Interpretability formalized

Interpreteerbaarheid geformaliseerd
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht
op gezag van de Rector Magnificus,
Prof. dr. W.H. Gispen,
ingevolge het besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 26 november 2004 des middags te 14:30 uur
door
Joost Johannes Joosten
geboren op 10 october 1972 te Diemen.

Promotoren:

Prof. dr. A. Visser (Faculteit der Wijsbegeerte, Universiteit Utrecht)

Prof. dr. D. de Jongh (Institute for Logic, Language and Computation, Universiteit van Amsterdam)

Copromotor:

Dr. Lev D. Beklemishev (Faculteit der Wijsbegeerte, Universiteit Utrecht)

Contents

Acknowledgments

Probably, a Georgian toast would be the best way to express my gratitude to all those who have supported me during my doctoral studies. It would be a long, long toast, with many, many names. Let me start with the most prominent among them.

First, I want to mention Albert Visser, my first promoter, who introduced me to the arithmetical aspects of interpretability. I have learned a great deal from his original approach to mathematics.

Dick de Jongh, my second promoter, has warmly and unceasingly supported me since I was an undergraduate student. He taught me provability logic and shared with me his expertise in surprising and beautiful arguments in modal logic.

But most of all, I have learned from Lev Beklemishev about all aspects of provability and interpretability, comprising fragments of arithmetic, modal logics, sequent calculi, cut elimination, models of arithmetic and ordinal notation systems. I feel very privileged to have enjoyed so many of his clear and patient explanations. Also, I cherish my memories of my stay in Moscow doing mathematics with Lev and Nikita at the dacha.

After a motivational dip in 2001, I rediscovered pleasure in doing mathematics by working together with other people. I consider my cooperation with Evan Goris during a period of over more than one year as especially fruitful. Without our project, my time as a PhD student would certainly have been less colorful. Thanks, Evan.

Working with other people has also been very enjoyable. Among them are Lev Beklemishev, Nick Bezhanishvili, Marta Bilkova, Marco Vervoort and Albert Visser. A special word of thanks is due to Marco Vervoort, not only for his computer support, but more important, for helping me out on the proof of the Bisimulation Lemma. Without his fundamental sequence, I would probably still be introducing subsidiary induction parameters.

Apart from my co-authors, there are many other colleagues that have been significant during my time as a PhD student. I would especially like to mention Vincent van Oostrom. Vincent, I think, is an indispensable ingredient for a well-running department. He is interested in the work of his colleagues and is helpful whenever he can be (which is quite often) at almost any hour of the day.

It has been nice sharing a room with Volodya Shavrukov, and I very much appreciate his careful reading of some of my texts.

I have found the cooperation and discussions within the theoretical philosophy department, and especially at our weekly lunch meetings, very stimulating. In this context, I would like to mention Mojtaba Aghaei, Mark van Atten, Jan Bergstra, Marta Bilkova, Sander Bruggink, Dirk van Dalen, Igor Douven, Philipp Gerhardy, Dimitri Hendriks, Herman Hendriks, John Kuiper, Menno Lievers, Janneke van Lith, Jaap van Oosten, Paul van Ulsen and Andreas Weiermann.

Working at the philosophy department in general has also been a very pleasant experience, and a word of thanks is due to the system group, the student administration, Yvonne Elverding and, of course, the students.

Let me proceed to the ILLC. During my time as a PhD student (and before), I have always had a pied-a-terre at the ILLC. Thus, I have profited enormously from the high density of logicians there. The atmosphere at the ILLC has always been pleasant, interdisciplinary and stimulating. It is impossible to mention all the good contacts I have had there. Nevertheless, I would like to mention 'a few': David Ahn, Loredana Afanasiev, Carlos Areces, Felix Bou, Boudewijn de Bruin, Annette Bleeker, Balder ten Cate, Ka Wo Chan, Amanda Collins, Sjoerd Druiven, David Gabelaia, Clemens Grabmayer, Spencer Gerhardt, René Goedman, Paul Harrenstein, Juan Heguiabehere, Aline Honingh, Eva Hoogland, Tanja Kasselaar, Clemens Kupke, Troy Lee, Fenrong Liu, Ingrid van Loon, Maricarmen Martinez, Mikos Massios, Marta Garcia Matos, Mark Theunisse, Siewert van Otterloo, Marc Pauli, Olivier Roy, Yoav Seginer, Darko Sarenac, Brian Semmes, Merlijn Sevenster, Neta Spiro, Dmitry Sustretov, Marjan Veldhuizen, Marco de Vries, Renata Wasserman and Jelle Zuidema.

A special word of thanks is also due to my family for staying true to me during such a long period of isolation. Thanks especially to Ineke, Toon, Sascha, Wolbert and Leila. Muchas gracias también a mi familia en España.

I am definitely not going to mention all my friends here who have not forgotten me. They will be mentioned in the cornucopia of real toasts in a near future. Here, I only mention Demetrio, Han, Roos and Samuel.

I thank my rowing team under the inspiring direction of Leif-Eric van der Leeden for reserving my seat during my absence.

Finally I would like to thank Emma: Muchas gracias por haberme apoyado con tanto amor y paciencia. No puedo imaginar como habria sido sin haberte tenido a mi lado.

Part I

Interpretability and arithmetic

Chapter 1

Introduction

Once upon a Venus transition . . .

1.1 Meta-mathematics and interpretations

As with all sciences, mathematics aims at a better description and understanding of reality. Now, the logician asks: what *is* mathematical reality? Mathematics deals with numbers, functions, shapes, circles, sets, etc. But who has ever touched a number? Who has ever seen a real circle?

The firm and unwavering mathematician answers: who cares? We have clear intuitions about what our mathematical entities are, and we use whatever evident properties (axioms) we get from our intuitions to build up mathematical knowledge. And in a sense, the mathematician is right. We all believe in certain basic mathematical truths, and the applicability of mathematics to other scientific disciplines only lends support to this belief.

Of course, some questions arise. How do we know that our axioms are all true? Which truths can we prove from these axioms? And again, what does *true* mean? If one wants to study these questions, one comes quite naturally to the study of those formal systems where well-defined parts of mathematical reasoning is captured.

Our questions then translate to questions about correctness and strength of these formal systems. Of course, these questions now become relative to other formal systems.

The next question is, how do we compare formal systems? Let us speak of theories from now on. We want to express that a theory S is at least as strong as a theory T. Clearly, this is the case if S proves all the theorems of T. But S and T might speak completely different languages.

In this case, the idea of a translation arises naturally. And translations combined with provability give rise to interpretations. In this dissertation, we shall *use* interpretations to compare theories. Furthermore, we shall also *study* interpretations as meta-mathematical entities.

Roughly, an interpretation j of a theory T into a theory S (we write $j : S \rhd T$) is a structure-preserving map, mapping axioms of T to theorems of S. *Structure-preserving* means that the map should commute with proof constructions and with logical connectives. For example, the constraints on the map should exclude the possibility that we simply map all axioms of T to some tautology of S, say $1 = 1$. Since an interpretation commutes with all proof constructions, it can easily be extended to a map sending all theorems of T to theorems of S.

A moment's reflection tells us that this is indeed a very reasonable way to say that a theory S is at least as strong as a theory T. And in mathematics and meta-mathematics, interpretations turn up time and again in different guises and for different purposes.

A famous and well known example is an interpretation of hyperbolic geometry in Euclidean geometry (e.g., the Beltrami-Klein model, see, for example, [Gre96]) to show the relative consistency of non-Euclidean geometry.

Another example, no less famous, is Gödel's interpretation of the theory of elementary syntax in arithmetic ([Göd31]) to show the incompleteness of, for example, Peano arithmetic.

Interpretations have also been used in partial realizations of Hilbert's programme and other attempts to settle foundational questions ([Sim88], [Fef88], [Nel86]).

For another occurrence of interpretations, we can think of translations of classical propositional calculus into intuitionistic propositional calculus. In this thesis, however, we will only consider interpretations between first order theories.

The notion of interpretability that we shall work with is the notion of relativized interpretability as studied by Tarski et al. in [TMR53]. They use interpretations to show undecidability of certain theories. It is not hard to see that U is undecidable if U interprets some essentially undecidable theory V.

However, it is not the case that $U \rhd V$ implies that U is undecidable whenever V is. For example, the undecidable theory of groups is interpretable in the decidable theory of Abelian groups. But all the theories we will be interested in are essentially undecidable anyway. Moreover, there is a notion, *faithful interpretability,* that does preserve undecidability. A theory V is faithfully interpretable in a theory U if there is a map which is an interpretation so that only theorems of V map to theorems of U.

As a matter of fact, there are many other variants of interpretations which will not be covered in this thesis, such as interpretations with parameters and many-dimensional interpretations. However, as we shall see, the notion of relativized interpretability we choose to work with has many desirable properties. We can distinguish four different approaches to the study of interpretability.

