context of modal logic, it is natural to ask whether it characterizes some nicely describable class of frames. We will now try to answer this question for the axiom S. However, as we do not have any conditions for a node to satisfy a formula of the form $A \owedge B$ (to investigate whether such conditions can be found is the purpose of the current chapter!), we will not arrive at fully fledged frame conditions. Instead, we will obtain certain minimal structural conditions that are sufficient for an ILM–model to validate axiom S.

Let $\mathcal{K}$ be the class of ILM–frames. Since we want our hypothetical ILMS–frames also to satisfy ILM, our question is whether there is some reasonable class $\mathcal{K}^{\owedge}$ of frames with $\mathcal{K}^{\owedge} \subseteq \mathcal{K}$, and such that

$$\mathcal{F} \in \mathcal{K}^{\owedge} \Rightarrow \mathcal{F} \Vdash \mathsf{S}.$$

Without having a structural truth condition for $\owedge$–formulas, it is not clear how one should proceed to investigate the other direction of frame characterizability. In this section, as well as in Chapter 5, we shall only be concerned with the direction of frame characterizability depicted above[2]. So let $\mathcal{M}$ be an ILM–model. We will see what structural characteristics $\mathcal{M}$ should possess in order for us to conclude that axiom S is valid in $\mathcal{M}$.

First, consider the direction of S which says that $(A \owedge B \rhd A) \wedge (A \owedge B \rhd B)$. Suppose that $\mathcal{M}, x \Vdash A \owedge B$ for some $x \in \mathcal{M}$. Using the semantics for $\rhd$, we see that for

[1] For more about the system GLS, see [**Boo93**].

[2] Under some circumstances, this might be the only direction we need. Suppose that we have a class of frames $\mathcal{K}^{\owedge}$ with $\mathcal{F} \in \mathcal{K}^{\owedge} \Rightarrow \mathcal{F} \Vdash \mathsf{S}$, but $\mathcal{F} \Vdash \mathsf{S} \Rightarrow \mathcal{F} \in \mathcal{K}'$ for some $\mathcal{K}' \neq \mathcal{K}^{\owedge}$, where of course $\mathcal{K}^{\owedge} \subset \mathcal{K}'$. Now if there is no modal formula that is valid in $\mathcal{K}^{\owedge}$ but not in $\mathcal{K}'$, we might still get soundness and completeness of ILMS w.r.t. the smaller class $\mathcal{K}^{\owedge}$ of frames.

any w s.t. wRx, there should be y, y' with $xS_wy \Vdash A$ and $xS_wy' \Vdash B$. Thus we get the first minimal condition:

$$x \Vdash A \between B \Rightarrow \forall w[wRx \Rightarrow \exists y, y'(xS_wy \Vdash A \ \wedge \ xS_wy' \Vdash B)]. \tag{69}$$

Suppose that we are able to find a condition $\mathcal{C}(A, B)$ for a node x in a model to satisfy the formula $A \between B$. Then the first minimal condition would become: if $x \vDash \mathcal{C}(A, B)$, then for all w with wRx, there are y, y' with $xS_wy \Vdash A$ and $xS_wy' \Vdash B$.

For the other direction of axiom S, let $w \Vdash (C \rhd A) \wedge (C \rhd B)$, and let x be s.t. $wRx \Vdash C$. Then we have y, y' with $xS_wy \Vdash A$ and $xS_wy' \Vdash B$. We want to conclude that there is some z with xS_wz and $z \Vdash A \between B$. Thus we get the second minimal condition[3]:

$$xS_wy \Vdash A \wedge xS_wy' \Vdash B \Rightarrow \exists z(xS_wz \Vdash A \between B). \tag{70}$$

Using our hypothetical condition $\mathcal{C}(A, B)$ again, (70) becomes: if $xS_wy \Vdash A$ and $xS_wy' \Vdash B$, then there exists z with xS_wz and $z \vDash \mathcal{C}(A, B)$.

If $\mathcal{M}$ is an ILM–model satisfying the minimal conditions (69) and (70), then clearly $\mathcal{M} \vDash \mathsf{S}$. We would want to find a condition $\mathcal{C}$ with two free variables, and such that if $\mathcal{M}$ is an ILM–model and we define for all $x \in \mathcal{M}$,

$$\mathcal{M}, x \Vdash A \between B :\Leftrightarrow x \vDash \mathcal{C}(A, B),$$

then $\mathcal{M}$ satisfies the minimal conditions (69) and (70) above. It would follow that $\mathcal{M} \Vdash \mathsf{S}$, and thus we would have a notion of an ILMS–model (i.e. an ILM–model where the truth values of $\between$–formulas are defined using the condition $\mathcal{C}$). In the following, we will look more closely at what such a condition $\mathcal{C}$ could be.

2.2. Relational semantics. In the context of modal logic, it is natural to expect the condition $\mathcal{C}(A, B)$ to mention the truth of A and B at nodes bearing a certain (fixed) relation to x in the underlying frame of the model $\mathcal{M}$. In this way, the truth value of $A \between B$ at x is completely determined by the truth values of A and B at nodes that stand in the relevant relation to x. If we are able to find such a condition $\mathcal{C}$, we say that we have a *relational semantics* for $\between$. The semantics for $\Box$ and for $\rhd$ is relational in this sense.

The only thing we require from $\mathcal{C}$ is that it only uses the relations of the underlying frame of the model. These could be R- or S-relations. However, we also allow *new* relations to be added to ILM–frames in order to deal with the new connective $\between$. We will see in Chapter 5 how such a freedom could in principle be used to get a relational semantics for ILMS. We explore a semantics that arises when adding to the ILM–frames a new relation Q, and using it to define the truth values of $\between$–formulas.

Given a condition $\mathcal{C}$ as described above, the minimal conditions from Section 2.1 can be transformed into actual frame conditions. The first one would say that if x stands in the specified relation to nodes y and y', then for all w with wRx, we have that xS_wy and xS_wy'. The second one would say that if xS_wy and xS_wy', then there is some z with xS_wz, and z stands in the specified relation to y and y'.

[3]See Section 4.1 for a discussion of the argument by which the second minimal condition is obtained.

An ILMS–frame could then be defined as an ILM–frame satisfying the above frame conditions. By definition, any model on an ILMS–frame would validate axiom S.

A relational semantics would thus give us more than just the notion of an ILMS–model — it would allow us to point out a class of frames such that any model whose underlying frame is from this class is automatically a model of ILMS. To establish such a strong connection between structure and validity of formulas is of course the desired outcome in a modal-logical context. However, we can also lower our ambitions and require a less strong connection between structure and validity. This brings us to set semantics.

2.3. Set semantics. In set semantics, the extension[4] of $A \mathbin{⍉} B$ in a model $\mathcal{M}$ is the output of a function $f_{⍉}$ given as inputs the extensions of A and B in $\mathcal{M}$. Let $[C]_{\mathcal{M}}$ denote the extension of C in $\mathcal{M}$. So the condition $\mathcal{C}(A, B)$ would become: being an element of $f_{⍉}([A]_{\mathcal{M}}, [B]_{\mathcal{M}})$. I.o.w. we would have:

$$\mathcal{M}, x \Vdash A \mathbin{⍉} B :\Leftrightarrow x \in f_{⍉}([A]_{\mathcal{M}}, [B]_{\mathcal{M}}). \tag{71}$$

Whereas in relational semantics, the truth value of $A \mathbin{⍉} B$ at x depended on the truth values of A and B at nodes bearing a certain relation to x, in set semantics it depends on the truth values of A and B at *all* nodes in the model.

Let $\mathcal{M}$ be an ILM–model where truth values of ⍉–formulas are defined as in (71). In order to guarantee that $\mathcal{M}$ satisfies all instances of axiom S, the function $f_{⍉}$ has to satisfy certain properties in $\mathcal{M}$. These properties are in fact just the minimal conditions of Section 2.1, translated into the terminology of set semantics.

i. if $x \in f_{⍉}(Y, Y')$, then for all w s.t. wRx, there are $y \in Y$ and $y' \in Y'$ with xS_wy and xS_wy'.

ii. if xS_wy and xS_wy' for some $y \in Y$, $y' \in Y'$, there is some z with xS_wz and $z \in f_{⍉}(Y, Y')$.

It is easy to see that if $\mathcal{M}$ is an ILM–model where the truth values of ⍉–formulas are determined through a function $f_{⍉}$ satisfying the above properties in $\mathcal{M}$, then $\mathcal{M} \Vdash$ S. Unlike in relational semantics, where the validity of S on a model can be guaranteed by structural properties of the underlying frame, in set semantics it can only be guaranteed by structural properties of the model itself.

3. Modest Modal Semantics

Both relational and set semantics would give us a uniform way of defining the truth values of ⍉–formulas in an ILM–model. If the model also satisfies the minimal conditions (69) and (70) above, it will automatically validate axiom S. Both of these approaches depend on us being able to find ILM–frames where the minimal conditions can be satisfied. We will show in Section 4 that the second minimal condition (70) cannot be successfully incorporated into ILM–frames, thus precluding the possibility of having either relational or set semantics for ⍉. If we still want to have a semantics for ILMS, we have to find some weaker notion of an ILMS–model. We will be modest: we only require an ILMS–model to validate axiom S.

[4]The extension of a formula A in a model $\mathcal{M}$ is the set of nodes in $\mathcal{M}$ where A is true.

DEFINITION 65. An ILMS–*model* is an ILM–model $\mathcal{M}$ where each instance of axiom S is forced at each node. If $\mathcal{D}$ is a set of formulas, then an ILMS–*model restricted to* $\mathcal{D}$ is an ILM–model $\mathcal{M}$ validating $(C \rhd A) \wedge (C \rhd B) \leftrightarrow C \rhd A \owedge B$ for all $C, A, B, (A \owedge B) \in \mathcal{D}$. ❑

If $\mathcal{D}$ is fixed or arbitrary, we also say *restricted* ILMS–*model* instead of ILMS–model restricted to $\mathcal{D}$. If $\mathcal{D}$ is finite, we can obtain an ILMS–model restricted to $\mathcal{D}$ by using the modal completeness of ILM. The idea is to treat the $\owedge$–formulas in $\mathcal{D}$ as new atoms, making sure that the model validates all instances of axiom S involving only formulas from $\mathcal{D}$ (as specified in Definition 65 above). Thus we obtain a model of ILM validating finitely many instances of axiom S. Now, the question is whether such a model can be extended to an ILMS–model as defined above, i.e. whether we can make *all* instances of S true at all nodes of $\mathcal{M}$.

The above method was used by Guaspari and Solovay ([**GS79**]) to prove modal completeness for the system R^-. A simplified proof was given by de Jongh in [**dJ87**]. The system R^- is GL plus witness comparison[5] $A \preceq B$ for $\Box$–formulas A and B. The behaviour of the witness comparison formulas is constrained by certain "order axioms". The models for R^- have no structural condition $\mathcal{C}$ of the form: $x \Vdash A \prec B \Leftrightarrow x \vDash \mathcal{C}(A, B)$. Instead, it is only required that each instance of the order axioms is forced at each node. In the proof of modal completeness, one first uses the modal completeness of GL to obtain a model validating order axioms involving only formulas from some finite set. After that, it is shown that such a restricted model can be extended to a model for the whole language, by proving a so-called *extension lemma*.

We will use the same strategy to obtain ILMS–models. However, we will only be able to prove an extension lemma for models of depth 2. This will be done in Section 5 below. Section 3.2 spells out how the modal completeness of ILM can be used to obtain restricted ILMS–models. As a preliminary, we will give an overview of the modal completeness proof of ILM by the construction method. A thorough exposition of this proof is given in the appendix.

3.1. Modal completeness of ILM. Let $\mathcal{K}$ be the class of ILM–frames. As usual in completeness proofs, the modal completeness of ILM is proved by contraposition. Given a sentence A with $\nvdash_{\mathsf{ILM}} A$, we will find an ILM–model $\mathcal{M}$ with $\mathcal{M} \nVdash A$.

We will follow the proof by construction method, as presented in [**GJ10**]. The main idea of the proof is as follows. If $\nvdash_{\mathsf{ILM}} A$, there is a maximal–ILM–consistent set Γ with $\neg A \in \Gamma$. We will build an ILM–frame $\mathcal{F}$, where every node x is labeled[6] with a maximal–ILM–consistent set $\nu(x)$. We start with a node w and $\nu(w) = \Gamma$; w will

[5]For a a Σ_1-sentence $\exists x \varphi(x)$ its *witness* is a number n with $\mathbb{N} \vDash \varphi(n)$. If A and B are Σ_1-sentences, we write $A \preceq B$ to express that the smallest witness for A is smaller than the smallest witness for B. Thus a Rosser sentence for PA can be seen as a sentence R with $\vdash_{\mathsf{PA}} R \leftrightarrow \Box\neg R \preceq \Box R$ (remember that $\Box A$ is a Σ_1–formula).

[6]We cannot identify nodes with the maximal consistent sets, as the latter might need to occur at several places in the model. This is related to ILM–models having a complicated notion of a bisimulation, with no obvious notion of a minimal bisimilar model. This means that bisimilar or modally equivalent nodes cannot be always identified. We will see an example of this in Section 4. See [**Vis98b**] for a definition of bisimilarity for IL–models.

be the root of $\mathcal{F}$. $\mathcal{F}$ will be built step by step, using the information contained in the maximal consistent sets labeling the nodes. Finally, $\mathcal{F}$ is transformed into a model $\mathcal{M}$ by defining a valuation on $\mathcal{F}$ as: $\mathcal{M}, x \Vdash p :\Leftrightarrow p \in \nu(x)$.

We want to build $\mathcal{F}$ in such a way that the harmony between truth in $\mathcal{M}$ and membership in the maximal consistent sets labeling the nodes extends to a larger set than just the propositional formulas. Our goal is to have for all x,

$$\mathcal{M}, x \Vdash A :\Leftrightarrow A \in \nu(x), \tag{72}$$

as then we could conclude that $\mathcal{M}, w \Vdash \neg A.$ and $\mathcal{M}$ would be our required countermodel to A. In order to ensure (72), it suffices to have a certain adequate set $\mathcal{D}$ containing A s.t. for all $B \in \mathcal{D}$, $x \in \mathcal{M}$,

$$\mathcal{M}, x \Vdash B \Leftrightarrow B \in \nu(x). \tag{73}$$

We call the equivalence in (73) a *truth lemma w.r.t.* $\mathcal{D}$.

Digression 66. Since ILM is not compact[7], we cannot in general expect to get a truth lemma w.r.t. the entire language. Hence we will only require (73) to hold w.r.t. an adequate set $\mathcal{D}$ which is big enough to allow inductive reasoning when proving the truth lemma, but small enough to block the incompactness phenomenon. For this, it suffices if $\mathcal{D}$ is finite, and closed under subformulas and single negations.

For later purposes, it is convenient if our adequate set $\mathcal{D}$ is also closed under boolean operations. Upon this requirement, $\mathcal{D}$ is of course not *actually* finite. However, closing off under boolean operations only adds a finite number of formulas which are not logically equivalent to formulas already in $\mathcal{D}$, so that $\mathcal{D}$ will only contain finitely many formulas up to logical equivalence. Such a set is "finite enough" for the purposes of blocking the incompactness phenomenon mentioned above. We will from now on often say that our adequate set is finite, even though in reality it only contains finitely many formulas up to logical equivalence. ❑

Let $\mathcal{D}$ be an adequate set containing A. When constructing $\mathcal{F}$, we will ensure:

1. a truth lemma holds in $\mathcal{F}$ w.r.t. $\mathcal{D}$
2. $\mathcal{F}$ is an ILM–frame

The idea of the construction is to approximate the truth lemma by eliminating so-called problems and deficiencies in $\mathcal{F}$. Let w be the root of $\mathcal{F}$ as above, and suppose that $\neg(B \rhd C) \in \mathcal{D} \cap \nu(w)$. For the truth lemma to hold, there has to be some x with wRx, s.t. $B \in \nu(x)$ and for all z with xS_wz, $\neg C \in \nu(z)$. If such an x does not exist, we say that $\neg(B \rhd C)$ is a *problem* in w. In order to eliminate this problem, we add to $\mathcal{F}$ an R-successor x of w with the required properties. The "tricky" part is to ensure that at no later stage in the construction, there will be a z with $xS_wz \Vdash C$, as this could ruin the truth of $\neg(B \rhd C)$ at w. This will be taken care of by using the notion of a *C-critical cone* above w. This C-critical cone contains x, and any z with xS_wz. In order to have the truth lemma, we have to guarantee that if z is in the C-critical cone above w, then $z \Vdash \neg C$. In general,

[7]In fact, compactness already fails for GL. This means that there exist maximal–GL–consistent sets which cannot be satisfied on a transitive conversely well-founded frame. The proof is by Fine and Rautenberg, and is treated in Chapter 7 of [**Boo93**].

there can be nothing S_w-above x that would demand for an S_w-transition to a node satisfying C.

Now suppose that $B \rhd D \in \mathcal{D} \cap \nu(w)$. For the truth lemma to hold, every R-successor x of w with $B \in \nu(x)$ has to have an S_w-successor z with $D \in \nu(z)$. If this is not the case, we say that $B \rhd D$ is a *deficiency* in w. In order to eliminate a deficiency, we add to $\mathcal{F}$ a node z with the required properties. It is clear that once a deficiency in w has been eliminated, it cannot reoccur.

When eliminating a problem or a deficiency in w, we introduce new R-successors of w, thus potentially generating many new deficiencies in w. Moreover, each node that we add to $\mathcal{F}$ potentially contains new problems and deficiencies itself. For the truth lemma to hold, all these problems and deficiencies have to be eliminated eventually.

$\mathcal{F}$ will be constructed as the union of an infinite chain of ILM–frames $\{F_n\}_{n\in\omega}$. Each F_n contains a deficiency or a problem less than its predecessor, thus with each F_n we can claim to be closer to the truth lemma (w.r.t. $\mathcal{D}$). In order to make sure that each F_n is an ILM–frame, F_n itself is also constructed as the union of an infinite chain $\{G_i\}_{i\in\omega}$, where each G_i is closer to being an ILM–frame than the previous one. In [**Joo98**] it is shown that with some care in the construction, we can even guarantee that the constructed frame $\mathcal{F}$ is finite. A more thorough exposition of the proof of the modal completeness of ILM by the construction method is given in the appendix.

3.2. Restricted modal completeness of ILMS. We will now explain how the modal completeness of ILM can be used to get modal completeness of ILMS w.r.t. restricted ILMS–frames.

Suppose that $\nvdash_{\mathsf{ILMS}} A$. Let $\mathcal{D}$ be the smallest set containing A closed under subformulas and boolean operations. Let $C_0 \ldots, C_n$ be the $\between$–formulas of $\mathcal{D}$, and let $p_1, \ldots, p_n$ be distinct atoms not occurring in $\mathcal{D}$. We translate the formulas in $\mathcal{D}$ into formulas of ILM: to every $B \in \mathcal{D}$, we associate a $B^\star$ in the language of ILM s.t. B is the result of substituting C_i for p_i throughout $B^\star$, with $0 \leq i \leq n$.

Let $\Psi^\star$ be the set of ILM–formulas consisting of all formulas of the form $B^\star \wedge \Box B^\star$, where B is $(C \rhd A) \wedge (C \rhd B) \leftrightarrow C \rhd A \between B$ with $C, A, B, (A \between B) \in \mathcal{D}$. We say that B is an instance of axiom S involving only formulas in $\mathcal{D}$. Note that since $\mathcal{D}$ is finite, also $\Psi^\star$ must be finite[8].

Since $\nvdash_{\mathsf{ILMS}} A$, it follows that $\nvdash_{\mathsf{ILM}} \bigwedge \Psi^\star \to A^\star$. Then suppose that there is an ILM–proof of $\bigwedge \Psi^\star \to A^\star$. Let Ψ be the result of replacing $B^\star \wedge \Box B^\star$ in $\Psi^\star$ by $B \wedge \Box B$. Then $\vdash_{\mathsf{ILMS}} \bigwedge \Psi \to A$, since ILMS has the same proof rules available as ILM, and it satisfies the substitution rule $B(p_i)/B(C_i)$. But clearly, $\vdash_{\mathsf{ILMS}} \bigwedge \Psi$, since S is an axiom scheme in ILMS, and ILMS has the rule of necessitation. Hence $\vdash_{\mathsf{ILMS}} A$, contradiction. By modal completeness for ILM, there is a finite ILM–model $\mathcal{M}^\star$ s.t. $\bigwedge \Psi^\star \wedge \neg A^\star$ is true at the root of $\mathcal{M}^\star$. We turn $\mathcal{M}^\star$ into a model $\mathcal{M}$ for

[8]In the case that $\mathcal{D}$ is not actually finite, but only has finitely many equivalence classes of formulas up to logical equivalence, we can just take $B^\star \wedge \Box B^\star$, where B is a representative of its equivalence class in $\mathcal{D}$.

the language of ILMS by defining, for all $B \in \mathcal{D}$,

$$\mathcal{M}, x \Vdash B :\Leftrightarrow \mathcal{M}^\star, x \Vdash B^\star \tag{74}$$

The forcing in $\mathcal{M}$ and $\mathcal{M}^\star$ agrees on ILM–formulas. Since $w \Vdash \bigwedge \Psi$, every instance of axiom S involving only formulas from $\mathcal{D}$ is valid in $\mathcal{M}$. To conclude, we have shown that if $\nvdash_{\mathsf{ILMS}} A$, then there is a restricted ILMS–model $\mathcal{M}$ with $\mathcal{M} \nVdash A$.

3.3. Restricted arithmetical completeness of ILMS. Let $\mathcal{M}$ be an ILMS–model restricted to an adequate set $\mathcal{D}$. $\mathcal{M}$ can be seen as just an ILM–model — think of the model $\mathcal{M}^\star$ above, where instead of $\between$–formulas we have just new atoms satisfying certain properties w.r.t. interpretability. Hence we can apply to $\mathcal{M}$ the arithmetical completeness of ILM, obtaining a restricted version of arithmetical completeness for ILMS. Let us illustrate this with an example.

Suppose that the unprovable[9] sentence A of ILMS we are interested in is $\top \between \top$. In order to apply modal completeness of ILM, we take an adequate set $\mathcal{D}$ containing $\top \between \top$ that is closed under subformulas and single negations (we forget about the boolean combinations for the moment). Our adequate set $\mathcal{D}$ will be rather simple: $\mathcal{D} = \{\top, \bot, \top \between \top, \neg(\top \between \top)\}$. We now have to consider all instances of axiom S involving only formulas from $\mathcal{D}$. It is easy to see that the only one of them which is not provable[10] in ILM is $\neg(\top \between \top) \rhd (\top \between \top)$. Let p be the proposition letter that we use as a representative for $\top \between \top$ in the context of ILM. Applying the modal completeness proof for ILM as described above, we get a finite ILM–model where the sentence $(\neg p \rhd p) \wedge \Box(\neg p \rhd p) \wedge \neg p$ is forced at the root w. Using the arithmetical completeness of ILM, we get an arithmetical realization $*$ s.t. $\nvdash_{\mathsf{PA}} (\neg p^* \rhd p^*) \wedge \Box(\neg p^* \rhd p^*) \to p^*$.

Remember that the propositional letter p was our representative, in the context of ILM, of the ILMS–formula $\top \between \top$. Now it is a valid question to ask whether the arithmetical sentence p^* obtained via the arithmetical completeness of ILM is in fact in the same degree as $\top \between \top$. It is not obvious why this should be the case. The real supremum of A and B has the property that for *all* sentences C, $(C \rhd A) \wedge (C \rhd B) \to C \rhd A \between B$. However, in our restricted ILMS–model only finitely many such sentences C are taken into account. Thus it might happen that the realization $*$ from the arithmetical completeness proof gives us as $(A \between B)^*$ an arithmetical sentence which is strictly above[11] the *real* supremum of A and B in the lattice of interpretability.

In our example, p^* will actually be in the degree of the real supremum $\top \between \top$. This is because $\vdash_{\mathsf{IL}} \neg p \rhd p \to \top \rhd p$ (using J3), whence by arithmetical soundness of ILM, $\vdash_{\mathsf{PA}} \neg p^* \rhd p^* \wedge \Box(\neg p^* \rhd p^*) \to \top \rhd p^*$. Since $\vdash_{\mathsf{IL}} p \rhd \top$ for all p, again by arithmetical soundness, $\vdash_{\mathsf{PA}} \neg p^* \rhd p^* \wedge \Box(\neg p^* \rhd p^*) \to \top \equiv p^*$. Thus PA proves that the arithmetical representative of $\top \between \top$ obtained by using the arithmetical completeness of ILM is indeed in the same degree as $\top \between \top$, i.e. in the degree of $\top$. Whether this is always the case is a question for future research.

---

[9] Since $\top \between \top$ is an Orey sentence under Visser's implementation (see Section 5.3 of Chapter 3), it cannot be provable in ILMS.

[10] The other ones are all of the form $A \rhd A$, $A \rhd \top$, $\bot \rhd A$, or just tautologies.

[11] Since $\mathcal{D}$ is required to be closed under subformulas, we will always take $A \between B \rhd A$ and $A \between B \rhd B$ into account, hence $(A \between B)^*$ cannot be lower than the real supremum.

3.4. Need for an extension lemma. Given the negative result of Section 4.2, the best we can hope for is that if $\nvdash_{\mathsf{ILMS}} A$, there is an ILM–model $\mathcal{M}$ where axiom S is valid and A is false, i.e. an ILMS–model in the sense of Definition 65. Above, we showed that if $\nvdash_{\mathsf{ILMS}} A$, there is an an ILM–model where A false, and finitely many instances of axiom S are valid, i.e. a restricted ILMS–model. If we could show that such a restricted ILMS–model can be extended to a proper ILMS–model, we would have completeness of ILMS w.r.t. the modest modal semantics. For this purpose, we would need to prove that any ILMS–model restricted to $\mathcal{D}$ can be extended to a model where *all*[12] instances of axiom S are valid. This is the so-called *extension lemma*. Guaspari and Solovay ([**GS79**]) use such an extension lemma to prove modal completeness for the system R^-.

In order get an idea what an extension lemma has to achieve, suppose that our adequate set contains the formulas $A \between B$ and $F \between G$. Since the method described in Section 3.2 treats these formulas as atoms, their truth values will be defined everywhere in the model $\mathcal{M}$ obtained by using the modal completeness of ILM. As a consequence, the forcing of the formulas $\Box(A \between B)$ and $(A \between B) \rhd (F \between G)$ is determined everywhere in the model, too. But if $\Box(A \between B) \between (A \between B \rhd F \between G)$ was not an element of $\mathcal{D}$, its forcing in $\mathcal{M}$ is undefined. After defining it, truth values of more $\Box$– and $\rhd$–formulas become defined, and thus we have to take care of their suprema in turn.

If we are able to show that the truth values of all $\between$–formulas can be defined eventually, such that the axiom S is valid in the resulting model, we say that we have an *extension lemma*. In Section 5, we will prove an extension lemma for restricted ILMS–models whose root satisfies $\Box\Box\bot$.

4. The Impossibility of a Structural Characterization

In this section, we will show that the second minimal condition from Section 2.1 cannot be successfully incorporated into ILM–frames, thus ruling out the possibility of having a relational or set semantics for ILMS.

4.1. A closer look at the second minimal condition. Let us recall how the second minimal condition (70) in Section 2.1 was obtained. We assumed that $w \Vdash (C \rhd A) \wedge (C \rhd B)$, and wanted to see what would allow us to conclude that $w \Vdash C \rhd A \between B$. So we took an x with $wRx \Vdash C$, whence the assumption gave us y and y' with $xS_wy \Vdash A$ and $xS_wy' \Vdash B$. We wanted to have some z with $xS_wz \Vdash A \between B$. Thus we formulated our minimal condition as:

$$xS_wy \Vdash A \wedge xS_wy' \Vdash B \Rightarrow \exists z(xS_wz \Vdash A \between B).$$

But note that in our argument we had assumed that $w \Vdash (C \rhd A) \wedge (C \rhd B)$ and $x \Vdash C$, whereas this assumption does not figure in the minimal requirement we obtained. So shouldn't the real minimal requirement be instead:

$$w \Vdash (C \rhd A) \wedge (C \rhd B) \wedge x \Vdash C \Rightarrow \exists z(xS_wz \Vdash A \between B)? \tag{75}$$

[12]By *all*, we here mean all instances in the language of the finitely many propositional letters contained in our adequate set $\mathcal{D}$.

But this "condition" is not informative — it merely rephrases axiom S in terms of the semantics for $\rhd$. It is not the kind of *structural* condition we are looking for.

Let us see what could go wrong as a result of us disregarding the initial assumptions of the argument. Suppose there are x, y and y' with $xS_wy \Vdash A$ and $xS_wy' \Vdash B$. Suppose further that there is no C with $w \Vdash (C \rhd A) \wedge (C \rhd B)$ and $x \Vdash C$. In this situation, the second minimal condition requires something that is not really necessary — axiom S could be valid even if there would be no z with xS_wz and $z \Vdash A \mathbin{\owedge} B$. As we will see, this unnecessary requirement on behalf of the second minimal condition can in fact lead to an inconsistent situation on an ILM–model.

On the other hand, the way we arrived at the second minimal condition is standard for obtaining frame conditions in modal logic. So why does it lead us astray here? Let us see under which conditions it would have been alright for us to have "forgetten" the assumptions $w \Vdash C \rhd A \wedge C \rhd B$ and $wRx \Vdash C$. Suppose that there is some sentence E that uniquely defines x in the model, in the sense that for all y, $y \Vdash E \Leftrightarrow y = x$. Now if $xS_wy \Vdash A$ and $xS_wy' \Vdash B$, then using the semantics of $\rhd$, $w \Vdash E \rhd A \wedge E \rhd B$. Hence to guarantee the validity of S, we would indeed need some z with $xS_wz \Vdash A \mathbin{\owedge} B$.

As we will see, the situation where the second minimal condition leads us into trouble is indeed one where two "essentially different" nodes in a model satisfy exactly the same sets of formulas. If the models we are dealing with are image finite, i.e. if for every w, the set $\{x \mid wRx\}$ is finite, then according to the Hennessy-Milner Theorem for interpretability logics (see [**dJ04**]), such nodes are also bisimilar. Now, in a normal modal-logical context, bisimilar nodes can always be identified, because of the existence of a minimal bisimilar model. This means that given a model $\mathcal{M}$, we can always find a model $\mathcal{M}'$ where all bisimilar states of $\mathcal{M}$ are identified. As a consequence, every node of $\mathcal{M}'$ is uniquely characterized by the set of formulas it satisfies. If moreover, $\mathcal{M}'$ is assumed to be finite, then each node in $\mathcal{M}'$ is uniquely defined by some formula. Hence if all that would hold for our models, the argumentation leading us to the second minimal condition would have been sound.

However, as mentioned in footnote 6 above, in IL–models one cannot in general identify two nodes that are bisimilar or satisfy the same sets of formulas. This is the reason why the standard modal-logical route to frame conditions fails in the case of ILMS. We will now see a concrete example of this.

4.2. Counterexample to the second minimal condition. In this section, we construct an ILMS–model (as in Definition 65.) which *cannot* be required to satisfy the second minimal condition. This shows that the latter is indeed too strong, as speculated above.

THEOREM 67. *There exists an* ILMS*-consistent set* $\mathcal{X}$ *of sentences such that no* ILM*-model satisfying* $\mathcal{X}$ *can satisfy the second minimal condition (70).*

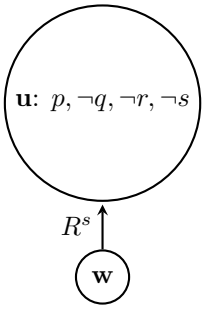
PROOF. Consider the following set of formulas:

$$\mathcal{X} = \{\neg(p \rhd s), \neg(p \rhd q), r \mathbin{\owedge} s \rhd q, p \rhd r, p \rhd q \vee s, \Box\Box\bot\}$$

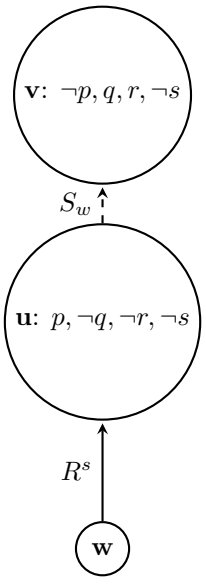
Let $\mathcal{D}$ be the smallest set containing $\mathcal{X}$ closed under subformulas and boolean combinations. We will use the method of Section 3.2 to get an ILMS–model $\mathcal{M}$

restricted to $\mathcal{D}$, where all formulas in $\mathcal{X}$ are true at the root w. In Section 5, we will show that (since $\Box\Box\bot \in \mathcal{X}$), $\mathcal{M}$ can be extended to an ILMS–model $\mathcal{M}'$, i.e. we can make all instances of S valid on $\mathcal{M}$. We will first go through the construction of $\mathcal{M}$ step by step.

In order for $\neg(p \rhd s)$ to be true at w, there has to be a node u with[13] $wR^s u \Vdash p$, and u is in the s-critical cone above w. Thus if $w \Vdash E \rhd s$, then $u' \Vdash \neg E$ for any u' with $uS_w u'$. In particular, $u' \Vdash \neg s$, $u' \Vdash \neg(r \between s)$, and $u' \Vdash \neg(r \wedge s)$. Thus[14]:



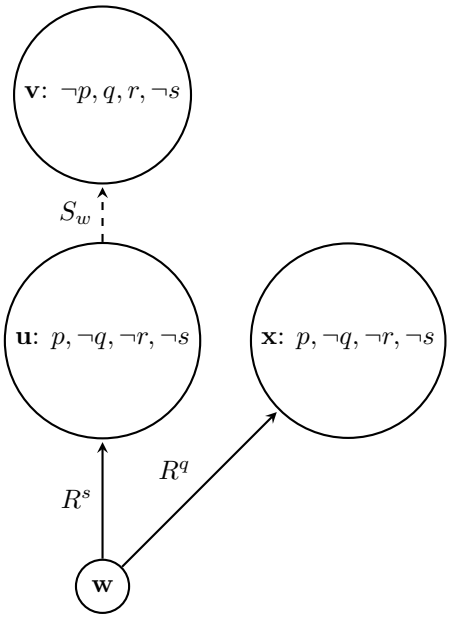
Sentences $p \rhd r$ and $p \rhd q \vee s$ are now deficiencies in w w.r.t. u. These deficiencies are eliminated by adding a node v with $uS_w v \Vdash r \wedge q$. Note that the deficiency caused by $p \rhd q \vee s$ needs to be eliminated by a node containing q, as every node S_w-above u has to satisfy $\neg s$. We are only free to choose the value of p. We get the following picture:



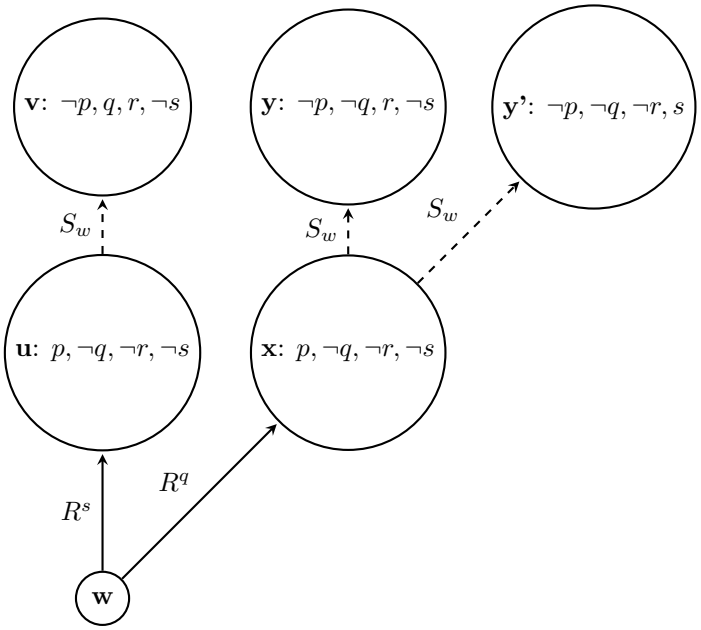
In order for $\neg(p \rhd q)$ to be true at w, there has to be a node x with $wR^q x \Vdash p$, and x is in the q-critical cone above w. Thus if $w \Vdash E \rhd q$ and $xS_w x'$, we have $x' \Vdash \neg E$. In particular $x' \Vdash \neg q$, $x' \Vdash \neg(r \between s)$, and $x' \Vdash \neg(r \wedge s)$. Thus:

[13]We write $wR^s u$ to indicate that R leads into an s-critical cone above w.