(A) To use interpretations in a series of case studies to relate the proof-theoretic strength of various theories to each other.

(B) To study the nature of interpretability, for example, by relating it to other meta-mathematical notions, like proof-theoretical ordinals, Π_1-conservativity, etc.

(C) To study the general behavioral properties of interpretability and to try to find logics describing this. This line of thinking leads to the study of interpretability logics.

(D) Interpretability induces a preorder on theories. This can be studied as such or by dividing it out to a partial order. This leads to the study of *degrees* or *chapters* (e.g., [Mon58], [Myc77], [Šve78]). We can also consider the category of theories where the interpretations (modulo some appropriate identification) are morphisms.

In this thesis, we shall touch on all four approaches. The emphasis, however, will be on interpretability logics. Two leading questions here concern the interpretability logic of all reasonable arithmetical (we shall call them *numberized*) theories, and the interpretability logic of Primitive Recursive Arithmetic.

1.1.1 Overview of this dissertation

Part I deals primarily with Items (B) and (C). To a lesser extent, we shall also touch on Item (D). The material from Part I comes in large part from [JV04a] and [JV04b].

Part II is completely dedicated to a modal study of interpretability logics. This part is joint work with Evan Goris with some contributions from Marta Bilkova. Chapters 5-7 are taken from [GJ04] and Chapter 8 contains results from [BGJ04].

In Part III we use interpretations in a case study on PRA. Also, we shall make some comments on the interpretability logic of PRA. Results from [Joo02], [Joo03a] and [Joo03b] have been included in this part. Subsection 12.3 is joint work with Lev Beklemishev and Marco Vervoort.

1.2 Preliminaries

As we already mentioned, our notion of interpretability is the one studied by Tarski et al in [TMR53]. The theories that we study in this dissertation are theories formulated in first order predicate logic. All theories have a finite signature that contains identity. For simplicity we shall assume that all our theories are formulated in a purely relational way. Here is the formal definition of a relative interpretation.

Definition 1.2.1. A relative interpretation k of a theory S into a theory T is a pair $\langle \delta, F\rangle$ for which the following holds. The first component δ, is a formula in the language of T with a single free variable. This formula is used to specify the domain of our interpretation. The second component, F, is a finite map that sends relation symbols R (including identity) from the language of S, to formulas $F(R)$ in the language of T. We demand for all R that the number of free variables of $F(R)$ equals the arity of R.[1] Recursively we define the translation φ^k of a formula φ in the language of S as follows.

- $(R(\vec{x}))^k = F(R)(\vec{x})$
- $(\varphi \wedge \psi)^k = \varphi^k \wedge \psi^k$ and likewise for other boolean connectives (this implies $\bot^k = \bot$)
- $(\forall x\ \varphi(x))^k = \forall x\ (\delta(x) \to \varphi^k)$ and analogously for the existential quantifier

Finally, we demand that $T \vdash \varphi^k$ for all axioms φ of S.

We assume the reader to have some familiarity with basic arithmetical theories like Buss's S^1_2, EA ($= \mathrm{I}\Delta_0 + \exp$), $\mathrm{I}\Sigma_1$, PA etcetera. We also assume some familiarity with arithmetical hierarchies as the Σ_n-sentences and the bounded arithmetical hierarchies like the Σ^b_n. (See for example, [Bus98] or [HP93].)

Moreover, we shall employ techniques and concepts necessary for the arithmetization of syntax. Thus, we shall work with provability predicates $\Box_U$ corresponding uniformly to arithmetical theories U.

We shall always write the formalized version of a concept in sans-serif style. For example, $\mathsf{Proof}_U(p, \varphi)$ stands for the formalization of "p is a U-proof of φ", $\mathsf{Con}(U)$ stands for the formalization of "U is a consistent theory" and so forth. Occasionally we shall employ truth predicates. Again, [Bus98] and [HP93] are adequate references.

1.2.1 A short word on coding

There are many good reasons to switch to formalized interpretability for our study. As we shall see, we can use formalized interpretability, just like Gödel used formalized provability, to study a theory and its limitations.

In a formalized setting it is straightforward to give a meaning to expressions involving iterated provability and interpretability statements. Moreover, by formalization we get access to powerful reasoning like the fixed-point lemma for arithmetic and so on.

Formalization calls for coding of syntax. At some places in this paper we shall need estimates of codes of syntactical objects. Therefore it is good to discuss the nature of the coding process we will employ. However we shall not consider the implementation details of our coding.

We shall code strings over some finite alphabet A with cardinality a. First we define an alphabetic order on A. Next we enumerate all finite strings over A in the following way. First we enumerate all strings of length 0, then of length 1, etcetera. For every n, we enumerate the strings of length n in alphabetic order. The coding of a finite string over A will just be its ordinal number in this enumeration. We shall now see some easy arithmetical properties of this coding. We shall often refrain from distinguishing syntactical objects and their codes.

[1] Formally, we should be more precise and specify our variables.

1. There are a^n many strings of length n.

2. There are $a^n + a^{n_1} \cdots + 1 = \frac{a^{n+1}-1}{a-1}$ many strings of length $\leq n$.

3. From (2) it follows that the code of a syntactical object of length n, is $\mathcal{O}(\frac{a^{n+1}-1}{a-1}) = \mathcal{O}(a^n)$ big.

4. Conversely, the length of a syntactical object that has code φ is $\mathcal{O}(|\varphi|)$ (logarithm of φ) big.

5. If φ and ψ are codes of syntactical objects, the concatenation $\varphi \star \psi$ of φ and ψ is $\mathcal{O}(\varphi \cdot \psi)$ big. For, $|\varphi \star \psi| = |\varphi| + |\psi|$, whence by (3), $\varphi \star \psi \approx a^{|\varphi|+|\psi|} = a^{|\varphi|} \cdot a^{|\psi|} = \varphi \cdot \psi$.

6. If φ and t are (codes of) syntactical objects, then $\varphi_x(t)$ is $\mathcal{O}(\varphi^{|t|})$ big. Here $\varphi_x(t)$ denotes the syntactical object that results from φ by replacing every (unbounded) occurrence of x by t. The length of φ is about $|\varphi|$. In the worst case, these are all x-symbols. In this case, the length of $\varphi_x(t)$ is $|\varphi| \cdot |t|$ and thus $\varphi_x(t)$ is $\mathcal{O}(a^{|\varphi|\cdot|t|}) = \mathcal{O}(t^{|\varphi|}) = \mathcal{O}(\varphi^{|t|})$ big.

We want to represent numbers by terms and then consider the code of the term. It is not a good idea to represent a number n by $\overbrace{S \ldots S}^{n \text{ times}} 0$. For, the length of this object is $n+1$ whence its code is about 2^{n+1} and we would like to avoid the use of exponentiation. In the setting of weaker arithmetics it is common practice to use so-called *efficient numerals*. These numerals are defined by recursion as follows. $\overline{0} = 0$; $\overline{2{\cdot}n} = (SS0) \cdot \overline{n}$ and $\overline{2{\cdot}n + 1} = S((SS0) \cdot \overline{n})$. Clearly, these numerals implement the system of dyadic notation.

1.2.2 Arithmetical theories

In this paper, we shall be mainly concerned with arithmetical theories. In doing so, formalization of interpretability becomes a routine matter. Moreover, it facilitates us to relate interpretability to other meta-mathematical notions that typically use arithmetic.

We do not demand that our theories are formulated in the language of arithmetic. Instead, we demand that some sufficiently strong fragment of number theory should be embeddable, viz. interpretable in our theories.

Reasonable arithmetical theories

As we have just agreed, our theories should contain a sufficient amount of arithmetic. Sufficient means here, enough to do coding and elementary arguments. On the other hand, we do not want to exclude many interesting weaker theories by demanding too much arithmetic.

In Subsection 1.2.1 we have seen that a substitution operation on codes of syntactical objects asks for a function of growth rate $x^{|x|}$. Reasonable arithmetical theories should thus also have such a function. In Buss's S^1_2 this is the smash function $\sharp$. In the theory $I\Delta_0 + \Omega_1$ this is the function $\omega_1(x)$. In this paper we choose[2] to work with S^1_2.

Definition 1.2.2. We will call a pair $\langle U, k\rangle$ a *numberized theory* if $k : U \rhd \mathsf{S}^1_2$. A theory U is *numberizable* or *arithmetical* if for some j, $\langle U, j\rangle$ is a numberized theory.

From now on, we shall only consider numberizable or numberized theories. Often however, we will fix a numberization j and reason about the theory $\langle U, j\rangle$ as if it were formulated in the language of arithmetic. Moreover, we shall most of the times work with sequential theories. Basically, sequentiality means that any finite sequence of objects can be coded.