[14]We will not indicate the truth value of $r \between s$, as it fill be false at every x that is R-above w.



Sentences $p \rhd r$ and $p \rhd q \vee s$ are now deficiencies in w w.r.t. x. To eliminate the deficiency caused by $p \rhd r$, we add a node $y \Vdash r$. Since $y \Vdash \neg(r \wedge s)$, we have to put $y \Vdash \neg s$. Thus in order to eliminate the deficiency caused by $p \rhd q \vee s$, another node y' has to be added. Since y' is in a q-critical area, the deficiency $p \rhd q \vee s$ needs to be eliminated by adding a node that satisfies s. Since $y' \Vdash \neg(r \wedge s)$, we have $y' \Vdash \neg r$. Again, we are free to choose the value of p. We get the following model:

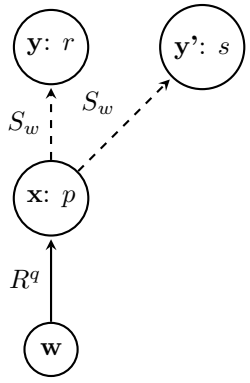


We will now check that $\mathcal{M}$ is indeed an ILMS–model restricted to $\mathcal{D}$, i.e. that all instances of axiom S involving only formulas from $\mathcal{D}$ are satisfied[15] in $\mathcal{M}$. Since for $x \neq w$, $x \Vdash \Box\bot$, it suffices to check that every such instance of S is true at w. Since $r \between s$ is the only $\between$formula in $\mathcal{D}$, we only have to check that $w \Vdash (r \between s \rhd r) \wedge (r \between s \rhd s)$, and if $C \in \mathcal{D}$, then $w \Vdash (C \rhd r) \wedge (C \rhd s) \rightarrow C \rhd r \between s$.

Since $r \between s$ is false everywhere R-above w, it is clear that $w \Vdash (r \between s \rhd r) \wedge (r \between s \rhd s)$. For the other direction, we will show that there is no $C \in \mathcal{D}$ with $w \Vdash C \rhd r \wedge C \rhd s$ and $w \Vdash \Diamond C$. But if $w \Vdash C \rhd r \wedge C \rhd s$ and $w \Vdash \Box\neg C$, then clearly $w \Vdash C \rhd A \between B$. Note that x is the only node in $\mathcal{M}$ that has S_w-access to an r-node and an s-node[16]. Hence if there is a C with the above properties, then we must have that $x \Vdash C$. However, u and x make true exactly the same formulas of $\mathcal{D}$. Thus if $x \Vdash C$, also $u \Vdash C$. But the assumption that $w \Vdash C \rhd s$ contradicts u being in the s-critical cone above w.

In the next section, we will show that $\mathcal{M}$ can be extended to a proper ILMS–model, i.e. we can make *all* instances of axiom S (in the language containing the propositional letters p, q, r and s) true at w. We will take this result as given for now. As a preparation for the proof, notice that it will be crucial to guarantee that x and u always satisfy exactly the same sentences. Then suppose that we define for some formula $A \between B$ that $x \Vdash (A \between B)$ and $u \Vdash \neg(A \between B)$. If now furthermore x is the only node R-above w where $A \between B$ is defined to be true, then $A \between B \rhd r \wedge A \between B \rhd s$ will be suddenly true at w. In order to validate axiom S, we need some z with $xS_wz \Vdash r \between s$. However, x is in the q-critical area above w, whence for all x' with xS_wx', we need that $x' \Vdash \neg(r \between s)$ (since $w \Vdash r \between s \rhd q$). If no new subsets of the model become definable during the process of extending $\mathcal{M}$ to an ILMS–model, this problem will not arise.

Finally, we will explain why the above model is a counterexample to the second minimal condition of Section 2.1. The second minimal condition is not satisfied in $\mathcal{M}$. We have that $xS_wy \Vdash r$ and $xS_wy' \Vdash s$, however there is no z such that $xS_wz \Vdash r \between s$. In fact, the second minimal condition must fail in *any* ILM–model where the sentences $\neg(p \rhd q)$, $p \rhd r$, $p \rhd s \vee q$, and $r \between s \rhd q$ are true at a point w. Necessarily, in such a model we will have the following situation:



[15]Remember from Section 3 that an instance of S is said to involve only formulas from $\mathcal{D}$ if it is of the form $(C \rhd A) \wedge (C \rhd B) \leftrightarrow C \rhd (A \between B)$ with $A, B, C, (A \between B) \in \mathcal{D}$.

[16]By an r-node we of course mean just a node where r is true.

There are y and y' s.t. $xS_wy \Vdash r$ and $xS_wy' \Vdash s$. However, there cannot be any z with $xS_wz \Vdash r \between s$. Since $w \Vdash r \between s \rhd q$, this would require a S_w-transition from z to a q-node, contradicting the fact that z (like all S_w-successors of x) is in the q-critical area above w.

To conclude, we have shown (taking the result of the next section for granted) that there is an ILMS–model $\mathcal{M}$ with the property that if we require $\mathcal{M}$ to satisfy the second minimal condition (70) then $\mathcal{M}$ will become inconsistent. ❑

5. Quest for an Extension Lemma

In this section, we will prove an extension lemma for models of depth 2. After that, we will give an example illustrating why the strategy used in this proof cannot be directly adapted to the more general case.

5.1. A small extension lemma. In this section, we will show that if the restricted ILMS–model obtained by modal completeness of ILM (as described in Section 3.2) has depth at most 2, then we can indeed extend it to a full ILMS–model. Recall from Section 3 that an ILMS–model restricted to $\mathcal{D}$ is an ILM–model that validates $(C \rhd A) \wedge (C \rhd B) \leftrightarrow C \rhd (A \between B)$ for all $A, B, C, (A \between B) \in \mathcal{D}$.

Remember that a set $\mathcal{D}$ of formulas is adequate if it finite, and closed under subformulas and single negations. Recall from the discussion in Section 3.1 that we can also require $\mathcal{D}$ to be closed under boolean operations. All that matters is that $\mathcal{D}$ contains only finitely many formulas up to provable equivalence. Throughout this section, we will thus require our adequate set $\mathcal{D}$ to be closed under boolean operations.

Lemma 5.1 (Extension lemma for models of depth 2). *Let $\mathcal{D}$ be an adequate set closed under boolean operations. Let $(\mathcal{M}, w)$ be a finite rooted ILMS–model restricted to $\mathcal{D}$, and let $\mathcal{M}, w \Vdash \Box\Box\bot$. Then $\mathcal{M}$ can be extended to an ILMS–model.*

We will first explain the intuitive idea behind the proof, introducing some terminology on the way. The simplicity of $\mathcal{M}$ is essential to the proof. If $x \neq w$, then $x \Vdash \Box\bot$. It follows that any boolean combination of $\Box$– and $\rhd$–formulas is true at x, and in particular any instance of axiom S. Thus we only need to ensure that $w \Vdash \mathsf{S}$. We will guarantee this in an infinite process, extending $\mathcal{M}$ to an ILMS–model $\mathcal{M}^{\mathsf{S}}$ in stages. At stage $t+1$, we define truth values for formulas of the form $A \between B$, with $A, B \in \mathcal{S}_t$. Once the forcing of a formula has been defined, it will not be changed at a later stage. The forcing of $\between$–formulas at w can be arbitrary, for example we can make all of them true at w. In the proof, we will just focus on the truth values of $\between$–formulas at nodes $x \neq w$.

If $x, y \neq w$, we say that x and y are *indistinguishable at stage t* if there is no $C \in \mathcal{S}_t$ with $x \Vdash C$ and $y \Vdash \neg C$. It is clear that $\sim_t$ is an equivalence relation. We write $[x]_{\sim_t}$ for the equivalence class of x under $\sim_t$. If it is not the case that $x \sim_t y$, we say that x and y are *distinguishable at stage t*. We say that a set X of nodes is *definable* (at stage t) if there is some formula C ($\in \mathcal{S}_t$) s.t. for all x, $x \in X \Leftrightarrow x \Vdash C$. In the process of extending $\mathcal{M}$ to an ILMS–model, we will guarantee that no new subsets of $\mathcal{M}$ become definable, i.o.w. that every definable

subset of $\mathcal{M}^{\mathsf{S}}$ is already definable by a formula in $\mathcal{D}$. In the light of our discussion in the end Section 4.2, it is desirable not to have new definable subsets, as these might require more supremum formulas to be true at certain nodes (in particular, supremum formulas whose forcing has already been defined).

If $\mathcal{S}$ is a set of formulas, then by $\mathcal{S}^+$ we denote the closure of $\mathcal{S}$ under boolean operations, $\Box$, and $\rhd$. Note that if the truth values of formulas in $\mathcal{S}$ are defined everywhere in a model, then also the truth values of formulas in $\mathcal{S}^+$ are automatically defined everywhere. We are now ready to prove Lemma 5.1.

PROOF. Let $\mathcal{S}_0 := \mathcal{D}^+$, and $\mathcal{S}_{t+1} := (\mathcal{S}_t \cup \{A \between B \mid A, B \in \mathcal{S}_t\})^+$.

At stage $t+1$, we define simultaneously the forcing of all $A \between B$ with $A, B \in \mathcal{S}_t$. So let $A, B \in \mathcal{S}_t$.

1. If there are A', B' with $w \Vdash \Box(A \leftrightarrow A') \wedge \Box(B \leftrightarrow B')$, and the forcing of $A' \between B'$ has already been defined at a previous stage, define $w \Vdash \Box(A \between B \leftrightarrow A' \between B')$. Since $x \neq w$ implies that wRx, $A \between B$ will then have the same extension as $A' \between B'$ at all nodes $x \neq w$. As a consequence, once the value of $\top \between \top$ has been defined, and if A and B are positive boolean combinations of $\Box$– and $\rhd$–formulas with $A \between B \notin \mathcal{D}$, we will have $w \Vdash \Box(\top \between \top \leftrightarrow A \between B)$.

2. If there are no A', B' as above, define $x \Vdash A \between B$ (for $x \neq w$) if

(76) for all x with $x \sim_t x'$, there are y, y' with $x' S_w y \Vdash A$ and $x' S_w y' \Vdash B$.

Otherwise, let $x \Vdash \neg(A \between B)$. Note that since $\sim_t$ is reflexive, if x fulfills the requirement in (76), then there must be y and y' with $x S_w y \Vdash A$ and $x S_w y' \Vdash B$.

Let $\mathcal{M}^{\mathsf{S}}$ be the resulting model, where the forcing of all $\between$–formulas has been defined. Note that according to clause 1 above, all newly defined $\between$–formulas in the model will satisfy extensionality, i.e. if $A \between B \notin \mathcal{D}$, then $\Box(A \leftrightarrow A') \wedge \Box(B \leftrightarrow B') \rightarrow \Box(A \between B \leftrightarrow A' \between B')$ is valid in $\mathcal{M}^{\mathsf{S}}$. Hence if also the forcing of $\between$–formulas from $\mathcal{D}$ was extensional in $\mathcal{M}$, $\between$ will be extensional in $\mathcal{M}^{\mathsf{S}}$.

Before showing that $\mathcal{M}^{\mathsf{S}} \Vdash \mathsf{S}$, it is convenient to show that no new subsets of $\mathcal{M} \setminus \{w\}$ become definable in the course of the process.

REMARK 68. Let $X \subsetneq \mathcal{M} \setminus \{w\}$, and suppose that X is closed under $\sim_t$. Then there is some $E \in \mathcal{D}$ s.t. for all x, $x \in X \Leftrightarrow x \Vdash E$.

Since $\mathcal{M}$ is finite, it suffices to show that for any $[x]_{\sim_t}$, there is some $E \in \mathcal{D}$ s.t. for all y, $y \in [x]_{\sim_t} \Leftrightarrow y \Vdash E$. Then if $X = \{[x_1]_{\sim_t}, \ldots, [x_n]_{\sim_t}\}$, we can take as E the disjunction[17] of the formulas in $\mathcal{D}$ defining $[x_1]_{\sim_t} \ldots [x_n]_{\sim_t}$. Note that then $E \in \mathcal{D}$, since $\mathcal{D}$ is closed under boolean operations. We will now prove the claim by induction on t.

For the base case, let $t = 0$. Since all $\Box$– and $\rhd$–formulas are true at all x with wRx, $[x]_{\sim_0}$ is uniquely characterized by a boolean combination of propositional letters and $\between$–formulas in $\mathcal{D}$. Since there are finitely many of those, and $\mathcal{D}$ is closed under boolean operations, it is clear that there is some $E \in \mathcal{D}$ that defines $[x]_{\sim_0}$. For the inductive step, assume that there is some $E \in \mathcal{D}$ with $y \in [x]_{\sim_t} \Leftrightarrow y \Vdash E$. We will show that $[x]_{\sim_t} = [x]_{\sim_{t+1}}$, whence E also defines $[x]_{\sim_{t+1}}$. If $[x]_{\sim_t} \neq [x]_{\sim_{t+1}}$,

[17] For this, we need $\mathcal{D}$ to only contain finitely many formulas up to provable equivalence.

then there must be some C whose truth value was defined at stage $t+1$, and such that C can distinguish between members of $[x]_{\sim_t}$. Clearly, C cannot be a boolean combination of $\Box$– and $\rhd$–formulas. But C also cannot be of the form $A \between B$, since by the procedure of defining the forcing of $A \between B$, it is either true at all y with $y \in [x]_{\sim_t}$, or at none of them. Hence such a C cannot exist, and $[x]_{\sim_t} = [x]_{\sim_{t+1}}$.

It follows that for any C, there is some $E \in \mathcal{D}$ s.t. $\mathcal{M}^{\mathsf{S}}, w \Vdash \Box(C \leftrightarrow E)$. Given C, let $X := \{w \neq x \in \mathcal{M} \mid x \Vdash C\}$, and let t be the stage where the forcing of C was defined. Note that X is closed under $\sim_t$. By what was shown above, there is some $E \in \mathcal{D}$ with $x \in X \Leftrightarrow x \Vdash E$. But then clearly $\mathcal{M}^{\mathsf{S}}, w \Vdash \Box(C \leftrightarrow E)$. ❑

We will now prove that $\mathcal{M}^{\mathsf{S}} \Vdash \mathsf{S}$. As mentioned above, it suffices to show that $\mathcal{M}^{\mathsf{S}}, w \Vdash \mathsf{S}$. So fix ILMS–formulas A, B and C. We need to show that

$$\mathcal{M}^{\mathsf{S}}, w \Vdash (C \rhd A) \wedge (C \rhd B) \leftrightarrow C \rhd A \between B.$$

Using the remark above, let $E \in \mathcal{D}$ be s.t. $w \Vdash \Box(E \leftrightarrow C)$. It is easy to see that it suffices to show: $\mathcal{M}^{\mathsf{S}}, w \Vdash (E \rhd A) \wedge (E \rhd B) \leftrightarrow E \rhd A \between B$. Let t be the stage where the forcing of $A \between B$ was defined. We will prove the claim by induction on t.

If $t = 0$, then $A \between B \in \mathcal{D}$. By closure properties of $\mathcal{D}$, also $A, B \in \mathcal{D}$. Since also $E \in \mathcal{D}$, we have $\mathcal{M}^{\mathsf{S}}, w \Vdash (E \rhd A) \wedge (E \rhd B) \leftrightarrow E \rhd A \between B$ by the assumption that $\mathcal{M}$ is an ILMS–model restricted to $\mathcal{D}$.

For the inductive step, assume that the value of $A \between B$ was defined at stage $t+1$. In case there were A' and B' with $w \Vdash \Box(A \leftrightarrow A') \wedge (B \leftrightarrow B')$, we defined $w \Vdash (A \between B) \leftrightarrow (A' \between B')$ (according to clause 1 of the process). Since by induction assumption, $\mathcal{M}^{\mathsf{S}}, w \Vdash (E \rhd A') \wedge (E \rhd B') \leftrightarrow E \rhd A' \between B'$, we clearly also have that $\mathcal{M}^{\mathsf{S}}, w \Vdash (E \rhd A) \wedge (E \rhd B) \leftrightarrow E \rhd A \between B$.

If there were no A', B' as above, the forcing of $A \between B$ in the model was defined according to condition (76). We will show the two directions of S separately. Let $w \Vdash (E \rhd A) \wedge (E \rhd B)$ and $wRx \Vdash E$ (with $E \in \mathcal{D}$ as above). Then there are y, y' with $xS_wy \Vdash A$ and $xS_wy' \Vdash B$. We want to show that $A \between B$ was defined to be true at x at stage $t+1$ — since xS_wx, this will be sufficient for our purpose. So let $x \sim_t x'$. Since[18] $E \in \mathcal{D}$, also $x' \Vdash E$. Since $w \Vdash (E \rhd A) \wedge (E \rhd B)$, we have some z, z' with $x'S_wz \Vdash A$ and $x'S_wz' \Vdash B$. Thus x fulfills the condition (76) above, and thus $A \between B$ was indeed defined to be true at x. For the other direction of S, we want to show that $\mathcal{M}^{\mathsf{S}}, w \Vdash (A \between B \rhd A) \wedge (A \between B \rhd B)$. So let $wRx \Vdash A \between B$. Since $A \between B$ was defined to be true at x, x fulfills condition (76), whence there must be y and y' with $xS_wy \Vdash A$ and $xS_wy' \Vdash B$, and we are done. ❑

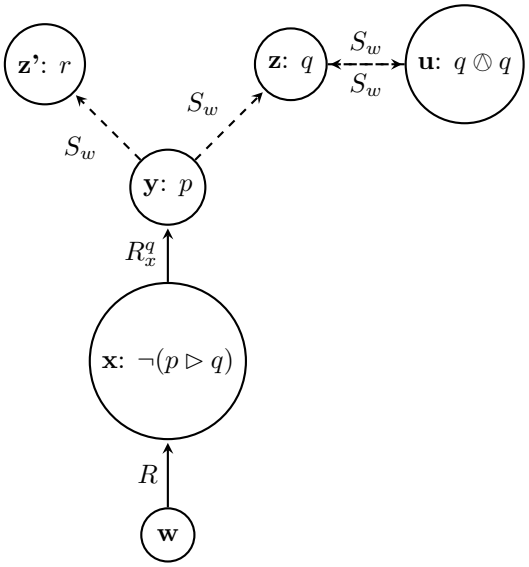
5.2. Models of depth > 2. We will show how the method used in the proof of Lemma 5.1 can fail in case of frames of depth > 2. Consider the following set of sentences:

$$\mathcal{X} = \{p \rhd q, p \rhd r, q \between q \rhd q, \Diamond\neg(p \rhd q)\}$$

Let $\mathcal{D}$ be the smallest set containing $\mathcal{X}$ closed under subformulas and boolean combinations. As before we use the method of Section 3.2 to get an ILMS–model $\mathcal{M}$ restricted to $\mathcal{D}$ where all formulas in $\mathcal{X}$ are true at the root w. Since $q \between q$ is

[18]Note that if E would be a formula whose truth value was defined at some later stage $t+k$, we would not be able to make this inference.

the only $\between$–formula in $\mathcal{D}$, it is the only $\between$–formula whose forcing is defined in $\mathcal{M}$. $\mathcal{M}$ might look as follows[19]:



Consider the formula $q \between r$ whose forcing is initially undefined in the model. If we were to follow the strategy of the previous section in defining its truth value, the only node where we could make it true is y, as it is the only node that has S-access to a q-node and an r-node (assuming that q and r are only true at z and z' respectively). However, supposing that we define $y \Vdash q \between r$, we would need some v with $yS_xv \Vdash q$, if axiom S were to be valid at x (since xRy and $q \between r \rhd q$). However this contradicts the fact that y and all its S_x-successors are in the q-critical cone above x. Hence we cannot make $q \between r$ true at y.

On the other hand, since $w \Vdash (p \rhd q) \wedge (p \rhd r)$ and $wRy \Vdash p$, there has to be some v with $yS_wv \Vdash q \between r$ (if S is to be valid). But in order to have $w \Vdash (q \between r \rhd r) \wedge (q \between r \rhd q)$, this v has to furthermore have the property that vS_wz and vS_wz' (again, assuming that q and r are only true at z and z' respectively). Except for y, there is no such node in the above model.

As a consequence, we cannot have an extension lemma as in the previous, where we just take an existing ILMS–model restricted to $\mathcal{D}$ and, without making any changes in the underlying structure of the model, define the forcing of all $\between$–formulas. Whether an alternative route exists in the general case remains a question for future research.

6. Arithmetical Completeness for a Simple Language

In this section, we give an arithmetical completeness result for ILM w.r.t. a very restricted set of formulas, namely the formulas constructed from a single propositional letter using only $\between$. An important consequence of the result is that all formulas in

[19]We only indicate the forcing of formulas which are important for the example. We write xR_x^qy to indicate that y is in the q-critical cone above x.

this simple language are arithmetically independent. To be more precise, there is an arithmetical implementation under which any two formulas of this language are even mutually inconsistent. This is already enough to establish that the arithmetical suprema are in general not idempotent (w.r.t. provability) or extensional. This section presupposes knowledge of the arithmetical completeness proof of ILM. This proof was found independently by Alessandro Berarducci ([**Ber90**]) and Volodya Shavrukov ([**Sha88**]).

Let $\{A_i\}_{0\leq i\in\omega}$ be an enumeration of all ILMS–formulas constructed from a single propositional letter p using only $\owedge$. We assume that $A_0 = p$. Consider the rooted ILM-frame $\mathcal{F} = \langle W, R, \{S_1\}\rangle$ with $W = \mathbb{N}\setminus\{0\}$, $1Ri$ for all $i > 1$, and iS_1j for all $i, j > 1$. Since S_1 is the only S-relation in $\mathcal{F}$, we will from now on write just iSj instead of iS_1j. We define a forcing relation on $\mathcal{F}$ by letting[20] $x \Vdash A_i :\Leftrightarrow x = i+2$. Thus p is true only at node 2, A_1 is true only at node 3, and so on. Let $\mathcal{M}$ be the resulting model. Using that $\mathcal{M}$ has depth 2, and that all nodes R-above 1 have access to all other nodes R-above 1, it is easy to see that axiom S is valid in $\mathcal{M}$.

By the Berarducci-Shavrukov Arithmetical Completeness Theorem for ILM, $\mathcal{M}$ can be embedded in PA. Let $*$ be the arithmetical realization that we obtain in the proof. Then we have that $p^* = (\ell = 2)$, and in general[21] $A_i^* = (\ell = (i+2))$. I.o.w. the sentence A_i is represented in arithmetic as the sentence $\ell = (i+2)$.

Since iSj for all $i, j > 1$, by the arithmetical completeness proof we get for all i and j that $\vdash_{\mathsf{PA}} \ell = (i+2) \equiv \ell = (j+2)$. This means that all sentences $\ell = i$ for $i > 1$ are in the same interpretability degree. As a consequence, if $A_k = A_i \owedge A_j$, then the sentence $\ell = (k+2)$ (the arithmetical representative of A_k) is in the same degree as the *real* supremum[22] of A_i^* and A_j^*, i.e. of $\ell = (i+2)$ and $\ell = (j+2)$.

Thus we can define an implementation $\owedge$ of the supremum in PA as follows: $B \owedge C := B \curlywedge C$, unless B and C are of the form $\ell = (i+2)$ and $\ell = (j+2)$ respectively for some $i, j \in \mathbb{N}$. In that case, take $B \owedge C$ to be the sentence $\ell = (k+2)$ for the k for which $A_k = A_i \owedge A_j$.

By the properties of the sentences $\ell = i$, we have that $\vdash \ell = (i+2) \to \ell \neq (j+2)$ if $i \neq j$. It follows that if $i \neq j$, then $\nvdash \ell = (i+2) \to \ell = (j+2)$, i.e. $\nvdash A_i^* \to A_j^*$. Thus all formulas in the small language are arithmetically independent. Note that the implementation $\owedge$ defined above is certainly not extensional. For example, if $\vdash A \leftrightarrow (\ell = i)$ but $A \neq (\ell = i)$, and $\vdash B \leftrightarrow (\ell = j)$ but $B \neq (\ell = j)$, then it is clear that not necessarily $\vdash_{\mathsf{PA}} A \owedge B \leftrightarrow (\ell = i) \owedge (\ell = j)$. In order to obtain independence results for extensional implementations, a more elaborate procedure is therefore needed.

---

[20] Note that just as in Section 3.2, the $\owedge$-formulas are treated as atoms.

[21] The sentence $\ell = i + 2$ can be said to "represent" the node $i + 2$ of the model $\mathcal{M}$ in arithmetic. Since A_i is only true at $i + 2$, A_i and $i + 2$ are represented by the same arithmetical sentence.

[22] See the discussion in Section 3.3

CHAPTER 5

A Relational Semantics for $\curlywedge$

As a disclaimer, we admit that the negative result of Theorem 67 has a nullifying effect on the results presented in this chapter. We will introduce a relational semantics for Visser's implementation $\curlywedge$, even though we have already established the existence of such a semantics as an impossibility. There are two reasons why we chose to present the semantics nevertheless. First, it is an interesting system of modal semantics on its own. It can cope with the failure of monotonicity of Visser's implementation, and furthermore the truth condition for $\curlywedge$–formulas is intuitively "in tune" with $\curlywedge$ being a supremum operator. Second, the match between what is valid in the semantics and what is provable in PA about $\curlywedge$ goes a surprisingly long way (although as we will see in the end of this chapter, it does not go *all* the way). In fact, the semantics has been fruitful for discovering new arithmetical principles for $\curlywedge$. The main idea behind the semantics is due to Frank Veltman.

1. Introducing the Semantics

Recall from Section 5 of Chapter 3 Visser's implementation $\curlywedge$ of the supremum in PA. Given sentences A and B of PA, $A \curlywedge B$ was defined as the sentence σ with the property that

$$\vdash_{\mathsf{PA}} \sigma \leftrightarrow \exists x(\Box_x \neg\sigma \wedge (\Diamond_x A \wedge \Diamond_x B)).$$

Alternatively, using the explicit form for $\curlywedge$, $A \curlywedge B$ can be seen as the sentence $\exists x(\Box_x(\Box_x A \vee \Box_x B) \wedge (\Diamond_x A \wedge \Diamond_x B))$. Throughout this chapter, the intended meaning of the modal symbol $\mathbin{\text{⍀}}$ is Visser's implementation $\curlywedge$. To be more precise, by an arithmetical realization for ILMS we will in this chapter mean a realization for ILM, plus the clause[1]:

$$(A \curlywedge B)^* = \exists x(\Box_x(\Box_x A^* \vee \Box_x B^*) \wedge (\Diamond_x A^* \wedge \Diamond_x B^*)).$$

For the sake of clarity, we will write $\curlywedge$ instead of $\mathbin{\text{⍀}}$ for the modal supremum symbol, and denote the resulting system by $\mathsf{ILMS}^{\curlywedge}$. Hence, the system $\mathsf{ILMS}^{\curlywedge}$ contains ILM, plus axiom S for $\curlywedge$:

$$(C \rhd A) \wedge (C \rhd B) \leftrightarrow C \rhd (A \curlywedge B). \tag{77}$$

We have seen in Section 5 of Chapter 3, and will see below, that PA proves more about $\curlywedge$ than just (77). As a consequence, we will have to add more axioms to the system $\mathsf{ILMS}^{\curlywedge}$ if we want to use it for capturing the arithmetical behaviour of $\curlywedge$. We shall not be concerned with the axiomatization of $\mathsf{ILMS}^{\curlywedge}$ here. Instead, we will try to find a relational semantics for this system. The goal is to have the validities of the semantics (w.r.t. $\curlywedge$) match exactly what is provable in PA about

[1]Note how the existence of an explicit form is convenient when defining arithmetical realizations.

$\curlywedge$. We will now introduce our most successful candidate for such a semantics. We define the truth values of $\curlywedge$–formulas by using a new relation Q, which we will add to ILM–frames.

DEFINITION 69. Let $\mathcal{F}$ be an ILM–frame with an additional relation Q. If $\mathcal{M}$ is a model on $\mathcal{F}$, then for all $x \in \mathcal{M}$,

$$x \Vdash A \curlywedge B :\Leftrightarrow$$

1. $\exists y, y'[xQy \Vdash A \wedge xQy' \Vdash B]$
2. $\forall x'[xRx' \Rightarrow \neg\exists u, u'(x'Qu \Vdash A \wedge x'Qu' \Vdash B)]$ ❑

There are two requirements that a node needs to satisfy in order to make true the formula $A \curlywedge B$. First, it needs to Q-see an A-node and a B-node. Second, it cannot have an R-successor that also Q-sees an A-node and a B-node. We will refer to these two requirements as *the first* and *the second truth condition (for $A \curlywedge B$)* respectively. Note that the second truth condition requires x to be R-maximal in the set of nodes that Q-see an A-node and a B-node. Given that it is a supremum operator that we are trying to model, this is a rather appealing feature to have.

DEFINITION 70. An $\mathsf{ILMS}^{\curlywedge}$–frame is a quadruple $\{W, R, \{S_x \mid x \in W\}, Q\}$, where $\{W, R, \{S_x \mid x \in W\}\}$ is an ILM–frame, and

i. if xQy, then xS_wy for every w with wRx

ii. if xS_wy and xS_wy', there is z with xS_wz, zQy and zQy'

iii. $R \subseteq Q$

iv. Q is converse well-founded ❑

Note that i and ii in Definition 70 are the frame-conditional equivalents of the minimal conditions discussed in Section 2.1 of Chapter 4. Since we showed that ILM–frames cannot in general be required to satisfy the second minimal condition, the definition of $\mathsf{ILMS}^{\curlywedge}$–frames (and thereby the whole current approach) is ill-founded. We will suppress this issue while exploring the features of our semantics for $\curlywedge$.

We will first show that axiom S is valid on $\mathsf{ILMS}^{\curlywedge}$–frames.

LEMMA 1.1. *Let $\mathcal{K}^{\curlywedge}$ be the class of $\mathsf{ILMS}^{\curlywedge}$–frames. If $\mathcal{F} \in \mathcal{K}^{\curlywedge}$, then*[2] *$\mathcal{F} \Vdash \mathsf{S}$.*

PROOF. Suppose that $\mathcal{F} \in \mathcal{K}^{\curlywedge}$, $\mathcal{M}$ is a model on $\mathcal{F}$, and $w \in \mathcal{M}$. To see that $w \Vdash (A \curlywedge B \rhd A) \wedge (A \curlywedge B \rhd B)$, let $wRx \Vdash A \curlywedge B$. By the first truth condition for $A \curlywedge B$, there are y, y' with $xQy \Vdash A$ and $xQy' \Vdash B$. By property i of $\mathsf{ILMS}^{\curlywedge}$–frames, also xS_wy and xS_wy'. For the other direction of S, let $w \Vdash (C \rhd A) \wedge (C \rhd B)$, and $wRx \Vdash C$. Then there are y, y' with $xS_wy \Vdash A$ and $xS_wy' \Vdash B$. By property ii of $\mathsf{ILMS}^{\curlywedge}$–frames, there is some z with xS_wz, zQy and zQy'. Then z satisfies the first truth condition for $A \curlywedge B$. If it also satisfies the second one, we are done. If not, then there must be z', u, and u' with zRz', $z'Qu \Vdash A$, and $z'Qu' \Vdash B$. Since R is converse well-founded, we can choose an R-maximal z' with this property. Then

[2]Due to the complexity of the truth condition for $\curlywedge$–formulas, the other direction of frame characterizability cannot be seen to hold so easily (see also footnote 2 in Chapter 4).

clearly z' fulfills both the first and the second minimal condition, so $z' \Vdash A \curlywedge B$. We now have $wRxS_wzRz'$, and hence also xS_wz', which is what we wanted. ❑

We only used properties i and ii of $\mathsf{ILMS}^{\curlywedge}$–frames in the above proof. The purpose of requirements iii and iv will be explained later in this chapter.

In the above proof, verifying $(A \curlywedge B \rhd A) \wedge (A \curlywedge B \rhd B)$ was a simple consequence of property i of $\mathsf{ILMS}^{\curlywedge}$–frames. Verifying the other direction of S is more involved. After assuming $w \Vdash (C \rhd A) \wedge (C \rhd B)$ and $wRx \Vdash C$, we need to find a node z' with $xS_wz' \Vdash A \curlywedge B$. Property ii of $\mathsf{ILMS}^{\curlywedge}$–frames only gives us a node z with xS_wz that satisfies the first truth condition for $A \curlywedge B$. Hence at this point we need to be able to argue as follows: if z satisfies the first truth condition for $A \curlywedge B$, but nevertheless $z \nVdash A \curlywedge B$, then there is some z' with $zS_wz' \Vdash A \curlywedge B$ (since by transitivity of S_w then also xS_wz'). Now it might seem that the requirement of R-maximality in the truth definition for $A \curlywedge B$ is too strong, as it allows us to conclude that even $zRz' \Vdash A \curlywedge B$. So why not require S-maximality in the truth definition for $\curlywedge$–formulas? Note that in order to conclude the existence of z', we had to use the converse well-foundedness of R. Since we cannot in general assume the S-relation to be converse well-founded, this is where the argument would break down were we to require S-maximality instead of R-maximality. Later in this chapter, we will see more evidence for why the relation w.r.t. which we require maximality in the truth definition for $\curlywedge$–formulas should be the R-relation.