As we want to do arithmetization of syntax, our theories should be coded in a simple way. We will assume that all our theories U have an axiom set that is decidable in polynomial time. That is, there is some formula $\mathsf{Axiom}_U(x)$ which is Δ^b_1 in S^1_2, with

- $\mathsf{S}^1_2 \vdash \mathsf{Axiom}_U(\varphi)$ iff φ is an axiom of U.

The choice of Δ^b_1-axiomatizations is also motivated by Lemma 1.2.3.

For already really weak theories we have Σ_1-completeness. However, proofs of Σ_1-sentences σ are multi-exponentially big, that is, 2^σ_n for some n depending on σ. (See e.g., [HP93].)

However, for $\exists\Sigma^b_1$-formulas we do have a completeness theorem (see [Bus98]). From now on, we shall often write a sup-index to a quantifier to specify the domain of quantification.

Lemma 1.2.3. *If $\alpha(x) \in \exists\Sigma^b_1$, then there is some standard natural number n such that*

$$\mathsf{S}^1_2 \vdash \forall x\ [\alpha(x) \to \exists\, p{<}\omega^n_1(x)\ \mathsf{Proof}_U(p, \alpha(\dot{x}))].$$

This holds for any reasonable arithmetical theory U. Moreover, we have also a formalized version of this statement.

$$\mathsf{S}^1_2 \vdash \forall^{\exists\Sigma^b_1}\alpha\, \exists n\ \Box_{\mathsf{S}^1_2}(\forall x\ [\alpha(x) \to \exists\, p{<}\omega^n_1(x)\ \mathsf{Proof}_U(p, \alpha(\dot{x}))]).$$

Reflexive theories

Many meta-mathematical statements involve the notion of reflexivity. A theory is reflexive if it proves the consistency of all of its finite subtheories. There exist various ways in which reflexivity can be formalized, and throughout literature we can find many different formalizations. For stronger theories, all these formalizations coincide. But for weaker theories, the differences are essential. We give some formalizations of reflexivity.

[2] The choice of S^1_2 is motivated as follows. Robinson's arithmetic Q is too weak for some of our arguments. On the other hand $I\Delta_0 + \Omega_1$ aka S_2 is not known to be finitely axiomatizable. However, with some care, we could have used $I\Delta_0 + \Omega_1$ as well.

1. $\forall n\, U \vdash \mathsf{Con}(U[n])$ where $U[n]$ denotes the conjunction of the first n axioms of U.

2. $\forall n\, U \vdash \mathsf{Con}(U{\restriction}n)$ where $\mathsf{Con}(U{\restriction}n)$ denotes that there is no proof of falsity using only axioms of U with Gödel numbers $\leq n$.

3. $\forall n\; U \vdash \mathsf{Con}_n(U)$ where $\mathsf{Con}_n(U)$ denotes that there is no proof of falsity with a proof p where p has the following properties. All non-logical axioms of U that occur in p have Gödel numbers $\leq n$. All formulas φ that occur in p have a logical complexity $\rho(\varphi) \leq n$.
Here ρ is some complexity measure that basically counts the number of quantifier alternations in φ. Important features of this ρ are that for every n, there are truth predicates for formulas with complexity n. Moreover, the ρ-measure of a formula should be more or less (modulo some poly-time difference, see Remark 1.2.8) preserved under translations. An example of such a ρ is given in [Vis93].

It is clear that (1) ⇒ (2) ⇒ (3). For the corresponding provability notions, the implications reverse. In this paper, our notion of reflexivity shall be the third one.

We shall write $\Box_{U,n}\varphi$ for $\neg\mathsf{Con}_n(U + \neg\varphi)$ or, equivalently, $\exists p\ \mathsf{Proof}_{U,n}(p,\varphi)$. Here, $\mathsf{Proof}_{U,n}(p,\varphi)$ denotes that p is a U-proof of φ with all axioms in p are $\leq n$ and for all formulas ψ that occur in p, we have $\rho(\psi) \leq n$.

Remark 1.2.4. An inspection of the proof of provable Σ_1-completeness (Lemma 1.2.3) gives us some more information. The proof p that witnesses the provability in U of some $\exists\Sigma_1^b$-sentence α, can easily be taken cut-free. Moreover, all axioms occurring in p are about as big as α. Thus, from α, we get for some n (depending linearly on α) that $\mathsf{Proof}_{U,n}(p,\alpha)$.

If we wish to emphasize the fact that our theories are not necessarily in the language of arithmetic, but just can be numberized, our formulations of reflexivity should be slightly changed. For example, (3) will for some $\langle U, j\rangle$ look like $j : U \rhd \mathsf{S}^1_2 + \{\mathsf{Con}_n(U) \mid n \in \omega\}$. This also explains the prominent role of the reflexivization functor $\mho_{(\cdot)}$ as studied in Subsection 2.2.

If U is a reflexive theory, we do not necessarily have any reflection principles. That is, we do not have $U \vdash \Box_V\varphi \to \varphi$ for some natural $V \subset U$ and for some natural class of formulae φ. We do have, however, a weak form of $\forall\Pi^b_1$-reflection. This is expressed in the following lemma.

Lemma 1.2.5. *Let U be a reflexive theory. Then*

$$\mathsf{S}^1_2 \vdash \forall^{\forall\Pi^b_1}\pi\,\forall n\ \Box_U\forall x\ (\Box_{U,n}\pi(\dot{x}) \to \pi(x)).$$

Proof. Reason in S^1_2 and fix π and n. Let m be such that we have (see Lemma 1.2.3 and Remark 1.2.4)

$$\Box_U\forall x\ (\neg\pi(x) \to \Box_{U,m}\neg\pi(\dot{x})).$$

Furthermore, let $k := \max\{n, m\}$. Now, reason in U, fix some x and assume $\Box_{U,n}\pi(x)$. Thus, clearly also $\Box_{U,k}\pi(x)$. If now $\neg\pi(x)$, then also $\Box_{U,k}\neg\pi(x)$, whence $\Box_{U,k}\bot$. This contradicts the reflexivity, whence $\pi(x)$. As x was arbitrary we get $\forall x\ (\Box_{U,n}\pi(x) \to \pi(x))$. $\dashv$

We note that this lemma also holds for the other notions of restricted provability we introduced in this subsection.

1.2.3 Interpretability in a weak meta-theory

To formalize insights about interpretability in weak meta-theories like S^1_2 we need to be very careful. Definitions of interpretability that are unproblematically equivalent in a strong theory like, say, $\mathrm{I}\Sigma_1$ diverge in weak theories. As we shall see, the major source of problems is the absence of $\mathrm{B}\Sigma_1$.

Here $\mathrm{B}\Sigma_1$ is the so-called collection scheme for Σ_1-formulae. Roughly, $\mathrm{B}\Sigma_1$ says that the range of a Σ_1-definable function on a finite interval is again finite. A mathematical formulation is $\forall\, x{\leq}u\, \exists y\, \sigma(x, y) \to \exists z\, \forall\, x{\leq}u\, \exists\, y{\leq}z\, \sigma(x, y)$ where $\sigma(x, y) \in \Sigma_1$ may contain other variables too. In this subsection, we study various divergent definitions of interpretability.

We start by making an elementary observation on interpretations. Basically, the next definition and lemma say that interpretations transform proofs into translated proofs.

Definition 1.2.6. Let k be a translation. By recursion on a proof p in natural deduction we define the translation of p under k, we write p^k. For this purpose, we first define $k(\varphi)$ for formulae φ to be[3] $\bigwedge_{x_i\in\mathsf{FV}(\varphi)} \delta(x_i) \to \varphi^k$. Here $\mathsf{FV}(\varphi)$ denotes the set of free variables of φ. Clearly, this set cannot contain more than $|\varphi|$ elements, whence $k(\varphi)$ will not be too big. Obviously, for sentences φ, we have $k(\varphi) = \varphi^k$.

If p is just a single assumption φ, then p^k is $k(\varphi)$. The translation of the proof constructions are defined precisely in such a way that we can prove Lemma 1.2.7 below. For example, the translation of

$$\frac{\varphi \quad \psi}{\varphi\wedge\psi}$$

will be

$$\dfrac{\dfrac{\dfrac{\dfrac{[\bigwedge_{x_i\in\mathsf{FV}(\varphi\wedge\psi)}\delta(x_i)]_1}{\bigwedge_{x_i\in\mathsf{FV}(\varphi)}\delta(x_i)} \qquad \bigwedge_{x_i\in\mathsf{FV}(\varphi)}\delta(x_i)\to\varphi^k}{\varphi^k} \qquad \dfrac{\mathcal{D}}{\psi^k}}{\varphi^k\wedge\psi^k}}{\bigwedge_{x_i\in\mathsf{FV}(\varphi\wedge\psi)}\delta(x_i)\to\varphi^k\wedge\psi^k}\ \to I, 1$$

[3] To be really precise we should say that, for example, we let smaller x_i come first in $\bigwedge_{x_i\in\mathsf{FV}(\varphi)} \delta(x_i)$.