2. Coping with Non-Monotonicity

As we saw in Section 5 of Chapter 3, Visser's implementation is not monotone, i.e. we do *not* have for all A and B,

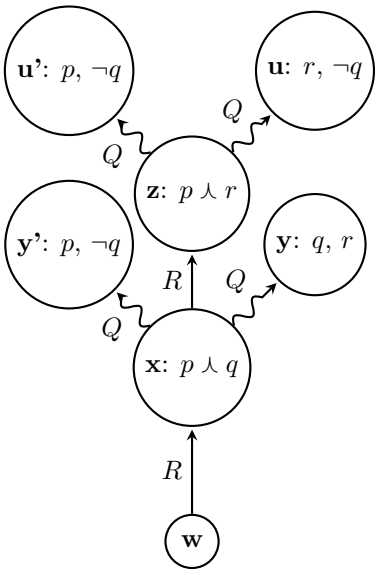
$$\vdash_{\mathsf{PA}} \Box(A \to A') \wedge \Box(B \to B') \to \Box(A \curlywedge B \to A' \curlywedge B').$$

If we want an exact match between what is provable in PA about $\curlywedge$ and what is valid in our semantics, our semantics should also be able to falsify

$$\Box(A \to A') \wedge \Box(B \to B') \to \Box(A \curlywedge B \to A' \curlywedge B'). \tag{78}$$

Our intricate truth condition for $A \curlywedge B$ allows us to find a countermodel to (78). Unlike when proving Lemma 1.1, in this countermodel we make essential use of the second truth condition for $\curlywedge$.

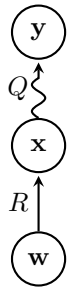
The model below falsifies $\Box(q \to r) \to \Box((p \curlywedge q) \to (p \curlywedge r))$. In particular, $w \Vdash \Box(q \to r)$, but $w \Vdash \Diamond((p \curlywedge q) \wedge \neg(p \curlywedge r))$, since $x \Vdash (p \curlywedge q) \wedge \neg(p \curlywedge r)$. We have $x \Vdash (p \curlywedge q)$, as x has Q-access to a p-node (i.e. y') and a q-node (i.e. y), and the only R-successor of x does not have Q-access to a q-node. The assumption that $w \Vdash \Box(q \to r)$ only implies that if $x \Vdash p \curlywedge q$, then x satisfies the first truth condition for $p \curlywedge r$ (as the q-node that it Q-sees is also an r-node). However, it does not satisfy the second truth condition for $p \curlywedge r$. This is because xRz, and z has Q-access to a p-node and an r-node.



The counterexample to the provability of (78) in PA was that we had:

$$\nvdash_{\mathsf{PA}} (\top \curlywedge \top) \curlywedge (\top \curlywedge \top) \to \top \curlywedge \top$$

Thus we should also be able to find a counterexample to this on our frames. In the model below, there is no z with yQz, whence $y \Vdash \neg(A \curlywedge B)$ for all A, B. Since x has no R-successors, the second truth condition for $\top \curlywedge \top$ cannot fail at x, whence $x \Vdash \top \curlywedge \top$. Consequently, $w \vDash \neg(\top \curlywedge \top)$ (since wRx). However, $w \Vdash (\top \curlywedge \top) \curlywedge (\top \curlywedge \top)$, since $wQx \Vdash \top \curlywedge \top$ (here we use that $R \subseteq Q$), and there is no z with $xQz \Vdash \top \curlywedge \top$ (since the only node Q-seen by x is y).



3. Other Properties of ILMS$^{\curlywedge}$–Frames

We will now explain why the ILMS$^{\curlywedge}$–frames were assumed to have properties iii and iv (in Definition 70). For this, we also have to go back to the arithmetical side and see what is provable about Visser's implementation in PA.

An immediate consequence of our truth definition for $\curlywedge$ is that our semantics validates $A \curlywedge B \to \Box\neg(A \curlywedge B)$ for all A and B. Fortunately, this is also provable in PA.

Fact 71. $\vdash_{\mathsf{PA}} A \curlywedge B \to \Box\neg(A \curlywedge B)$.

PROOF. Using the fixed point version of $\curlywedge$, $\neg(A \curlywedge B)$ is the sentence θ with the property that $\vdash_{\mathsf{PA}} \theta \leftrightarrow \forall x(\Box_x\theta \to \Box_x\neg A \vee \Box_x\neg B)$. We will show by contraposition $\vdash_{\mathsf{PA}} \Diamond(A \curlywedge B) \to \neg(A \curlywedge B)$, i.e. that $\vdash \neg\Box\theta \to \theta$. So suppose that $\Diamond(A \curlywedge B)$, i.e. $\neg\Box\theta$, where θ is as above. By monotonicity (in our chosen stratification of provability, see Section 2 of Chapter 3), we have $\forall x\neg\Box_x\theta$, whence by propositional logic $\forall x(\Box_x\theta \to (\Box_x\neg A \vee \Box_x\neg B))$, i.e. θ. ❑

Remember from Section 1 Chapter 4 that $\vdash_{\mathsf{ILMS}} \Box\neg(A \curlywedge B) \to \Box(\Box\neg A \vee \Box\neg B)$. Combining this with Fact 71, we see that our semantics should validate for all A and B, $A \curlywedge B \to \Box(\Box\neg A \vee \Box\neg B)$. It is easy to see that for this to hold, we have to require that $R \subseteq Q$.

We now turn to property iv of $\mathsf{ILMS}^{\curlywedge}$–frames: the converse well-foundedness of Q. In GL, the need for the converse well-foundedness of R is due to the validity of Löb's Theorem in arithmetic. Similarly, the need for the converse well-foundedness of Q on $\mathsf{ILMS}^{\curlywedge}$–frames is triggered by a "Löb-like" principle that holds for $\curlywedge$ in PA. According to Löb's Theorem, if $\vdash_{\mathsf{PA}} \Box A \to A$, then already $\vdash_{\mathsf{PA}} A$. The Löb-like principle for $\curlywedge$ is: if $\vdash_{\mathsf{PA}} A \to A \curlywedge A$, then $\vdash_{\mathsf{PA}} \neg A$. Like Löb's Theorem, this can be formalized inside PA.

FACT 72. $\vdash_{\mathsf{PA}} \Box(A \to A \curlywedge A) \to \Box\neg A$.

PROOF. We will use the explicit version of Visser's implementation to show something stronger: $\vdash_{\mathsf{PA}} \Box_x(A \to A \curlywedge A) \to \Box_x\neg A$.

$$
\begin{aligned}
&\vdash_{\mathsf{PA}} \Box_x(A \to A \curlywedge A)\\
&\quad\to \Box_x(A \to \exists y(\Box_y\Box_y\neg A \wedge \Diamond_y A)) && \text{(def. of } \curlywedge)\\
&\quad\to \Box_x(A \to \exists y \geq x(\Box_y\Box_y\neg A \wedge \Diamond_y A)) && \text{(reflection)}\\
&\quad\to \Box_x(A \to \exists y \geq x\Diamond_y A)\\
&\quad\to \Box_x(\forall y \geq x\Box_y\neg A \to \neg A)\\
&\quad\to \Box_x(\Box_x\neg A \to \neg A) && \text{(monotonicity)}\\
&\quad\to \Box_x\neg A && \text{(Löb's Theorem)}
\end{aligned}
$$ ❑

By contraposition, $\vdash \Diamond A \to \Diamond(A \wedge \neg(A \curlywedge A))$. Note that since $\vdash_{\mathsf{ILMS}^{\curlywedge}} A \rhd A \curlywedge A$, also $\vdash_{\mathsf{ILMS}^{\curlywedge}} \Diamond A \to \Diamond(A \curlywedge A)$ by axiom J4 of ILM. Thus if we add Fact 72 as an axiom to $\mathsf{ILMS}^{\curlywedge}$, then

$$\vdash_{\mathsf{ILMS}^{\curlywedge}} \Diamond A \to \Diamond(A \curlywedge A) \wedge \Diamond\neg(A \curlywedge A). \tag{79}$$

From the point of view of modal semantics, this means that every point with an R-successor has at least two R-successors. From the point of view of arithmetic, (79) is in fact a good property for a provability logic to have. The sentence $\Diamond(A \curlywedge A) \wedge \Diamond\neg(A \curlywedge A)$ is the formalization of the assertion $A \curlywedge A$ is independent from PA. Substituting $\top$ for A, we get that

$$\vdash_{\mathsf{ILMS}^{\curlywedge}} \neg\Box\bot \to \neg\Box(\top \curlywedge \top) \wedge \neg\Box\neg(\top \curlywedge \top). \tag{80}$$

Remember that $\top \curlywedge \top$ is an Orey sentence, and as such independent from PA. Thus the sentence in (80) just expresses that if PA is consistent, then the Orey-sentence $\top \curlywedge \top$ is independent from PA. The fact that this is provable in $\mathsf{ILMS}^{\curlywedge}$ signifies an increase in expressive power over ILM. There is no sentence B of ILM

s.t. $\vdash_{\mathsf{ILM}} \neg\Box\bot \to \neg\Box B \wedge \neg\Box\neg B$, as can be easily checked by using the soundness of ILM w.r.t. its modal semantics.

We will now show that if Q is converse well-founded, then the Löb-like principle for $\curlywedge$ is guaranteed to be valid.

LEMMA 3.1. . *If $\mathcal{F} \in \mathcal{K}^{\curlywedge}$, then $\mathcal{F} \Vdash \Diamond A \to \Diamond(A \wedge \neg(A \curlywedge A))$. (Where $\mathcal{K}^{\curlywedge}$ is the class of $\mathsf{ILMS}^{\curlywedge}$–frames as above).*

PROOF. Let $\mathcal{F} \in \mathcal{K}^{\curlywedge}$, and let $\mathcal{M}$ be a model on $\mathcal{F}$ with $\mathcal{M}, w \Vdash \Diamond A$. Then there is some x with $wRx \Vdash A$. Consider the set[3] $\mathcal{X} := \{y \mid xQ^{\mathsf{tr}}y \Vdash A\}$. If $\mathcal{X} = \emptyset$, then there is no y with $xQy \Vdash A$, whence certainly $x \Vdash \neg(A \curlywedge A)$, and $w \Vdash \Diamond(A \wedge \neg(A \curlywedge A))$. If $\mathcal{X} \neq \emptyset$, then by the converse well-foundedness of Q, $\mathcal{X}$ has a Q-maximal element y. Since $y \in \mathcal{X}$, we have that $xQ^{\mathsf{tr}}y$, whence also xS_wy by properties of Q and transitivity of S_w. But then also wRy. Thus in order to show that $w \Vdash \Diamond(A \wedge \neg(A \curlywedge A))$, it suffices to show that $y \Vdash A \wedge \neg(A \curlywedge A)$. We have $y \Vdash A$ by the fact that $y \in \mathcal{X}$. Now if yQy', then $y' \notin \mathcal{X}$ by Q-maximality of y. But clearly $xQ^{\mathsf{tr}}y'$, hence the reason that y' is not in $\mathcal{X}$ must be that $y' \nVdash A$, whence y does not fulfill the first truth condition for $A \curlywedge A$, so certainly $y \Vdash \neg(A \curlywedge A)$. ❑

We will now discuss some further properties of our semantics. It is easy to check that our semantics validates the extensionality of $\curlywedge$, i.e. that $\mathsf{ILMS}^{\curlywedge}$–frames validate

$$\Box(A \leftrightarrow A') \wedge \Box(B \leftrightarrow B') \to \Box(A \curlywedge B \leftrightarrow A' \curlywedge B').$$

Recall from Section 5.4 of Chapter 3 that two directions of (almost) distributivity were provable in PA for Visser's supremum $\curlywedge$:

1. $C \curlywedge (A \vee B) \to (C \curlywedge A) \vee (C \curlywedge B)$
2. $(C \vee A) \curlywedge (C \vee B) \to (C \curlywedge C) \vee (A \curlywedge B)$

As it turns out, these are exactly the directions of distributivity which are valid on $\mathsf{ILMS}^{\curlywedge}$–frames. We will show this for 1. The proof for the validity of 2. is similar. Let $w \Vdash C \curlywedge (A \vee B)$. By the first truth condition, there are y, y' with $wQy \Vdash C$ and $wQy' \Vdash A \vee B$. Suppose w.l.o.g. that $y' \Vdash A$. Then w fulfills the first truth condition for $C \curlywedge A$. To see that it also fulfills the second, suppose that there are x, u and u' with wRx, $xQu \Vdash C$ and $xQu' \Vdash A$. But then x fulfills the first truth condition for $C \curlywedge (A \vee B)$, contradicting that $C \curlywedge (A \vee B)$ was true at w. To see that the other direction fails, we need to show that (w.l.o.g.) $C \curlywedge A \to (C \curlywedge (A \vee B))$ is not valid in the semantics. Note that $C \curlywedge A \to (C \curlywedge (A \vee B))$ would be valid in the semantics if the semantics could not cope with the failure of monotonicity (since $A \to A \vee B$). In fact, the countermodel to $C \curlywedge A \to (C \curlywedge (A \vee B))$ is almost the same as the countermodel to monotonicity of $\curlywedge$ in Section 2.

Combining Fact 71 with 2 above, we get $(C \vee A) \curlywedge (C \vee B) \to \Box\neg(C \curlywedge C) \vee \Box\neg(A \curlywedge B)$. However, as we will first see in arithmetic, something stronger holds:

FACT 73. $\vdash (C \vee A) \curlywedge (C \vee B) \to \Box\neg(C \curlywedge C) \wedge \Box\neg(A \curlywedge B)$

[3] By Q^{tr}, we denote the transitive closure of Q.

PROOF.

$$\begin{aligned}
&\vdash (C \vee A) \curlywedge (C \vee B)\\
&\quad \to \exists x(\Box_x(\Box_x(\neg C \wedge \neg A) \vee \Box_x(\neg C \wedge \neg B)) \wedge\\
&\qquad (\Diamond_x(C \vee A) \wedge \Diamond_x(C \vee B))) && \text{(def. of } \curlywedge)\\
&\quad \to \exists x \Box_x(\Box_x(\neg C \wedge \neg A) \vee \Box_x(\neg C \wedge \neg B)) && \text{(propositional logic)}\\
&\quad \to \exists x \Box_x((\Box_x \neg C \wedge \Box_x \neg A) \vee (\Box_x \neg C \wedge \Box_x \neg B)) && \text{(distributivity of } \Box \text{ over } \wedge)\\
&\quad \to \exists x \Box_x(\Box_x \neg C \wedge (\Box_x \neg A \vee \Box_x \neg B)) && \text{(distributivity of } \vee \text{ over } \wedge)\\
&\quad \to \exists x(\Box_x \Box_x \neg C \wedge \Box_x(\Box_x \neg A \vee \Box_x \neg B)) && \text{(distributivity of } \Box \text{ over } \wedge)\\
&\quad \to \exists x \Box_x \Box_x \neg C \wedge \exists x \Box_x(\Box_x \neg A \vee \Box_x \neg B) && \text{(predicate logic)}
\end{aligned}$$

According to Theorem 53 we have for all x, and for all A and B (using the explicit form of $\curlywedge$):

$$\vdash_{\mathsf{PA}} \Box_x \neg(A \curlywedge B) \leftrightarrow \Box_x(\Box_x \neg A \vee \Box_x \neg B).$$

Hence $\vdash_{\mathsf{PA}} \Box_x \Box_x \neg C \to \Box_x \neg(C \curlywedge C)$, and $\vdash_{\mathsf{PA}} \Box_x(\Box_x \neg A \vee \Box_x \neg B) \to \Box_x \neg(A \curlywedge B)$, Now $\vdash_{\mathsf{PA}} \Box_x \neg(C \curlywedge C) \to \Box \neg(C \curlywedge C)$ and $\vdash_{\mathsf{PA}} \Box_x \neg(A \curlywedge B) \to \Box \neg(A \curlywedge B)$. ❑

It is easy that Fact 73 is valid on ILMS$^{\curlywedge}$–frames. If $x \Vdash \Diamond(A \curlywedge B)$, then x cannot satisfy the second truth condition for $(C \vee A) \curlywedge (C \vee B)$, so certainly $x \nVdash (C \vee A) \curlywedge (C \vee B)$.

According to Theorem 37, PA does not in general verify that the supremum of two consistent theories is consistent, i.e. we do *not* have for all A, B:

$$\vdash_{\mathsf{PA}} \Diamond A \wedge \Diamond B \to \Diamond(A \curlywedge B).$$

The following fact can be seen as PA's way to approximate this unprovable truth.

FACT 74. $\vdash_{\mathsf{PA}} \Diamond A \wedge \Diamond B \to (A \curlywedge B) \vee \Diamond(A \curlywedge B)$.

PROOF. We argue in PA. Assume $\Diamond A \wedge \Diamond B$ and $\neg(A \curlywedge B)$, i.e. (using the fixed point version of $\curlywedge$) the sentence θ with $\theta \leftrightarrow \forall x(\Box_x \theta \to \Box_x \neg A \vee \Box_x \neg B)$. Suppose for contradiction that $\Box\theta$ (i.e. $\Box \neg(A \curlywedge B)$). Then there is some x with $\Box_x \theta$. By the assumption that θ and the properties of θ, we get $\Box_x \neg A \vee \Box_x \neg B$, i.e. $\Box \neg A \vee \Box \neg B$, contradicting the assumption that $\Diamond A \wedge \Diamond B$. ❑

To see that Fact 74 is valid on ILMS$^{\curlywedge}$–frames, note that if $w \Vdash \Diamond A \wedge \Diamond B$, then w satisfies the first truth condition for $A \owedge B$. For this, we need that $R \subseteq Q$. So if $w \Vdash \neg(A \curlywedge B)$, then it must be that w does not satisfy the second truth condition for $A \curlywedge B$. But then there are x, u and, u' with wRx, $xQy \Vdash A$, and $xQy' \Vdash B$. Since R is converse well-founded, we can take the maximal x with this property. But then $x \Vdash A \curlywedge B$, whence $w \Vdash \Diamond(A \curlywedge B)$. Note that in order to verify this principle, we really have to consider R-maximal worlds in the truth definition for $\curlywedge$–formulas.

The following fact is an immediate consequence of Fact 74 (by contraposition, taking $\top$ for A and B), thus it holds both in PA and in the semantics.

FACT 75. $\neg(\top \curlywedge \top) \wedge \Box \neg(\top \curlywedge \top) \to \Box\bot$. *In particular, for any A and B, we have* $\neg(\top \curlywedge \top) \wedge \Box \neg(\top \curlywedge \top) \to \Box \neg(A \curlywedge B)$.

As a consequence of our truth definition for $\curlywedge$, if some node x satisfies $\top \curlywedge \top$, and y is an R-successor of x, then all $\curlywedge$–formulas have to be false at y. Otherwise, x would not satisfy the second truth condition for $\top \curlywedge \top$. In modal terms, this becomes $\top \curlywedge \top \to \Box\neg(A \curlywedge B)$. Indeed, this turns out to be verifiable in PA.

FACT 76. $\vdash_{\mathsf{PA}} \top \curlywedge \top \to \Box\neg(A \curlywedge B)$ *for any* A, B.

PROOF. We argue in PA. Assume $\top \curlywedge \top$, i.e. $\exists x(\Box_x\Box_x\bot \wedge \Diamond_x\top)$. In particular $\exists x\Box_x\Box_x\bot$. We will show that $\Box_x\neg(A \curlywedge B)$, i.e. (using the explicit form of $\curlywedge$)

$$\Box_x\forall y(\Box_y(\Box_y\neg A \vee \Box_y\neg B) \to (\Box_y\neg A \vee \Box_y\neg B)).$$

Since $\Box_x\Box_x\bot$, by monotonicity $\Box_x\forall y \geq x\Box_y\bot$, whence $\Box_x\forall y \geq x(\Box_y\neg A \vee \Box_x\neg B)$. By propositional logic, we get

$$\Box_x\forall y \geq x(\Box_y(\Box_y\neg A \vee \Box_y\neg B) \to (\Box_y\neg A \vee \Box_y\neg B)). \tag{81}$$

By reflection,

$$\Box_x\forall y < x(\Box_y(\Box_y\neg A \vee \Box_y\neg B) \to (\Box_y\neg A \vee \Box_y\neg B)). \tag{82}$$

Combining (81) and (82), we get

$$\Box_x\forall y(\Box_y(\Box_y\neg A \vee \Box_y\neg B) \to (\Box_y\neg A \vee \Box_y\neg B)),$$

which is what we wanted to show. ❑

4. A Problem

As we saw, the validity of $\neg(\top \curlywedge \top) \wedge \Box\neg(\top \curlywedge \top) \to \Box\neg(A \curlywedge B)$ in $\mathsf{ILMS}^{\curlywedge}$–frames matches what is provable in PA about $\curlywedge$. But the antecedent of this implication has another important consequence on $\mathsf{ILMS}^{\curlywedge}$–frames. If $x \Vdash \neg(\top \curlywedge \top)$ and $x \Vdash \Box\neg(\top \curlywedge \top)$, then the reason why x does not make $\top \curlywedge \top$ true must be that it does not satisfy the *first* truth condition for $\top \curlywedge \top$. Since if x satisfies the first truth condition for $\top \curlywedge \top$ but nevertheless $x \nVdash \top \curlywedge \top$, then it must be that $x \Vdash \Diamond(\top \curlywedge \top)$. But if x does not satisfy the first truth condition for $\top \curlywedge \top$, then it clearly cannot satisfy the first truth condition for *any* formula $A \curlywedge B$. Thus

$$\neg(\top \curlywedge \top) \wedge \Box\neg(\top \curlywedge \top) \to \neg(A \curlywedge B) \tag{83}$$

is valid on $\mathsf{ILMS}^{\curlywedge}$–frames. Unfortunately, the arithmetic counterpart of (83) cannot be verified in PA. According to Theorem 61,

$$\nvdash_{\mathsf{PA}} \neg(\top \curlywedge \top) \wedge \Box\neg(\top \curlywedge \top) \to \neg((\top \curlywedge \top) \curlywedge (\top \curlywedge \top)). \tag{84}$$

Trying to change the semantics so that it would be able to falsify (83) gets us into a vicious circle. The reasoning that lead us to conclude the validity of (83) on $\mathsf{ILMS}^{\curlywedge}$–frames (i.e. the reasoning which we would want to block) also allowed us to conclude the validity of some of the above principles. In particular, in order to establish the validity of Fact 74 we had to argue as follows: if x satisfies the first truth condition for $A \curlywedge B$ but nevertheless $x \vDash \neg(A \curlywedge B)$, then $x \Vdash \Diamond(A \curlywedge B)$. However if (83) is to be false at a node x, then since $x \Vdash A \curlywedge B$, x will satisfy the first truth condition for $\top \curlywedge \top$, however we are not allowed to conclude that $x \Vdash \Diamond(\top \curlywedge \top)$. Even if our attempts at fixing the situation would succeed, by the results of the previous chapter a relational semantics for $\curlywedge$ is impossible in any case, thus we will better leave this castle in the air at this point.

CHAPTER 6

Conclusions and Future Research

To conclude, we will summarize what has been achieved in this thesis, and point out some open questions as well as directions for future research.

1. Summary

In this thesis, we investigated the supremum in the lattice of degrees of finite extensions of PA, i.e. the lattice $(\mathrm{V}_{\mathsf{PA}}, \rhd)$. Our goal was to extend the system ILM — the interpretability logic of PA — with a new modality $\between$ whose intended arithmetical meaning is the supremum in $(\mathrm{V}_{\mathsf{PA}}, \rhd)$, and find a modal semantics for the resulting system ILMS. The system ILMS contains ILM plus the defining axiom for $\between$: $(C \rhd A) \wedge (C \rhd B) \leftrightarrow C \rhd A \between B$.

We studied the supremum from the arithmetical (Chapter 3) as well as from the modal perspective (Chapters 4 an Chapter 5). Our research did not yield a straightforward fulfillment of the original goal. Instead, the most important result of this thesis is negative in nature: there is no structural characterization of ILM–frames which satisfy the defining axiom for $\between$. We only proved modal completeness of the logic ILMS w.r.t. a quite modest notion of semantics, and for a very simple case. As for the arithmetical side of the supremum, we have established that the notion of an arithmetical realization for ILMS is more involved than that for ILM or for GL. Most notably, the properties of the logic ILMS depend on the implementation, i.e. the intended arithmetical meaning of $\between$. Two such implementations — Švejdar's implementation and Visser's implementation — were studied in Chapter 3. We established that the arithmetical language needed to express these implementations is significantly more complex than the arithmetical language corresponding to the language of GL.

To conclude, our contribution is not one that involves clear and expected solutions to well-defined problems — in fact, several results in this thesis have been unexpected, and even surprising. Instead, we have pointed out important methodological considerations, and narrowed down the horizon of possibilities concerning the system ILMS. With this, we hope to have prepared the ground for future investigations of the supremum.

2. Questions for Future Research

This section lists a number of unanswered questions that we touched upon throughout the thesis, as well as possible directions for future research.

2.1. Arithmetic.

1. Let $T \supseteq \mathrm{I}\Delta_0 + \mathrm{SUPEXP}$ be a finitely axiomatizable theory. Let A and B be sentences in the language of T. Is there a sentence $A \owedge B$ in the language of T with the property that for all C,

$$\vdash_T (C \rhd_T A) \wedge (C \rhd_T B) \leftrightarrow C \rhd_T A \owedge B?$$

   A positive answer to this question brings us in a good position for trying to extend the logic ILP with a supremum operator. See the discussion in Section 3.4 of Chapter 2. This question was listed as an open question already by Švejdar in [**Šve78**].

2. Devise a modal logic for Friedman's density argument for the lattice $(\mathrm{F}, \rhd)$. As explained in Section 3 of Chapter 1, this question was the starting point for the research presented in this thesis. A positive answer to question 1 would be helpful for this purpose. If a positive answer can not be found, then devising a logic for Friedman's density argument would require a conceptually novel approach. For a start, we would have to figure out how the procedure of making the languages disjoint could be represented in a modal setting. Again, see the discussion in Section 3.4 of Chapter 2.

3. Does Švejdar's implementation have an explicit form? As pointed out in Section 4.2 of Chapter 3, the strategy of using the fixed point algorithm for GL to find an explicit fixed point does not seem to work for Švejdar's implementation.

4. Is Švejdar's implementation $\sqcap$ distributive w.r.t. provability? Also, is it associative or idempotent w.r.t. provability? Some of these properties were discussed in Section 4.2 of Chapter 3. The same questions can be asked for Visser's implementation $\curlywedge$ (we do have at least one of the directions of distributivity for $\curlywedge$). In general, it seems to be rather difficult to prove properties of the implementations w.r.t. provability.

5. Is there an analogue of the "puzzling result" of Section 4.3 for Visser's implementation? We were not able to show in general that if a sentence σ is equivalent to the fixed point version of $\curlywedge$, then it is itself a fixed point of the fixed point version of $\curlywedge$. We can only prove something weaker (see Theorem 55). However, we do not have a concrete counterexample, as we do in the case of Švejdar's implementation.

6. Can there be an implementation of the supremum in PA that is monotone? According to Corollary 62, Visser's implementation is not monotone. We feel that it is unlikely that Švejdar's implementation is monotone, although we do not have a counterexample. Thus, maybe it is in the nature of implementations not to be monotone?

7. A possible way to study fixed point phenomena concerning our implementations (i.e. Švejdar's implementation and Visser's implementation) would be to use modal logic. This approach has been successful for Rosser sentences (see [**GS79**], and also [**Smo85**]). In order to express our implementations, we would need to have a modal language with $\Box$, $\triangle$, and $\forall$, where $\forall$ means $\forall x$, $\triangle$ means $\Box_x$, and x is a fixed variable. As usual, $\exists A$ would be an abbreviation for $\neg\forall\neg A$, and $\triangledown A$ for $\neg\triangle\neg A$. Possibly, the language should be even richer. Using the Orey-Hájek

characterization, we can define interpretability in this language as

$$A \rhd B :\Leftrightarrow \forall \Box(A \to \triangledown B).$$

Visser's implementation could be defined as

$$\exists\, [\triangle(\triangle\neg A \vee \triangle\neg B) \wedge (\triangledown A \wedge \triangledown B)].$$

This language has already been investigated by Franco Montagna in [**Mon87**]. We would want to answer the following questions concerning this language:

a) When do fixed point equations have an explicit fixed point? Can the explicit fixed point always be calculated according to the GL-algorithm?

b) Under which conditions are fixed points unique and extensional?

c) When do we have that if σ is equivalent to a fixed point of $A(Y)$, then σ is itself a fixed point of $A(Y)$?

2.2. Modal logic.

1. Is ILMS complete w.r.t. the modest modal semantics? In Section 5.1 of Chapter 4, we proved modal completeness of ILMS w.r.t. the modest modal semantics for a very simple case.
2. According to Theorem 67, there cannot be a relational semantics for the system ILMS that would extend the usual semantics for ILM. But could there be some other notion of relational semantics for ILMS, i.e. one that would not extend the usual semantics for ILM?
3. As explained in Section 3.3 of Chapter 4, the arithmetical completeness of ILM gives us a restricted version of arithmetical completeness for ILMS. If $A \owedge B$ has a truth value in an ILM–model, we obtain an arithmetical representative of the sentence $A \owedge B$. We saw that it some cases, this sentence is the real supremum of (the arithmetical representatives of) A and B. But is this always the case?
4. While the semantics presented in Chapter 5 has a good match with Visser's implementation, the match is not perfect (see Section 4 of Chapter 5). Is there a way to modify Visser's implementation, or the semantics, in order to overcome this imperfection?
5. Is ILMS really the *minimal* logic for the supremum? It is possible that there is some property which is provable in PA for all implementations, but is still not a theorem of ILMS. In this case, ILMS would only be minimal in the modal, not in the arithmetical sense.

APPENDIX A

Modal Completeness of ILM by the Construction Method

We will prove modal completeness of the system ILM. We will follow the proof by construction method, as presented in [**GJ10**]. For a short overview of the proof, see Section 3.1 of Chapter 4.

1. The System ILM (Remainder)

Recall from Section 2.5 of Chapter 2 that the logic ILM is IL plus Montagna's principle M: $A \rhd B \to A \wedge \Box C \rhd B \wedge \Box C$. For convenience, we will repeat here the definition of an ILM-frame.

Definition 77. An ILM-*frame* is a tuple $\langle W, R, S\rangle$, where W is a non-empty countable set of nodes, R is a binary relation on W, and S a set of binary relations on W, indexed by the elements of W. The R and S relations satisfy the following requirements:

1. R is converse well-founded
2. $xRyRz \Rightarrow xRz$
3. $yS_xz \Rightarrow xRy$ and xRz
4. $xRy \Rightarrow yS_xy$
5. $xRyRz \Rightarrow yS_xz$
6. $xS_wyS_wz \Rightarrow xS_wz$
7. $xS_wyRz \Rightarrow xRz$

❑

Remember that item 7 corresponds to axiom M. An ILM–model is a model whose underlying frame is an ILM–frame.

Definition 78. A set Γ is ILM–*consistent* if $\Gamma \nvdash_{\mathsf{ILM}} \bot$. An ILM–consistent set Γ is *maximal*–ILM–*consistent* if for all A, either $A \in \Gamma$ or $\neg A \in \Gamma$. ❑

From now on, we will often just say "consistent" and "maximal consistent" instead of "ILM-consistent" and "maximal-ILM-consistent" respectively.

Lemma 1.1. *Every consistent set can be extended to a maximal consistent one.*

Proof. This is just Lindenbaum's Lemma for ILM. ❑

2. Modal Completeness: Introduction

Let $\mathcal{K}$ be the class of ILM–frames. As usual in completeness proofs, the modal completeness of ILM is proved by contraposition. Given a sentence A s.t. $\nvdash_{\mathsf{ILM}} A$, we will find an ILM-model $\mathcal{M}$ with $\mathcal{M} \nVdash A$.

The general idea of the proof is as follows. If $\nvdash_{\mathsf{ILM}} A$, there is a maximal–ILM–consistent set Γ with $\neg A \in \Gamma$. We will build an ILM-frame $\mathcal{F}$, where every node x is labeled with a maximal–ILM–consistent set $\nu(x)$. We start with a node w and $\nu(w) = \Gamma$; w will be the root of $\mathcal{F}$. $\mathcal{F}$ will be built step by step, using the information contained in the maximal consistent sets labeling the nodes. Finally, $\mathcal{F}$ will be transformed into a model $\mathcal{M}$ by defining a valuation on $\mathcal{F}$ as: $\mathcal{M}, x \Vdash p :\Leftrightarrow p \in \nu(x)$.

We want to build $\mathcal{F}$ in such a way that the harmony between truth in $\mathcal{M}$ and membership in the maximal consistent sets labeling the nodes extends to a larger set than just the propositional formulas. In particular, we want to be able to conclude that $\mathcal{M}, w \Vdash \neg A$ on the basis that $\neg A \in \nu(w)$.

Definition 79. A set $\mathcal{D}$ of formulas is *adequate* if it is finite[1], and closed under subformulas and single negations. ❑

Let $\mathcal{D}$ be an adequate set containing A. If $\mathcal{M}$ is defined as above, our goal is to have

$$\mathcal{M}, x \Vdash B \Leftrightarrow B \in \nu(x). \tag{85}$$

We call the equivalence in (85) a *truth lemma w.r.t.* $\mathcal{D}$. When constructing $\mathcal{F}$, we thus need to ensure:

1. a truth lemma holds in $\mathcal{F}$ w.r.t. $\mathcal{D}$

2. $\mathcal{F}$ is an ILM–frame

$\mathcal{F}$ will be constructed as the union of an infinite chain of ILM–frames $\{F_n\}_{n\in\omega}$. With each F_n, we come closer to the truth lemma (w.r.t. $\mathcal{D}$). In order to make sure that each F_n is an ILM–frame, F_n itself is also constructed as the union of an infinite chain $\{G_i\}_{i\in\omega}$, where each G_i is closer to being an ILM–frame than the previous one.

3. Preparing the Construction

This section introduces the tools we will need for our construction. In modal completeness proofs, the nodes of the countermodel are often taken to be maximal consistent sets. However, in the context of interpretability logic it is sometimes necessary to use the same maximal consistent set in different places of the model. Therefore we will not identify a node x with a maximal consistent set, but rather label it with a maximal consistent set $\nu(x)$. We will also label some R transitions with formulas: if xRy and $\nu(\langle x, y\rangle) = B$, then y leads into a B-critical cone above x (the notion of a B-critical cone will be defined below).

[1]It is not necessary for $\mathcal{D}$ to be actually finite; it is also sufficient if it contains only finitely many formulas up to provable equivalence. See the discussion in Section 3.1 of Chapter 2.