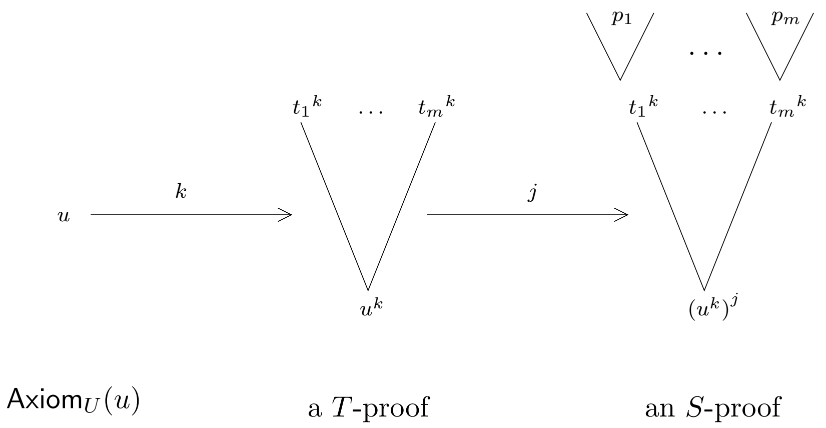

Figure 1.1: Transitivity of interpretability

where $\mathcal{D}$ is just a symmetric copy of the part above φ^k. We note that the translation of the proof constructions is available[4] in S^1_2, as the number of free variables in $\varphi \wedge \psi$ is bounded by $|\varphi \wedge \psi|$.

Lemma 1.2.7. *If p is a proof of a sentence φ with assumptions in some set of sentences Γ, then for any translation k, p^k is a proof of φ^k with assumptions in Γ^k.*

Proof. Note that the restriction on sentences is needed. For example

$$\frac{\forall x\ \varphi(x) \quad \forall x\ (\varphi(x) \to \psi(x))}{\psi(x)}$$

but

$$\frac{(\forall x\ \varphi(x))^k \quad (\forall x\ (\varphi(x) \to \psi(x)))^k}{\delta(x) \to \psi^k(x)}$$

and in general $\nvdash (\delta(x) \to \psi^k) \leftrightarrow \psi^k$. The lemma is proved by induction on p. To account for formulas in the induction, we use the notion $k(\varphi)$ from Definition 1.2.6, which is tailored precisely to let the induction go through. $\dashv$

Remark 1.2.8. The proof translation leaves all the structure invariant. Thus, there is a provably total (in S^1_2) function f such that , if p is a U, n-proof of φ, then p^k is a proof of φ^k, where p^k has the following properties. All axioms in p^k are $\leq f(n,k)$ and all formulas ψ in p^k have $\rho(\psi) \leq f(n,k)$.

There are various reasons to give, why we want the notion of interpretability to be transitive, that is, $S \rhd U$ whenever $S \rhd T$ and $T \rhd U$. The obvious way of proving this would be by composing (doing the one after the other) two

[4] More efficient translations on proofs are also available. However they are less uniform.

In S^1_2:

$$\begin{tikzcd}
j : U \rhd_t V \arrow[dd, shift left] \arrow[dr, dashed, shift left, "\mathrm{exp}"] & & \\
& j : U \rhd_{sa} V \arrow[ul, shift left] \arrow[dl, shift left] \arrow[r, shift left] & j : U \rhd_{st} V \arrow[l, shift left] \\
j : U \rhd_a V \arrow[uu, shift left, "\times" marking] \arrow[ur, dashed, shift left, "\mathrm{B}\Sigma_1"'] & &
\end{tikzcd}$$

Figure 1.2: Versions of relative interpretability

interpretations. Thus, if we have $j : S \rhd T$ and $k : T \rhd U$ we would like to have[5] $j \circ k : S \rhd U$.

If we try to perform this proof as depicted in Figure 1.1, at a certain point we would like to collect the S-proofs $p_1, \cdots, p_m$ of the j-translated T-axioms used in a proof of a k-translation of an axiom u of U, and take the maximum of all such proofs. But to see that such a maximum exists, we precisely need Σ_1-collection.

However, it is desirable to also reason about interpretability in the absence of $\mathrm{B}\Sigma_1$. A trick is needed to circumvent the problem of the unprovability of transitivity (and many other elementary desiderata).

One way to solve the problem is by switching to a notion of interpretability where the needed collection has been built in. This is the notion of smooth (axioms) interpretability as in Definition 1.2.9. In this thesis we shall mean by interpretability, unless mentioned otherwise, always smooth interpretability. In the presence of $\mathrm{B}\Sigma_1$ this notion will coincide with the earlier defined notion of interpretability, as Theorem 1.2.10 tells us.

Definition 1.2.9. We define the notions of axioms interpretability $\rhd_a$, theorems interpretability $\rhd_t$, smooth axioms interpretability $\rhd_{sa}$ and smooth theorems interpretability $\rhd_{st}$.

$$\begin{array}{lcl}
j : U \rhd_a V & := & \forall v\, \exists p\ (\mathsf{Axiom}_V(v) \to \mathsf{Proof}_U(p, v^j)) \\
j : U \rhd_t V & := & \forall \varphi\, \forall p\, \exists p'\ (\mathsf{Proof}_V(p, \varphi) \to \mathsf{Proof}_U(p', \varphi^j)) \\
j : U \rhd_{sa} V & := & \forall x\, \exists y\, \forall v{\leq}x\, \exists p{\leq}y\ (\mathsf{Axiom}_V(v) \to \mathsf{Proof}_U(p, v^j)) \\
j : U \rhd_{st} V & := & \forall x\, \exists y\, \forall \varphi{\leq}x\, \forall p{\leq}x\, \exists p'{\leq}y\ (\mathsf{Proof}_V(p, \varphi) \to \mathsf{Proof}_U(p', \varphi^j))
\end{array}$$

Theorem 1.2.10. *In S^1_2 we have all the arrows as depicted in Figure 1.2: Versions of relative interpretability. The dotted arrows indicate that an additional*

[5] A formal definition of $j \circ k$ is given in Section 3.1.

condition is needed in our proof; the condition written next to it. The arrow with a cross through it, indicates that we know that the implication fails in S^1_2.

Proof. We shall only comment on the arrows that are not completely trivial.

- $T \vdash j : U \rhd_a V \to j : U \rhd_{sa} V$, if $T \vdash \mathrm{B}\Sigma_1$. So, reason in T and suppose $\forall v\, \exists p\ (\mathsf{Axiom}_V(v) \to \mathsf{Proof}_U(p, v^j))$. If we fix some x, we get $\forall\, v{\leq}x\, \exists p\ (\mathsf{Axiom}_V(v) \to \mathsf{Proof}_U(p, v^j))$. By $\mathrm{B}\Sigma_1$ we get the required $\exists y\, \forall\, v{\leq}x\, \exists\, p{\leq}y\ (\mathsf{Axiom}_V(v) \to \mathsf{Proof}_U(v^j))$. It is not clear if $T \vdash \mathrm{B}\Sigma_1^-$, parameter-free collection, is a necessary condition.

- $\mathsf{S}^1_2 \nvdash j : U \rhd_a V \to j : U \rhd_t V$. A counter-example is given in [Vis91].

- $T \vdash j : U \rhd_t V \to j : U \rhd_{sa} V$, if $T \vdash \mathrm{exp}$. If T is reflexive, we get by Corollary 2.1.9 that $\vdash U \rhd_t V \leftrightarrow U \rhd_{sa} V$. However, different interpretations are used to witness the different notions of interpretability in this case. If $T \vdash \mathrm{exp}$, we reason as follows. We reason in T and suppose that $\forall \varphi\, \forall p\, \exists p'\ (\mathsf{Proof}_V(p, \varphi) \to \mathsf{Proof}_U(p', \varphi^j))$. We wish to see

$$\forall x\, \exists y\, \forall\, v{\leq}x\, \exists\, p{\leq}y\ (\mathsf{Axiom}_V(v) \to \mathsf{Proof}_U(v^j)). \tag{1.1}$$

So, we pick x arbitrarily and consider[6] $\nu := \bigwedge_{\mathsf{Axiom}_V(v_i) \wedge v_i \leq x} v_i$. Notice that in the worst case, for all $y \leq x$, we have $\mathsf{Axiom}_V(y)$, whence the length of ν can be bounded by $x \cdot |x|$. Thus, ν itself can be bounded by x^x, which exists whenever $T \vdash \mathrm{exp}$. Clearly, $\exists p\ \mathsf{Proof}_V(p, \nu)$ whence by our assumption $\exists p'\ \mathsf{Proof}_U(p', \nu^j)$. In a uniform way, with just a slightly larger proof p'', every ${v_i}^j$ can be extracted from the proof p' of ν^j. We may take this $p'' \approx y$ to obtain (1.1). It is not clear if $T \vdash \mathrm{exp}$ is a necessary condition.