DEFINITION 80. A *labeled frame* is a quadruple $\langle W, R, S, \nu\rangle$. Here W is a non-empty set of worlds, R a binary relation on W, and S a set of binary relations on W indexed by elements of W. The function ν assigns to each $x \in W$ a maximal ILM–consistent set of sentences $\nu(x)$. To some pairs $\langle x, y\rangle$ with xRy, ν assigns a formula $\nu(\langle x, y\rangle)$. ❑

If $\nu(\langle x, y\rangle) = B$, we will write xR^By. Thus an R^B transition is just an R transition labeled by the formula B. Note that a labeled frame F does not have to be an ILM–frame, or even an IL–frame.

When defining relations R and S between nodes x and y of the frame, we want the maximal consistent sets $\nu(x)$ and $\nu(y)$ to be related in a coherent way. For this purpose, we will define the following relations between maximal consistent.

DEFINITION 81. Let Γ and Δ be maximal ILM–consistent sets.

i. $\Gamma \prec \Delta := \Box A \in \Delta \Rightarrow A, \Box A \in \Delta$, and there is some $E \in \mathcal{D}$ s.t. $\Box E \in \Delta\backslash\Gamma$

ii. $\Gamma \prec_B \Delta := A \rhd B \in \Gamma \Rightarrow \neg A\, \Box\neg A \in \Delta$ and there is some $E \in \mathcal{D}$ s.t. $\Box E \in \Delta\backslash\Gamma$ ❑

iii. $\Gamma \subseteq_\Box \Delta := \Box A \in \Gamma \Rightarrow \Box A \in \Delta$

Note that if $\Gamma \prec_B \Delta$, then also $\Gamma \prec \Delta$. Furthermore, $\Gamma \prec_B \Delta \prec \Delta'$ implies $\Gamma \prec_B \Delta'$. We will refer to $\prec$ as the *successor* relation, and to $\prec_B$ as the *B-critical successor* relation.

The following definition helps us to enforce the truth lemma for formulas of the form $\neg(A \rhd B)$. If $x \Vdash \neg(A \rhd B)$, there has to be some y s.t. $xRy \Vdash A$ and for all z, if yS_xz, then $z \nVdash B$. The B-critical cone above x contains all nodes which are S_x-accessible from y. All of them have to be B-critical successors of x.

DEFINITION 82. Let x be a node in a labeled frame. The *B-critical cone above x*, we write $\mathcal{C}^B_x$, contains y with xR^By, and is closed under R, S_x and $R \circ S^{\mathsf{tr}}$ transitions[2]. The *generalized B-cone above x*, we write $\mathcal{G}^B_x$, contains $\mathcal{C}^B_x$, and is closed under R and S_w (for arbitrary w) transitions. ❑

The above definition is redundant for IL–frames, where closure under R transitions follows from closure under S_x transitions. However, we want to use the notion of B-critical cones also in the context of frames which are *not* IL–frames. Demanding closure under R, S_x and $R \circ S^{\mathsf{tr}}$ transitions is motivated by the fact that in an ILM–frame, all S_x transitions will then remain inside $\mathcal{C}^B_x$. Consequently, if we want to ensure that $x \Vdash \neg(A \rhd B)$, then $\mathcal{C}^B_x$ will be a "good" place for having an R successor y of x with $y \Vdash A$ — given that we can guarantee that all labels in $\mathcal{C}^B_x$ contain $\neg B$, and not B.

Note that since $\mathcal{C}^B_x \subseteq \mathcal{G}^B_x$ by definition, $\mathcal{G}^B_x \cap \mathcal{G}^C_x = \emptyset$ implies $\mathcal{C}^B_x \cap \mathcal{C}^C_x = \emptyset$ for all B and C.

The notion of adequacy will help us to guarantee that the labels of nodes related via an R or S transition are coherently related themselves. Clauses iii and iv help us to enforce the truth lemma for formulas of the form $\neg(A \rhd B)$. All frames we construct will be adequate in this sense.

---

[2]If S is a relation, we write S^{tr} for the transitive closure of S. The $\circ$ is the composition operator for relations, i.e. $xR \circ Sy$ means that there is some w s.t. $xSwRy$.

DEFINITION 83. An *adequate* frame is a labeled frame with the following properties:

i. $xRy \Rightarrow \nu(x) \prec \nu(y)$

ii. $yS_xz \Rightarrow \nu(y) \subseteq_\Box \nu(z)$

iii. $y \in \mathcal{C}_x^B \Rightarrow \nu(x) \prec_B \nu(y)$

iv. $A \neq B \Rightarrow \mathcal{G}_x^A \cap \mathcal{G}_x^B = \emptyset$ ❑

The notions of problems and deficiencies allow us to approximate the truth lemma step by step. Whenever we eliminate a $\mathcal{D}$-problem or a $\mathcal{D}$-deficiency, we get closer to the truth lemma w.r.t. $\mathcal{D}$. If the set $\mathcal{D}$ is clear or fixed, we will just speak of problems and deficiencies.

DEFINITION 84. A $\mathcal{D}$-problem is a pair $\langle x, \neg(A \rhd B)\rangle$ s.t. $\neg(A \rhd B) \in \nu(x) \cap \mathcal{D}$, but there is no $y \in \mathcal{C}_x^B$ s.t. $A \in \nu(y)$. ❑

DEFINITION 85. A $\mathcal{D}$-deficiency is a triple $\langle x, y, C \rhd D\rangle$ s.t xRy, $C \rhd D \in \nu(x) \cap \mathcal{D}$, $C \in \nu(y) \cap \mathcal{D}$, but for no z s.t. yS_xz we have $D \in \nu(z)$. ❑

4. Overview

This section gives an overview of the construction. Let A be a sentence s.t. $\nvdash_{\mathsf{ILM}} A$, let Γ be a maximal consistent set containing $\neg A$, and let $\mathcal{D}$ be the smallest adequate set containing $\neg A$. Define a labeled frame $F_0 := \langle \{w\}\, \emptyset, \{\emptyset\}\, , \langle w, \Gamma\rangle\rangle$ Note that F_0 is adequate. We will now extend F_0 to an adequate ILM–frame $\mathcal{F}$ containing no problems or deficiencies. It is easy to see that then the truth lemma holds on $\mathcal{F}$.

LEMMA 4.1. *Let $\langle W, R, S, \nu\rangle$ be an adequate labeled frame. Let $\mathcal{M}$ be the model induced by letting $\mathcal{M}, x \Vdash p :\Leftrightarrow p \in \nu(x)$. Then a truth lemma holds in $\mathcal{M}$ w.r.t. $\mathcal{D}$ iff there are no $\mathcal{D}$-problems or $\mathcal{D}$-deficiencies in $\mathcal{F}$.*

PROOF. Immediate. The only non-trivial part of the truth lemma could be reformulated as "there are no problems or deficiencies". ❑

As said above, we will construct $\mathcal{F}$ as the limit of a possibly infinite chain $\{F_n\}_{n\in\omega}$, of ILM–frames. Furthermore, we require each F_n to be adequate. Fix an ordering on the set $\mathcal{P}$ of possible $\mathcal{D}$-problems and -deficiencies in current and future worlds[3]. When going from F_n to F_{n+1}, we will eliminate the problem or deficiency in F_n which is minimal w.r.t. this ordering, guaranteeing that it will not recur in the future. By construction $\mathcal{F} := \bigcup_{n\in\omega} F_n$, does not contain problems or deficiencies.

Apart from eliminating problems and deficiencies, we have to guarantee that each F_n is an ILM–frame. As we will see, a problem or deficiency in F_n is eliminated by adding to F_n a new node together with an appropriate label, as well as a new R or a new S relation. E.g. to eliminate the problem $\langle a, \neg(A \rhd B)\rangle$, we add to W a new node b with aRb, $A \in \nu(b)$, and $b \in \mathcal{C}_a^B$. The resulting frame G is not necessarily an ILM–frame. For example, wRa does not imply wRb in G, i.e. the R relation is not transitive. In order to come back to an adequate ILM–frame F_{n+1}, we have to close off under the frame conditions of ILM.

[3]We will not bother with the exact technical details here.

For this, we will first show that even though G is not an adequate ILM–frame, it is "close enough". The meaning of "close enough" here is captured in the notion of a quasi–ILM–frame, which will be defined in the next section. Second, we will show that every frame which is "close enough" to being an adequate ILM–frame can in fact be extended to an adequate ILM–frame. This extension of G will be the F_{n+1} that we are looking for.

5. Quasi–Frames

DEFINITION 86. An adequate frame $G = \langle W, R, S, \nu \rangle$ is a *quasi–*ILM*–frame* (or just quasi–frame) if the following properties hold:

i. R is converse well-founded

ii. $yS_xz \Rightarrow xRy$ and xRz

iii. $R^{\mathsf{tr}} \circ S^{\mathsf{tr}}$ is converse well-founded ∎

LEMMA 5.1. *Let $G_0 = \langle W, R, S, \nu \rangle$ be a quasi–frame. There is an adequate frame F_{n+1} extending G.*

PROOF. We will construct F_{n+1} as the union of an infinite chain of quasi–frames $\{G_j\}_{j \in \omega}$. We define an imperfection on a quasi frame G_j to be a tuple γ having one of the following forms:

i. $\gamma = \langle 0, a, b, c \rangle$ with $G_j \vDash aRbRc$ but $G_j \nvDash aRc$

ii. $\gamma = \langle 1, a, b, \rangle$ with $G_j \vDash aRb$ but $G_j \nvDash bS_ab$

iii. $\gamma = \langle 2, a, b, c, d \rangle$ with $G_j \vDash bS_acS_ad$ but $G_j \nvDash bS_ad$

iv. $\gamma = \langle 3, a, b, c \rangle$ with $G_j \vDash aRbRc$ but $G_j \nvDash bS_ac$

v. $\gamma = \langle 4, a, b, c, d \rangle$ with $G_j \vDash bS_acRd$ but $G_j \nvDash bRd$

Thus an imperfection on G_j is just a violation of an ILM–frame condition on G_j. Imperfections will be eliminated step by step. Each G_j in the chain will have at least one imperfection less than its predecessor, and the union F_{n+1} will have no imperfections at all. Fix some ordering of the imperfections[4]. To go from $G_j = \langle W_j, R_j, S_j, \nu \rangle$ to G_{j+1}, we choose the least imperfection γ in the ordering. Depending on the form of γ, G_{j+1} will be defined as:

i. $\langle W_j, R_j \cup \{a, c\}, S_j, \nu \rangle$

ii. $\langle W_j, R_j, S_j \cup \{a, b, b\}, \nu \rangle$

iii. $\langle W_j, R_j, S_j \cup \{a, b, d\}, \nu \rangle$

iv. $\langle W_j, R_j \cup \{a, c\}, S_j \cup \{a, b, c\}, \nu \rangle$

v. $\langle W_j, R_j \cup \{b, d\}, S_j, \nu \rangle$

[4]New imperfections will arise during the process. For example, when we set aRc in case i, we will have a new imperfection since $G_{j+1} \nvDash cS_ac$. We will also add the new imperfections to the ordering. We will not bother with the technical details here.

The proof is by induction. It has to be established that in all cases, G_{j+1} will remain a quasi–frame. Thus we have to check that G_{j+1} is adequate, and satisfies the conditions of Definition 86. In total, 35 cases have to be checked, but most of them are straightforward. We will illustrate the general strategy behind the proofs with a few examples.

Suppose that G_{j+1} is defined by case v, i.e. $G_{j+1} = \langle W_j, R_j \cup \{b, d\}, S_j, \nu \rangle$.

We will show that R is converse well-founded. Suppose that there is an infinite sequence s.t. $G_{j+1} \vDash z_1 R z_2 R \ldots$. Replace every occurrence of bRd in the sequence by bS_acRd and leave the rest unchanged. If there are infinitely many S_a transitions in the new sequence, we get a contradiction to the assumption that $R^{\mathsf{tr}} \circ S^{\mathsf{tr}}$ is converse well-founded on G_j. If not, we get a contradiction to the assumption that R is converse well-founded on G_j.

For adequacy, we have to check (among other things) that $y \in \mathcal{C}^B_x \Rightarrow \nu(x) \prec_B \nu(y)$, and that if $A \neq B$, then $\mathcal{C}^B_x \neq \mathcal{C}^A_x$. We will instead prove that the critical and generalized cones of G_j and G_{j+1} are the same[5]. Suppose $G_{j+1} \vDash y \in \mathcal{C}^B_x$. Then there are $z_1 \ldots, z_m$ s.t. $G_{j+1} \vDash x R^B z_1 (S_x \cup R \cup (R \circ S^{\mathsf{tr}})) z_2 \ldots z_m (S_x \cup R \cup (R \circ S^{\mathsf{tr}}))y$. We transform $z_1 \ldots, z_m$ into a new sequence $w_1 \ldots, w_n$ by replacing each occurrence of bRd by bS_acRd. Thus $b(R \circ S^{\mathsf{tr}})d$. We leave the rest of the sequence unchanged. Clearly, $G_j \vDash x R^B w_1 (S_x \cup R \cup (R \circ S^{\mathsf{tr}})) w_2 \ldots w_n (S_x \cup R \cup (R \circ S^{\mathsf{tr}}))y$, whence $G_j \vDash y \in \mathcal{C}^B_x$. A similar proof strategy works for showing that the generalized cones remain unchanged.

In the end, we also have to make sure that $F_{n+1} := \bigcup_{j \in \omega} G_j$ is an adequate ILM–frame.

By construction, F_{n+1} has no imperfections, so it satisfies the corresponding ILM–frame conditions. Furthermore, it is clear that the conditions on adequacy are preserved under unions of chains. The only non-obvious thing is the converse well-foundedness of R. To prove that, we show that for all j,

$$G_j \vDash xRy \Rightarrow G_0 \vDash x(R^{\mathsf{tr}} \circ S^{\mathsf{tr},\mathsf{refl}})^{\mathsf{tr}} y.$$

We will first show how this gives us the desired result. Suppose for contradiction that $F_{n+1} \vDash x_1 R x_2 R x_3 \ldots$. Each of the R relations in the chain was defined at some G_j. By the above claim, we can reduce each of them to an $(R^{\mathsf{tr}} \circ S^{\mathsf{tr},\mathsf{refl}})^{\mathsf{tr}}$ relation in G_0. So we get an infinite chain $\{y_i\}_{i \in \omega}$ s.t.

$$G_0 \vDash y_0 S^{\mathsf{tr},\mathsf{refl}} y_1 R^{\mathsf{tr}} y_2 S^{\mathsf{tr},\mathsf{refl}} y_3 R^{\mathsf{tr}} y_4 S^{\mathsf{tr},\mathsf{refl}} \ldots$$

If there are infinitely many S^{tr} transitions in this sequence, we get a contradiction to the converse well-foundedness of $R^{\mathsf{tr}} \circ S^{\mathsf{tr}}$ on G_0. If not, we get a contradiction to the converse well-foundedness of R on G_0.

The claim $G_j \vDash xRy \Rightarrow G_0 \vDash x(R^{\mathsf{tr}} \circ S^{\mathsf{tr},\mathsf{refl}})^{\mathsf{tr}} y$ is proven by induction on j. The proof is straightforward, once we note that we have to prove the stronger claim: $G_j \vDash xRy \vee xS_z y \Rightarrow G_0 \vDash x(R^{\mathsf{tr}} \circ S^{\mathsf{tr},\mathsf{refl}})^{\mathsf{tr}} y$.

This finishes the proof of Lemma 5.1. ❑

[5]In fact, the critical and generalized cones will remain unchanged throughout the whole process of building F_{n+1}. In all cases, the process might at most shorten the path that a node has to take to enter into a critical or generalized cone.

6. Elimination of Problems and Deficiencies

In this section, we explain how to eliminate problems and deficiencies in an adequate ILM–frame F_n. This is done by adding new nodes and relations to F_n. We show that appropriate labels can be found for the new nodes, and that the resulting frame G is a quasi–frame.

Recall that a problem in F_n is a a pair $\langle a, \neg(A \rhd B)\rangle$ s.t. $\neg(A \rhd B) \in \nu(a) \cap \mathcal{D}$, but there is no $y \in \mathcal{C}_a^B$ s.t. $A \in \nu(y)$. In order to eliminate this problem, we will add a new world b to F_n, with $b \in \mathcal{C}_a^B$ and $A \in \nu(b)$. As $b \in \mathcal{C}_a^B$, and as we want the resulting frame to be adequate, we have to make sure that $\nu(a) \prec_B \nu(b)$. Corollary 88 of Theorem 87 shows that an appropriate label can be found for b.