- $\mathsf{S}^1_2 \vdash j : U \rhd_{sa} V \to j : U \rhd_{st} V$. So, we wish to see that

$$\forall x\, \exists y\, \forall\, \varphi{\leq}x\, \forall\, p{\leq}x\, \exists\, p'{\leq}y\ (\mathsf{Proof}_V(p, \varphi) \to \mathsf{Proof}_U(p', \varphi^j))$$

from the assumption that $j : U \rhd_{sa} V$. So, we pick x arbitrarily. If now for some $p \leq x$ we have $\mathsf{Proof}_V(p, \varphi)$, then clearly $\varphi \leq x$ and all axioms v_i of V that occur in p are $\leq x$. By our assumption, we can find a y_0 such that we can find proofs $p_i \leq y_0$ for all the ${v_i}^j$. Now, with some sloppy notation, $(p^j)_{{v_i}^j}(p_i)$ is a proof for φ^j. This proof can be estimated (again with sloppy notations).

$$(p^j)_{{v_i}^j}(p_i) \leq (p^j)_{{v_i}^j}(y_0) \leq (p^j)^{|y_0|} \leq (x^j)^{|y_0|}$$

The latter bound is clearly present in S^1_2. ⊣

We note that we have many admissible rules from one notion of interpretability to another. For example, by Buss's theorem on the provably total recursive

[6] To see that ν exists, we seem to also use some collection; we collect all the $v_i \leq x$ for which $\mathsf{Axiom}_V(v_i)$. However, it is not hard to see that we can consider ν also without collection.

functions of S^1_2, it is not hard to see that

$$\mathsf{S}^1_2 \vdash j : U \rhd_a V \Rightarrow \mathsf{S}^1_2 \vdash j : U \rhd_t V.$$

In the rest of this thesis, we shall at most places no longer write subscripts to the $\rhd$'s. Our reading convention is then that we take that notion of interpretability that is best to perform the argument. Often this is just smooth interpretability $\rhd_s$, which from now on is the name for $\rhd_{sa}$.

Moreover, in [Vis91] some sort of conservation result concerning $\rhd_a$ and $\rhd_s$ is proved. For a considerable class of formulas φ and theories T, and for a considerable class of arguments we have that $T \vdash \varphi_a \Rightarrow T \vdash \varphi_s$. Here φ_a denotes the formula φ using the notion $\rhd_a$ and likewise for φ_s. Thus indeed, in many cases a sharp distinction between the notions involved is not needed.

We could also consider the following notion of interpretability.

$$j : U \rhd_{st_1} V := \forall x\, \exists y\, \forall\, \varphi{\leq}x\, \exists\, p'{\leq}y\ (\Box_V \varphi \to \mathsf{Proof}_U(p', \varphi^j))$$

Clearly, $j : U \rhd_{st_1} V \to U \rhd_{st} V$. However, for the reverse implication one seems to need $\mathrm{B}\Pi^-_1$. Also, a straightforward proof of $U \vdash \mathsf{id} : U \rhd_{st_1} U$ seems to need $\mathrm{B}\Pi^-_1$. Thus, the notion $\rhd_{st_1}$ seems to say more on the nature of a theory than on the nature of interpretability.

1.2.4 Interpretations and models

We can view interpretations $j : U \rhd V$ as a way of defining uniformly a model $\mathcal{N}$ of V inside a model $\mathcal{M}$ of U. Interpretations in foundational papers mostly bear the guise of a uniform model construction.

Definition 1.2.11. Let $j : U \rhd V$ with $j = \langle \delta, F \rangle$. If $\mathcal{M} \models U$, we denote by $\mathcal{M}^j$ the following model.

- $|\mathcal{M}^j| = \{x \in |\mathcal{M}| \mid \mathcal{M} \models \delta(x)\}/\equiv$, where $a \equiv b$ iff $\mathcal{M} \models a =^j b$.
- $\mathcal{M}^j \models R(\alpha_1, \ldots, \alpha_n)$ iff $\mathcal{M} \models F(R)(a_1, \ldots, a_n)$, for some $a_1 \in \alpha_1, \ldots ,$ $a_n \in \alpha_n$.

The fact that $j : U \rhd V$ is now reflected in the observation that, whenever $\mathcal{M} \models U$, then $\mathcal{M}^j \models V$.

On many occasions viewing interpretations as uniform model constructions provides the right heuristics.

1.3 Cuts and induction

Inductive reasoning is a central feature of everyday mathematical practice. We are so used to it, that it enters a proof almost unnoticed. It is when one works with weak theories and in the absence of sufficient induction, that its all pervading nature is best felt.

A main tool to compensate for the lack of induction are the so-called definable cuts. They are definable initial segments of the natural numbers that possess some desirable properties that we could not infer for all numbers to hold by means of induction.

The idea is really simple. So, if we can derive $\varphi(0) \wedge \forall x\ (\varphi(x) \to \varphi(x+1))$ and do not have access to an induction axiom for φ, we just consider $J(x)$: $\forall\, y{\leq}x\ \varphi(y)$. Clearly J now defines an initial segment on which φ holds. As we shall see, for a lot of reasoning we can restrict ourselves to initial segments rather than quantifying over all numbers.

1.3.1 Basic properties of cuts

Throughout the literature one can find some variations on the definition of a cut. At some places, a cut is only supposed to be an initial segment of the natural numbers. At other places some additional closure properties are demanded. By a well known technique due to Solovay (see for example [HP93]) any definable initial segment can be shortened in a definable way, so that it has a lot of desirable closure properties. Therefore, and as we almost always need the closure properties, we include them in our definition.

Definition 1.3.1. A definable U-cut is a formula $J(x)$ with only x free, for which we have the following.

1. $U \vdash J(0) \wedge \forall x\ (J(x) \to J(x+1))$
2. $U \vdash J(x) \wedge y{\leq}x \to J(y)$
3. $U \vdash J(x) \wedge J(y) \to J(x+y) \wedge J(x \cdot y)$
4. $U \vdash J(x) \to J(\omega_1(x))$

We shall sometimes also write $x \in J$ instead of $J(x)$. A first fundamental insight about cuts is the principle of *outside big, inside small.* Although not every number x is in J, we can find for every x a proof p_x that witnesses $x \in J$.

Lemma 1.3.2. *Let T and U be reasonable arithmetical theories and let J be a U-cut. We have that*

$$T \vdash \forall x\ \Box_U J(x).$$

Actually, we can have the quantifier over all cuts within the theory T, that is

$$T \vdash \forall^{U\text{-}\mathsf{Cut}} J\, \forall x\ \Box_U J(x).$$

Proof. Let us start by making the quantifier $\forall^{U\text{-}\mathsf{Cut}} J$ a bit more precise. By $\forall^{U\text{-}\mathsf{Cut}} J$ we shall mean $\forall J\ (\Box_U \mathsf{Cut}(J) \to \ldots)$. Here $\mathsf{Cut}(J)$ is the definable function that sends the code of a formula χ with one free variable to the code of the formula that expresses that χ defines a cut.

For a number a, we start with the standard proof of $J(0)$. This proof is combined with $a{-}1$ many instantiations of the standard proof of $\forall x\ (J(x) \to J(x{+}1))$. In the case of weaker theories, we have to switch to efficient numerals to keep the bound of the proof within range. $\dashv$

Remark 1.3.3. The proof sketch actually tells us that (provably in S^1_2) for every U-cut J, there is an $n \in \omega$ such that $\forall x \ \Box_{U,n} J(x)$.

Lemma 1.3.4. *Cuts are provably closed under terms, that is*

$$T \vdash \forall^{U\text{-}\mathsf{Cut}} J \, \forall^{\mathsf{Term}} t \ \Box_U \forall \, \vec{x} {\in} J \ t(\vec{x}) \in J.$$

Proof. By an easy induction on terms, fixing some U-cut J. Prima facie this looks like a Σ_1-induction but it is easy to see that the proofs have poly-time (in t) bounds, whence the induction is $\Delta_0(\omega_1)$. ⊣

As all U-cuts are closed under $\omega_1(x)$ and the smash function $\sharp$, simply relativizing all quantors to a cut is an example of an interpretation of S^1_2 in U. We shall always denote both the cut and the interpretation that it defines by the same symbol.

1.3.2 Cuts and the Henkin construction

It is well known that we can perform the Henkin construction in a rather weak meta-theory. As the Henkin model has a uniform description, we can link it to interpretations. The following theorem makes this precise.

Theorem 1.3.5. *If $U \vdash \mathsf{Con}(V)$, then $U \rhd V$.*

Early treatments of this theorem were given in [Wan51] [HB68]. A first fully formalized version was given in [Fef60]. A proof of Theorem 1.3.5 would closely follow the Henkin construction.

Thus, first the language of V is extended so that it contains a witness $c_{\exists x \varphi(x)}$ for every existential sentence $\exists x \ \varphi(x)$. Then we can extend V to a maximal consistent V' in the enriched language, containing all sentences of the form $\exists x \varphi(x) \to \varphi(c_{\exists x \varphi(x)})$. This V' can be seen as a term model with a corresponding truth predicate. Clearly, if $V \vdash \varphi$ then $\varphi \in V'$. It is not hard to see that V' is representable (close inspection yields a Δ_2-representation) in U.

At first sight the argument uses quite some induction in extending V to V'. Miraculously enough, the whole argument can be adapted to S^1_2. The trick consists in replacing the use of induction by employing definable cuts as is explained in Section 1.3. We get the following theorem. With $\Box^J_U \varphi$ we shall denote that φ has a U-proof p with $p \in J$. Similarly we know how to read $\Diamond^J_U \varphi$.

Theorem 1.3.6. *For any numberizable theories U and V, we have that*

$$\mathsf{S}^1_2 \vdash \Box_U \mathsf{Con}(V) \to \exists k \ (k : U \rhd V \ \& \ \forall \varphi \ \Box_U(\Box_V \varphi \to \varphi^k)).$$

Proof. A proof can be found in [Vis91]. Actually something stronger is proved there. Namely, that for some standard number m we have

$$\forall \varphi \, \exists \, p {\leq} \omega_1^m(\varphi) \ \mathsf{Proof}_U(p, \Box_V \varphi \to \varphi^k).$$

⊣

As cuts have nice closure properties, many arguments can be performed within that cut. The numbers in the cut will so to say, play the role of the normal numbers. It turns out that the whole Henkin argument can be carried out using only the consistency on a cut.

We shall write $\Box_T^J\varphi$ for $\exists p{\in}J\ \mathsf{Proof}_T(p,\varphi)$. Thus, it is also clear what $\Diamond_T^J\varphi$ and $\mathsf{Con}^J(V)$ mean.

Theorem 1.3.7. *We have Theorem 1.3.6 also in the following form.*

$$T \vdash \Box_U \mathsf{Con}^I(V) \to \exists k\ (k : U \rhd V\ \&\ \forall\varphi\ \Box_U(\Box_V\varphi \to \varphi^k))$$

Here I is any (possibly non-standard) U-cut that is closed under $\omega_1(x)$.

Proof. By close inspection of the proof of Theorem 1.3.6. All operations on hypothetical proofs p can be bounded by some $\omega_1^k(p)$, for some standard k. As I is closed under $\omega_1(x)$, all the bounds remain within I. ⊣

We conclude this subsection with two asides, closely related to the Henkin construction.

Lemma 1.3.8. *Let U contain S^1_2. We have that $U \vdash \mathsf{Con}(\mathsf{Pred})$. Here, $\mathsf{Con}(\mathsf{Pred})$ is a natural formalization of the statement that predicate logic is consistent.*

Proof. By defining a simple (one-point) model within S^1_2. ⊣

Remark 1.3.9. If U has full induction, then it holds that $U \rhd V$ iff V is interpretable in U by some interpretation that maps identity to identity.

Proof. Suppose $j : U \rhd V$ with $j = \langle \delta, F\rangle$. We can define $j' := \langle \delta', F'\rangle$ with $\delta'(x) := \delta(x) \wedge \forall\, y{<}x\ (\delta(y) \to y{\neq^j}x)$. F' agrees with F on all symbols except that it maps identity to identity. By the minimal number principle we can prove $\forall x\ (\delta(x) \to \exists x'\ (x'{=^j}x) \wedge \delta'(x))$, and thus $\forall\vec{x}\ (\delta'(\vec{x}) \to (\varphi^j(\vec{x}) \leftrightarrow \varphi^{j'}(\vec{x})))$ for all formulae φ. ⊣

It is not the case that the implication in Remark 1.3.9 can be reversed. For, if U is reflexive, contains $\mathrm{I}\Delta_2$ (or $\mathrm{L}\Delta_2$) and $U \rhd V$, the following reasoning can be performed. By reflexivity of U (and the totality of exp), we get by Lemma 2.1.2 of the Orey-Hájek characterization that $\forall x\ \Box_U \mathsf{Con}_x(V)$. We can now perform the Henkin construction (Lemma 2.1.1). This yields an interpretation where all symbols of V get a Δ_2-translation. Thus, by $\mathrm{I}\Delta_2$ we can prove $\forall x\ (\delta(x) \to \exists x'\ (x'{=^j}x) \wedge \delta'(x))$ and obtain an interpretation that maps identity to identity. There exist plenty of reflexive extensions of $\mathrm{I}\Delta_2$ that do not contain full induction. An example is $\mathrm{I}\Sigma_3^R$.

1.3.3 Pudlák's lemma

Pudlák's lemma is central to many arguments in the field of interpretability logics. It provides a means to compare a model $\mathcal{M}$ of U and its internally defined model $\mathcal{M}^j$ of V if $j : U \rhd V$. If U has full induction, this comparison is fairly easy.

Theorem 1.3.10. *Suppose $j : U \rhd V$ and U has full induction. Let $\mathcal{M}$ be a model of U. We have that $\mathcal{M} \preceq_{\mathsf{end}} \mathcal{M}^j$ via a definable embedding.*

Proof. If U has full induction and $j : U \rhd V$, we may by Remark 1.3.9 actually assume that j maps identity in V to identity in U. Thus, we can define the following function.

$$f := \left\{ \begin{array}{l} 0 \mapsto 0^j \\ x+1 \mapsto f(x) {+^j} 1^j \end{array} \right.$$

Now, by induction, f can be proved to be total. Note that full induction is needed here, as we have a-priori no bound on the complexity of 0^j and $+^j$. Moreover, it can be proved that $f(a+b) = f(a){+^j} f(b)$, $f(a \cdot b) = f(a) \cdot^j f(b)$ and that $y {\leq^j} f(b) \rightarrow \exists\, a{<}b\ f(a) = y$. In other words, that f is an isomorphism between its domain and its co-domain and the co-domain is an initial segment of $\mathcal{M}^j$. $\dashv$

If U does not have full induction, a comparison between $\mathcal{M}$ and $\mathcal{M}^j$ is given by Pudlák's lemma, first explicitly mentioned in [Pud85]. Roughly, Pudlák's lemma says that in the general case, we can find a definable U-cut I of $\mathcal{M}$ and a definable embedding $f : I \longrightarrow \mathcal{M}^j$ such that $f[I] \preceq_{\mathsf{end}} \mathcal{M}^j$.

In formulating the statement we have to be careful as we can no longer assume that identity is mapped to identity. A precise formulation of Pudlák's lemma in terms of an isomorphism between two initial segments can for example be found in [JV00]. We have chosen here to formulate and prove the most general syntactic consequence of Pudlák's lemma, namely that I and $f[I]$, as substructures of $\mathcal{M}$ and $\mathcal{M}^j$ respectively, make true the same Δ_0-formulas.

In the proof of Pudlák's lemma we shall make the quantifier $\exists^{j,J\text{-}\mathsf{function}} h$ explicit. It basically means that h defines a function from a cut J to the $=^j$-equivalence classes of the numbers defined by the interpretation j.

Lemma 1.3.11 (Pudlák's Lemma).

$$\mathsf{S}^1_2 \vdash j : U \rhd V \rightarrow \exists^{U\text{-}\mathsf{Cut}} J\, \exists^{j,J\text{-}\mathsf{function}} h\, \forall^{\Delta_0} \varphi\ \Box_U \forall \vec{x} \in J\ (\varphi^j(h(\vec{x})) \leftrightarrow \varphi(\vec{x}))$$

Moreover, the h and J can be obtained uniformly from j by a function that is provably total in S^1_2.

Proof. Again, by $\exists^{U\text{-}\mathsf{Cut}} J$ we shall mean $\exists J\ \Box_U \mathsf{Cut}(J)$, where $\mathsf{Cut}(J)$ is the definable function that sends the code of a formula χ to the code of a formula that expresses that χ defines a cut. We apply a similar strategy for quantifying over j, J-functions. The defining property for a relation H to be a j, J-function is

$$\forall\, \vec{x}, y, y' {\in} J\ (H(\vec{x}, y)\ \&\ H(\vec{x}, y') \rightarrow y {=^j} y').$$

We will often consider H as a function and write for example $\psi(h(\vec{x}))$ instead of $\forall y\ (H(\vec{x}, y) \rightarrow \psi(y))$.

The idea of the proof is very easy. Just map the numbers of U via h to the numbers of V so that 0 goes to 0^j and the mapping commutes with the successor relation. If we want to prove a property of this mapping, we might run into problems as the intuitive proof appeals to induction. And sufficient induction is precisely what we lack in weaker theories.