THEOREM 87. *Let Γ be a maximal* ILM*–consistent set s.t. $\neg(A \rhd B) \in \Gamma$. Let C be s.t. $C \rhd B \in \Gamma$. Then the set $\{\neg C, \Box\neg C, A, \Box\neg A\}$ is consistent.*

PROOF. Suppose for contradiction that $(\neg C \wedge \Box\neg C) \wedge (A \wedge \Box\neg A) \vdash \bot$. Then

$$\begin{aligned}
A \wedge \Box\neg A &\vdash C \vee \Diamond C \\
&\vdash A \wedge \Box\neg A \to C \vee \Diamond C \\
&\vdash \Box(A \wedge \Box\neg A \to C \vee \Diamond C) && \text{(necessitation)} \\
&\vdash A \wedge \Box\neg A \rhd C \vee \Diamond C && \text{(J1)} \\
&\vdash A \wedge \Box\neg A \rhd C && \text{(J5, J3)} \\
&\vdash A \rhd C && \text{(lemma 2.1)}
\end{aligned}$$

Thus $A \rhd C \in \Gamma$. Since $C \rhd B$ in Γ, by J2 also $A \rhd B \in \Gamma$, contradiction. ❑

COROLLARY 88. *Let Γ be a maximal* ILM*–consistent set s.t. $\neg(A \rhd B) \in \Gamma$. Then there exists a maximal* ILM*–consistent set Δ s.t. $\Gamma \prec_B \Delta$ and A, $\Box\neg A \in \Delta$.*

PROOF. We want Δ to be a maximal consistent extension of

$$S := \{\neg C, \Box\neg C \mid C \rhd B \in \Gamma\} \cup \{A, \Box\neg A\}\,.$$

The first set guarantees the first half of the definition for $\Gamma \prec_B \Delta$. For the other half, note that $\Box\neg A \notin \Gamma$. By Lemma 2.1, $\Box\neg A \in \Gamma$ would imply $A \rhd \bot \in \Gamma$, and thus also (as $\bot \rhd B \in \Gamma$) $A \rhd B \in \Gamma$ by J2. Suppose for contradiction that S is inconsistent. By compactness, there is k s.t.

$$(\neg C_1 \wedge \ldots \wedge \neg C_k) \wedge (\Box\neg C_1 \wedge \ldots \wedge \Box\neg C_k) \wedge (A \wedge \Box\neg A) \vdash \bot.$$

Since $\Box$ commutes with $\wedge$, we get

$$(\neg C_1 \wedge \ldots \wedge \neg C_k) \wedge \Box(\neg C_1 \wedge \ldots \wedge \neg C_k) \wedge (A \wedge \Box\neg A) \vdash \bot,$$

and thus

$$(A \wedge \Box\neg A) \vdash (C_1 \vee \ldots \vee C_k) \vee \Diamond(C_1 \vee \ldots \vee C_k).$$

Let $C =: C_1 \vee \ldots \vee C_k$. Since for all $i \leq k$, $C_i \rhd B \in \Gamma$, by J3 $C \rhd B \in \Gamma$. But then

$$(\neg C \wedge \Box\neg C) \wedge (A \wedge \Box\neg A) \vdash \bot$$

contradicting Theorem 87. ❑

Recall that a deficiency in F_n is a triple $\langle a, b, C \rhd D\rangle$ s.t aRb, $C \rhd D \in \nu(a) \cap \mathcal{D}$, $C \in \nu(b) \cap \mathcal{D}$, but for no z s.t. $bS_a z$ we have $D \in \nu(z)$. To eliminate this deficiency, we will add a new node c to F_n with $D \in \nu(c)$ and $bS_a c$. As we want the resulting frame to be adequate, we need to make sure that $\nu(b) \subseteq_\Box \nu(c)$, and that if $b \in \mathcal{C}_x^B$ for some B and x, then $\nu(x) \prec_B \nu(c)$. Corollary 90 of the following theorem shows that an appropriate label can be found for c.

THEOREM 89. *Let Γ be a maximal* ILM*–consistent set s.t. $C \rhd D \in \Gamma$, and let Δ be s.t. $\Gamma \prec_B \Delta$ and $C \in \Delta$. There exists a maximal* ILM*–consistent set Δ' with $\Gamma \prec_B \Delta'$ and $D, \Box\neg D \in \Delta'$.*

PROOF. If $\neg(D \rhd B) \in \Gamma$, then by Corollary 88 there is a Δ' s.t. $\Gamma \prec_B \Delta'$ and $D, \Box\neg D \in \Delta'$. If $\neg(D \rhd B) \notin \Gamma$, then by maximal consistency $D \rhd B \in \Gamma$, whence also $C \rhd B \in \Gamma$ by $\mathsf{J2}$. But then $\Gamma \prec_B \Delta$ implies $\neg C, \Box\neg C \in \Delta$, contradiction. ❑

COROLLARY 90. *Let Γ and Δ be maximal* ILM*–consistent sets s.t. $\Gamma \prec_B \Delta$, $C \rhd D \in \Gamma$, and $C \in \Delta$. There exists a maximal* ILM*–consistent set Δ' s.t. $\Gamma \prec_B \Delta'$, $D, \Box\neg D \in \Delta'$ and $\Delta \subseteq_\Box \Delta'$.*

PROOF. We show that for any $\Box E \in \Delta$, there is a Δ' with $\Gamma \prec_B \Delta'$ and $D, \Box\neg D, \Box E \in \Delta'$. The desired result follows by compactness, and by commutation of boxes and conjunctions (as the proof of Corollary 88.). As $C \rhd D \in \Gamma$ and Γ is maximal ILM–consistent, also $C \wedge \Box E \rhd D \wedge \Box E \in \Gamma$. Clearly, we have that $C \wedge \Box E \in \Delta$. By Corollary 88, we find a Δ' s.t. $\Gamma \prec_B \Delta'$, $D, \Box E, \Box(\neg D \vee \neg\Box E)$. But already in GL, we have that $(D \wedge \Box E \wedge \Box(\neg D \vee \neg\Box E)) \to \Box\neg D$ (using that $\Box E \to \Box\Box E$). Hence also $\Box\neg D \in \Delta'$ as required. ❑

We will now show how a problem or a deficiency in F_n can be eliminated in such a way as to yield a quasi–ILM–frame G. As said before, we will always eliminate the smallest element of $\mathcal{P}$ which is indeed a problem or a deficiency in F.

Problems. Suppose that the least element of $\mathcal{P}$ which is indeed a problem or deficiency in F_n is a problem $\langle a, \neg(A \rhd B)\rangle$. Using Corollary 88, we find a maximal consistent set Δ s.t. $\nu(a) \prec_B \Delta$ and $A, \Box\neg A \in \Delta$. 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 have to check that G is a quasi–ILM–frame, i.e. that it is adequate, and satisfies the conditions of Definition 86.

We will prove here the only case where some work has to be done. For adequacy, we have to show that $G \vDash y \in \mathcal{C}_x^E \Rightarrow G \vDash \nu(x) \prec_E \nu(y)$. So suppose that $G \vDash y \in \mathcal{C}_x^E$. We only need to consider the case where $y = b$, as in all other cases $F_n \vDash y \in \mathcal{C}_x^E \Leftrightarrow G \vDash y \in \mathcal{C}_x^E$. In case $x = a$ and $E = B$, we get the property by the choice of $\nu(b)$. So suppose that $x \neq a$. If $a \in \mathcal{C}_x^E$, then we have $\nu(x) \prec_E \nu(a)$ by adequacy of F_n. Since $\nu(a) \prec \nu(b)$, this implies $\nu(x) \prec_E \nu(b)$. In case $a \notin \mathcal{C}_x^E$, there must be some $w \in \mathcal{C}_x^E$ s.t. $wS^{\mathsf{tr}}a$. By adequacy[6] of F_n, $\nu(w) \subseteq_\Box \nu(a)$. Thus $\nu(x) \prec_E \nu(w) \subseteq_\Box \nu(a) \prec \nu(b)$. It is easy to see that then also $\nu(x) \prec_E \nu(b)$.

[6]Note that the adequacy condition $yS_x w \to y \subseteq_\Box w$, implies that $yS^{\mathsf{tr}}w \Rightarrow \nu(y) \subseteq_\Box \nu(z)$.

Deficiencies. Suppose that the least element of $\mathcal{P}$ which is indeed a problem or deficiency in F_n is a deficiency $\langle a, b, C \rhd D\rangle$. Define B to be the formula s.t. $b \in \mathcal{C}^B_a$. If B does not exist, take B to be $\bot$. If B exists, it must be unique since F_n is adequate. Also by adequacy of F_n, we have that $\nu(a) \prec_B \nu(b)$. By Corollary 90 we find a Δ' s.t. $\nu(a) \prec_B \Delta'$, $D, \Box\neg D \in \Delta'$, and $\nu(b) \subseteq_\Box \Delta'$. Fix some $c \notin W$ and define

$$G = \langle W \cup \{c\}, R \cup \{\langle a, c\rangle\}, S \cup \{\langle a, b, c\rangle\}, \nu \cup \{\langle c, \Delta'\rangle\}\rangle$$

We again have to check that G is a quasi–ILM–frame.

Checking that $G \vDash y \in \mathcal{C}^E_x \Rightarrow G \vDash \nu(x) \prec_E \nu(y)$ is similar as in the case where we eliminated a problem. The only additional way how C would get into $\mathcal{C}^E_x$ is via the S_a transition from b, however then by definition of $\mathcal{C}^E_x$ we have that $x = a$ and thus $E = B$ by the choice of B.

To see that $A \neq B \Rightarrow \mathcal{G}^A_x \cap \mathcal{G}^A_x = \emptyset$, note that for $y \in W$, $F_i \vDash y \in \mathcal{G}^A_x \Leftrightarrow G \vDash y \in \mathcal{C}^A_x$. Thus we only have to consider the possibility that $c \in \mathcal{G}^A_x \cap \mathcal{G}^B_x$. Then we must have that $a \in \mathcal{G}^A_x$ or $b \in \mathcal{G}^A_x$, and $a \in \mathcal{G}^B_x$ or $b \in \mathcal{G}^B_x$. The only problematic case is when (w.l.o.g.) $a \in \mathcal{G}^A_x$ and $b \in \mathcal{G}^B_x$. But since aRb and $\mathcal{G}^A_x$ is closed under R transitions, this implies that also $b \in \mathcal{G}^A_x$, thus by adequacy of F_n we must have that $A = B$.

7. Rounding up

We have seen that if we have a problem (deficiency) on an adequate frame F_n, then there is an adequate frame F_{n+1} extending F_n and not containing that problem (deficiency). Thus we can construct a chain of adequate frames $\{F_n\}_{n\in\omega}$, s.t. each F_n will contain less deficiencies or problems than its predecessor[7]. We define F_0 as in Section 4. It is clear that F_0 is an adequate ILM–frame. We will now show that all problems and deficiencies are eliminated permanently. Then it is clear that $\mathcal{F} := \bigcup_{n\in\omega} F_n$ does not contain any problems or deficiencies, whence the truth-lemma w.r.t. $\mathcal{D}$ holds on $\mathcal{F}$.

It is easy to see that deficiencies, once eliminated, do not recur (because of their $\forall\exists$-nature). With problems, we have to do more work. Suppose that a problem $\langle a, \neg(A \rhd B)\rangle$ has been eliminated from F_n, i.e. we added some $b \in \mathcal{C}^B_a$ with $A \in \nu(b)$. Let $m \geq n$. If bS_ax in F_m, we have by definition that $x \in \mathcal{C}^B_a$. By adequacy of F_m, we get that $\nu(a) \prec_B \nu(x)$, whence $\neg B \in \nu(x)$.

What is left to show is that $\mathcal{F}$ is an adequate frame. It is clear that the conditions for adequacy are preserved under taking unions of chains. The same applies for all ILM–frame conditions except for converse well-foundedness of R. To see that R is converse well-founded on F, note that if $F_n \vDash x_1Rx_2Rx_3 \ldots Rx_m$, then by adequacy $x_1 \prec x_2 \prec \ldots \prec x_m$. But if $x_i \prec x_j$, there is some $E \in \mathcal{D}$ s.t. $\Box D \in \nu(x_j)\backslash\nu(x_i)$. Thus we must have that $m \leq |\mathcal{D}|$. Io.w. the lengths of R-sequences in F_n are bounded by $|\mathcal{D}|$, which is finite. Hence also the lengths of R-sequences in F are bounded by $|\mathcal{D}|$. Thus F is an ILM–frame where a truth lemma holds w.r.t. $\mathcal{D}$, and we are done with the proof.

[7]If F_n contains no problems or deficiencies, we define $F_{n+1} = F_n$.

Bibliography

[Avi03] J. Avigad. Number theory and elementary arithmetic. *Philosophia Mathematica*, 11:257–284, 2003.

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

[Boo93] G. Boolos. *The logic of provability*. Cambridge University Press, Cambridge, 1993.

[Car34] R. Carnap. *Logische Syntax der Sprache*. Schriften zur wissenschaftlichen Weltauffassung. Verlag von Julius Springer, Wien, 1934.

[dJ87] D.H.J. de Jongh. A simplification of a completeness proof of Guaspari and Solovay. *Studia Logica*, 46:187–192, 1987.

[dJ04] R. de Jonge. IL-modellen en bisimulaties. Technical Report X-04 -06, ILLC, University of Amsterdam, 2004.

[dJJM91] D.H.J. de Jongh, M. Jumelet, and F. Montagna. On the proof of Solovay's theorem. *Studia Logica*, 50:51–70, 1991.

[dJV90] D.H.J. de Jongh and F. Veltman. Provability logics for relative interpretability. In [**Pet90**], pages 31–42, 1990.

[dJV91] D.H.J. de Jongh and A. Visser. Explicit fixed points in interpretability logic. *Studia Logica*, 50:39–50, 1991.

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

[FoM] Foundations of mathematics mailing list. `http://www.cs.nyu.edu/pipermail/fom/1999-April/003014.html`. Accessed: 4/07/2012.

[Fri07] H. Friedman. Interpretations according to Tarski. This is one of the 2007 Tarski Lectures at Berkeley. The lecture is available at `http://www.math.osu.edu/~friedman.8/pdf/Tarski1,052407.pdf`, 2007.

[GJ10] E. Goris and J.J. Joosten. Self provers and Σ_1-sentences. *Logic Journal of IGPL*, 0(0):1–22, 2010.

[GS79] D. Guaspari and R.M. Solovay. Rosser sentences. *Annals of Mathematical Logic*, 16:81–99, 1979.

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

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

[Hen11] P. Henk. A New Perspective on the Arithmetical Completeness of GL. Technical Report X-11-06, ILLC, University of Amsterdam, 2011.

[HP91] P. Hájek and P. Pudlák. *Metamathematics of First-Order Arithmetic*. Perspectives in Mathematical Logic. Springer, Berlin, 1991.

[Joo98] J.J. Joosten. *Towards the Interpretability Logic of all Reasonable Arithmetical Theories*. Master's Thesis, ILLC, University of Amsterdam, 1998.

[Joo04] J.J. Joosten. *Interpretability Formalised*. PhD Thesis, University of Utrecht, 2004.

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

[KW07] Richard Kaye and Tin Lok Wong. On interpretations of arithmetic and set theory. *Notre Dame Journal of Formal Logic*, 48(4):497–510, 2007.

[Lin79] P. Lindström. Some results on interpretability. In F.V. Jensen, Mayoh B.H., and Moller K.K., editors, *Proceedings of the 5th Scandinavian Logic Symposium 1979*, pages 329–361, Aalborg, 1979. Aalborg University Press.

[Löb55] M.H. Löb. Solution of a Problem of Leon Henkin. *Journal of Symbolic Logic*, 20:115–118, 1955.

[McA75] K. McAloon. Formules de Rosser pour ZF. *Comptes Rendus hebdomadaires des Séances de l'Académie des Sciences*, Sér. A-B 281(16):A669–A672, 1975.

[Mon87] F. Montagna. Provability in finite subtheories of PA and relative interpretability: A modal investigation. *The Journal of Symbolic Logic*, 52:494–511, 1987.

[Ono87] H. Ono. Reflection principles in fragments of Peano arithmetic. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 33:317–333, 1987.

[Ore61] S. Orey. Relative interpretations. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 7:146–153, 1961.

[Pet90] P.P. Petkov, editor. *Mathematical logic, Proceedings of the Heyting 1988 summer school in Varna, Bulgaria*. Plenum Press, Boston, 1990.

[Ros36] J.B. Rosser. Extensions of some theorems of Gödel and Church. *Journal of Symbolic Logic*, 1:87–91, 1936.

[Ros84] G.F. Rose. *Subrecursion: Functions and Hierarchies*. Clarendon Press, Oxford, 1984.

[SA12] V.Yu Shavrukov and Visser. A. Uniform density in lindenbaum algebras. 2012.

[Seg71] K. Segerberg. An essay in classical modal logic. *Uppsala: Filosofiska Freningen och Filosofiska Institutionen vid Uppsala Universitet*, 1971.

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

[Sha94] V.Yu. Shavrukov. A smart child of Peano's. *Notre Dame Journal of Formal Logic*, 35:161–185, 1994.

[Sie85] W. Sieg. Fragments of arithmetic. *Annals of Pure and Applied Logic*, 28:33–71, 1985.

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

[Smo85] C. Smoryński. *Self-Reference and Modal Logic*. Universitext. Springer, New York, 1985.

[Smo89] C. Smoryński. Arithmetic analogues of McAloon's unique Rosser sentences. *Archive for Mathematical logic*, 28:1–21, 1989.

[Sol76] R.M. Solovay. Provability interpretations of modal logic. *Israel Journal of Mathematics*, 25:287–304, 1976.

[Šve78] V. Švejdar. Degrees of interpretability. *Commentationes Mathematicae Universitatis Carolinae*, 19:789–813, 1978.

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

[Ver93] L.C. Verbrugge. *Efficient metamathematics*. ILLC-disseration series 1993-3, Amsterdam, 1993.

[Vis] A. Visser. Interpretability logic. In P.P. Petkov, editor, *Mathematical logic, Proceedings of the Heyting 1988 summer school in Varna, Bulgaria*.

[Vis91] A. Visser. The formalization of interpretability. *Studia Logica*, 51:81–105, 1991.

[Vis98a] A. Visser. An Overview of Interpretability Logic. In M. Kracht, M. de Rijke, H. Wansing, and M. Zakharyaschev, editors, *Advances in Modal Logic, vol 1*, CSLI Lecture Notes, no. 87, pages 307–359. Center for the Study of Language and Information, Stanford, 1998.

[Vis98b] A. Visser. Interpretations over Heyting's Arithmetic. In E. Orłowska, editor, *Logic at Work*, Studies in Fuzziness and Soft Computing, pages 255–284. Physica-Verlag, Heidelberg/New York, 1998.

[Wan51] H. Wang. Arithmetical models for formal systems. *Methodos*, 3:217–232, 1951.