The way out here is to just put all the properties that we need our function h to possess into its definition. Of course, then the work is in checking that we still have a good definition. The definition being good means here that the set of numbers on which h is defined induces a definable U-cut.

In a sense, we want an (definable) initial part of the numbers of U to be isomorphic under h to an initial part of the numbers of V. Thus, h should definitely commute with successor, addition and multiplication. Moreover, the image of h should define an initial segment, that is, be closed under the smaller than relation. All these requirements are reflected in the definition of $\mathsf{Goodsequence}$.

$$\begin{aligned}\mathsf{Goodsequence}(\sigma,x,y) \quad := \quad & \mathsf{lh}(\sigma)=x+1 \wedge \sigma_0{=^j}0^j \wedge \sigma_x{=^j}y\\ & \wedge \forall\, i{\leq}x\ \delta(\sigma_i)\\ & \wedge \forall\, i{<}x\ (\sigma_{i+1}{=^j}\sigma_i{+^j}1^j)\\ & \wedge \forall\, k{+}l{\leq}x\ (\sigma_k{+^j}\sigma_l{=^j}\sigma_{k+l})\\ & \wedge \forall\, k{\cdot}l{\leq}x\ (\sigma_k{\cdot^j}\sigma_l{=^j}\sigma_{k\cdot l})\\ & \wedge \forall a\ (a{\leq^j}y \rightarrow \exists\, i{\leq}x\ \sigma_i{=^j}a)\end{aligned}$$

$$\begin{aligned}H(x,y) \quad := \quad & \exists\sigma\ \mathsf{Goodsequence}(\sigma,x,y)\\ & \wedge \forall\sigma'\,\forall y'\ (\mathsf{Goodsequence}(\sigma',x,y')\rightarrow y{=^j}y')\end{aligned}$$

$$J'(x) := \forall\, x'{\leq}x\,\exists y\ H(x',y)$$

Finally, we define J to be the closure of J' under $+$, $\cdot$ and $\omega_1(x)$. Now that we have defined all the machinery we can start the real proof. The reader is encouraged to see at what place which defining property is used in the proof. We do note here that the defining property $\forall\, i{\leq}x\ \delta(\sigma_i)$ is not used in the proof here. We shall need it in the proof of Lemma 2.1.6.

We first note that $J'(x)$ indeed defines a U-cut. For $\Box_U J'(0)$ you basically need sequentiality of U, and the translations of the identity axioms and properties of 0.

To see $\Box_U\forall x\ (J'(x)\rightarrow J'(x+1))$ is also not so hard. It follows from the translation of basic properties provable in V, like $x=y\rightarrow x+1=y+1$ and $x+(y+1)=(x+y)+1$, etc.

We should now see that h is a j,J-function. This is actually quite easy, as we have all the necessary conditions present in our definition. Thus, we have

$$\Box_U\forall\, x,y{\in}J\ (h(x){=^j}h(y)\leftrightarrow x=y) \tag{1.2}$$

The $\leftarrow$ direction reflects that h is a j,J-function. The $\rightarrow$ direction follows from elementary reasoning in U using the translation of basic arithmetical facts provable in V. So, if $x \neq y$, say $x < y$, then $x+(z+1) = y$ whence $h(x){+^j}h(z+1){=^j}h(y)$ which implies $h(x){\neq^j}h(y)$.

We are now to see that for our U-cut J and for our j, J-function h we indeed have that[7]

$$\forall^{\Delta_0}\varphi\ \Box_U \forall\, \vec{x}{\in}J\ (\varphi^j(h(\vec{x})) \leftrightarrow \varphi(\vec{x})).$$

First we shall proof this using a seemingly Σ_1-induction. A closer inspection of the proof shall show that we can provide at all places sufficiently small bounds, so that actually an $\omega_1(x)$-induction suffices. We first proof the following claim.

Claim 1. $\forall^{\mathsf{Term}}t\ \Box_U \forall\, \vec{x}, y \in J\ (t^j(h(\vec{x})){=^j}h(y) \leftrightarrow t(\vec{x}) = y)$

Proof. The proof is by induction on t. The basis is trivial. To see for example $\Box_U \forall\, y{\in}J\ (0^j{=^j}h(y) \leftrightarrow 0 = y)$ we reason in U as follows. By the definition of h, we have that $h(0){=^j}0^j$, and by (1.2) we moreover see that $0^j{=^j}h(y) \leftrightarrow 0 = y$. The other basis case, that is, when t is an atom, is precisely (1.2).

For the induction step, we shall only do $+$, as $\cdot$ goes almost completely the same. Thus, we assume that $t(\vec{x}) = t_1(\vec{x}) + t_2(\vec{x})$ and set out to prove

$$\Box_U \forall\, \vec{x}, y{\in}J\ ({t_1}^j(h(\vec{x})){+^j}{t_2}^j(h(\vec{x})){=^j}h(y) \leftrightarrow t_1(\vec{x}) + t_2(\vec{x}) = y).$$

Within U:

- ← If $t_1(\vec{x}) + t_2(\vec{x}) = y$, then by Lemma 1.3.4, we can find y_1 and y_2 with $t_1(\vec{x}) = y_1$ and $t_2(\vec{x}) = y_2$. The induction hypothesis tells us that ${t_1}^j(h(\vec{x})){=^j}h(y_1)$ and ${t_2}^j(h(\vec{x})){=^j}h(y_2)$. Now by (1.2), $h(y_1+y_2){=^j}h(y)$ and by the definition of h we get that

$$\begin{array}{rll} h(y_1+y_2) & =^j & h(y_1){+^j}h(y_2) \\ & {=^j}_{\mathsf{i.h.}} & {t_1}^j(h(\vec{x})){+^j}{t_2}^j(h(\vec{x})) \\ & =^j & (t_1(h(\vec{x})) + t_2(h(\vec{x})))^j. \end{array}$$

- → Suppose now ${t_1}^j(h(\vec{x})){+^j}{t_2}^j(h(\vec{x})){=^j}h(y)$. Then clearly ${t_1}^j(h(\vec{x})){\leq^j}h(y)$ whence by the definition of h we can find some $y_1 \leq y$ such that ${t_1}^j(h(\vec{x})){=^j}h(y_1)$ and likewise for t_2 (using the translation of the commutativity of addition). The induction hypothesis now yields $t_1(\vec{x}) = y_1$ and $t_2(\vec{x}) = y_2$. By the definition of h, we get
$h(y){=^j}h(y_1){+^j}h(y_2){=^j}h(y_1+y_2)$, whence by (1.2), $y_1 + y_2 = y$, that is, $t_1(\vec{x}) + t_2(\vec{x}) = y$.

⊣

We now prove by induction on $\varphi \in \Delta_0$ that

$$\Box_U \forall\, \vec{x}{\in}J\ (\varphi^j(h(\vec{x})) \leftrightarrow \varphi(\vec{x})). \tag{1.3}$$

[7] We use $h(\vec{x})$ as short for $h(x_0), \cdots, h(x_n)$.

For the basis case, we consider that $\varphi \equiv t_1(\vec{x}) + t_2(\vec{x})$. We can now use Lemma 1.3.4 to note that

$$\Box_U \forall\, \vec{x}{\in}J\ (t_1(\vec{x}) = t_2(\vec{x}) \leftrightarrow \exists\, y{\in}J\ (t_1(\vec{x}) = y \wedge t_2(\vec{x}) = y))$$

and then use Claim 1, the transitivity of = and its translation to obtain the result.

The boolean connectives are really trivial, so we only need to consider bounded quantification. We show (still within U) that

$$\forall\, y, \vec{z}{\in}J\ (\forall\, x{\leq^j}h(y)\ \varphi^j(x, h(\vec{z})) \leftrightarrow \forall\, x{\leq}y\ \varphi(x, \vec{z})).$$

$\leftarrow$ Assume $\forall\, x{\leq}y\ \varphi(x, \vec{z})$ for some $y, \vec{z} \in J$. We are to show $\forall\, x{\leq^j}h(y)\ \varphi^j(x, h(\vec{z}))$. Now, pick some $x{\leq^j}h(y)$ (the translation of the universal quantifier actually gives us an additional $\delta(x)$ which we shall omit for the sake of readability). Now by the definition of h we find some $y' \leq y$ such that $h(y') = x$. As $y' \leq y$, by our assumption, $\varphi(y', \vec{z})$ whence by the induction hypothesis $\varphi^j(h(y'), h(\vec{z}))$, that is $\varphi^j(x, h(\vec{z}))$. As x was arbitrarily $\leq^j h(y)$, we are done.

$\rightarrow$ Suppose $\forall\, x{\leq^j}h(y)\ \varphi^j(x, h(\vec{z}))$. We are to see that $\forall\, x{\leq}y\ \varphi(x, \vec{z}))$. So, pick $x \leq y$ arbitrarily. Clearly $h(x){\leq^j}h(y)$, whence, by our assumption $\varphi^j(h(x), h(\vec{z}))$ and by the induction hypothesis, $\varphi(x, \vec{z})$.

In the proof of Lemma 1.3.11 we have used twice a Σ_1-induction; In Claim 1 and in proving (1.3). But in both cases, at every induction step, a constant piece p' of proof is added to the total proof. This piece looks every time the same. Only some parameters in it have to be replaced by subterms of t. So, the addition to the total proof can be estimated by $p'_a(t)$ which is about $\mathcal{O}(t^k)$ for some standard k. Consequently there is some standard number l such that

$$\forall\, \varphi{\in}\Delta_0\, \exists\, p{\leq}\varphi^l\ \mathsf{Proof}_U(p, \forall\, \vec{x}{\in}J\ (\varphi^j(h(\vec{x})) \leftrightarrow \varphi(\vec{x})))$$

and indeed, our induction was really but a bounded one. Note that we dealt with the bounded quantification by appealing to the induction hypothesis only once, followed by a generalization. So, fortunately we did not need to apply the induction hypothesis to all $x{\leq}y$, which would have yielded an exponential blow-up. ⊣

Remark 1.3.12. Pudlák's lemma is valid already if we employ the notion of theorems interpretability rather than smooth interpretability. If we work with theories in the language of arithmetic, we can do even better. In this case, axioms interpretability can suffice. In order to get this, all arithmetical facts whose translations were used in the proof of Lemma 1.3.11 have to be promoted to the status of axiom. However, a close inspection of the proof shows that these facts are very basic and that there are not so many of them.

If j is an interpretation with $j : \alpha \rhd \beta$, we shall sometimes call the corresponding isomorphic cut that is given by Lemma 1.3.11, the *Pudlák* cut of j and denote it by the corresponding upper case letter J.

Chapter 2

Characterizations of interpretability

In this chapter we shall relate the notion of relative interpretability to other notions, familiar in the context of meta-mathematics, like consistency assertions and Π_1-conservativity. Typically, these notions are formulated using arithmetic. Thus, our theories should be related to arithmetic too. In this section we employ two ways of relating our original theory U to arithmetic.

In the first section we do so by fixing some interpretation (numberization) j of S^1_2 in U. In the second section we use a map $\mho_{(\cdot)}$ assigning arithmetical theories $\mho_U$ to arbitrary theories U.

In Section 2.1 we are mainly concerned with the so-called Orey-Hájek characterizations of interpretability. We give detailed proofs and study the conditions needed in them. We shall work with theories as if they were formulated in the language of arithmetic. That is, we consider theories U with a fixed numberization $n : U \rhd \mathsf{S}^1_2$.

A disadvantage of doing so is clearly that our statements may be somehow misleading; when we think of, e.g., ZFC we do not like to think of it as coming with a fixed numberization.

On the other hand, there is the advantage of perspicuity and readability. For example, our notion of Π_1-conservativity refers to *arithmetical* Π_1-sentences and thus makes explicit use of some fixed interpretation.

In Section 2.2 we consider our map $\mho_U$ and study it as a functor between categories. In doing so, many characterizations get a more elegant formulation and proof. Our results have a direct bearing on the categories we study. In this subsection we shall be explicit about the numberizations used.

Finally, in Section 2.3 we give a model-theoretic characterization of interpretability.

$$\begin{tikzcd}
U \rhd V \arrow[rr, shift right, "2.1.2"'] \arrow[dr, shift right, "2.1.6"'] & & \forall n\ U \vdash \mathsf{Con}_n(V) \arrow[ll, shift right, "2.1.1"'] \arrow[dl, shift right, "2.1.3"'] \\
& \forall^{\forall\Pi_1^b}\pi\ (\Box_V \pi \rightarrow \Box_U \pi) \arrow[ul, shift right, "2.1.5"'] \arrow[ur, shift right, "2.1.4"'] &
\end{tikzcd}$$

Figure 2.1: Characterizations of interpretability

2.1 The Orey-Hájek characterizations

We consider the diagram from Figure 2.1. It is well known that all the implications hold when both U and V are reflexive. This fact is referred to as the Orey-Hájek characterizations ([Fef60], [Ore61], [Háj71], [Háj72]) for interpretability. However, for the Π_1-conservativity part, we should also mention work by Guaspari, Lindström and Pudlák ([Gua79], [Lin79], [Lin84], [Pud85]).

In this section we shall comment on all the implications in Figure 2.1, and study the conditions on U, V and the meta-theory, that are necessary or sufficient.

Lemma 2.1.1. *In* S^1_2 *we can prove* $\forall n\ \Box_U \mathsf{Con}_n(V) \rightarrow U \rhd V$.

Proof. The only requirement for this implication to hold, is that $U \vdash \mathsf{Con}(\mathsf{Pred})$. But, by our assumptions on U and by Lemma 1.3.8 this is automatically satisfied.

Let us first give the informal proof. Thus, let $\mathsf{Axiom}_V(x)$ be the formula that defines the axiom set of V.

We now apply a trick due to Feferman and consider the theory V' that consists of those axioms of V up to which we have evidence for their consistency. Thus, $\mathsf{Axiom}_{V'}(x) := \mathsf{Axiom}_V(x) \wedge \mathsf{Con}_x(V)$.

We shall now prove that $U \rhd V$ in two steps. First, we will see that

$$U \vdash \mathsf{Con}(V'). \tag{2.1}$$

Thus, by Theorem 1.3.5 we get that $U \rhd V'$. Second, we shall see that

$$V = V'. \tag{2.2}$$

To see (2.1), we reason in U, and assume for a contradiction that $\mathsf{Proof}_{V'}(p, \bot)$ for some proof p. We consider the largest axiom v that occurs in p. By assumption we have (in U) that $\mathsf{Axiom}_{V'}(v)$ whence $\mathsf{Con}_v(V)$. But, as clearly $V' \subseteq V$, we see that p is also a V-proof. We can now obtain a cut-free proof p' of $\bot$. Clearly $\mathsf{Proof}_{V,v}(p', \bot)$ and we have our contradiction.

If V' is empty, we cannot consider v. But in this case, $\mathsf{Con}(V') \leftrightarrow \mathsf{Con}(\mathsf{Pred})$, and by assumption, $U \vdash \mathsf{Con}(\mathsf{Pred})$.

We shall now see (2.2). Clearly $\mathbb{N} \models \mathsf{Axiom}_{V'}(v) \rightarrow \mathsf{Axiom}_V(v)$ for any $v \in \mathbb{N}$. To see that the converse also holds, we reason as follows.

Suppose $\mathbb{N} \models \mathsf{Axiom}_V(v)$. By assumption $U \vdash \mathsf{Con}_v(V)$, whence $\mathsf{Con}_v(V)$ holds on any model $\mathcal{M}$ of U. We now observe that $\mathbb{N}$ is an initial segment of (the numbers of) any model $\mathcal{M}$ of U, that is,

$$\mathbb{N} \preceq_{\mathsf{end}} \mathcal{M}. \tag{2.3}$$

As $\mathcal{M} \models \mathsf{Con}_v(V)$ and as $\mathsf{Con}_v(V)$ is a Π_1-sentence, we see that also $\mathbb{N} \models \mathsf{Con}_v(V)$. By assumption we had $\mathbb{N} \models \mathsf{Axiom}_V(v)$, thus we get that $\mathbb{N} \models \mathsf{Axiom}_{V'}(v)$. We conclude that

$$\mathbb{N} \models \mathsf{Axiom}_V(x) \leftrightarrow \mathsf{Axiom}_{V'}(x) \tag{2.4}$$

whence, that $V = V'$. As $U \vdash \mathsf{Con}(V')$, we get by Theorem 1.3.5 that $U \rhd V'$. We may thus infer the required $U \rhd V$.

It is not possible to directly formalize the informal proof. At (2.4) we concluded that $V = V'$. This actually uses some form of Π_1-reflection which is manifested in (2.3). The lack of reflection in the formal environment will be compensated by another sort of reflection, as formulated in Theorem 1.3.6.

Moreover, to see (2.1), we had to use a cut elimination. To avoid this, we shall need a sharper version of Feferman's trick.

Let us now start with the formal proof sketch. We shall reason in U. Without any induction we conclude $\forall x\ (\mathsf{Con}_x(V) \rightarrow \mathsf{Con}_{x+1}(V))$ or $\exists x\ (\mathsf{Con}_x(V) \wedge \Box_{V,x+1}\bot)$. In both cases we shall sketch a Henkin construction.

If $\forall x\ (\mathsf{Con}_x(V) \rightarrow \mathsf{Con}_{x+1}(V))$ and also $\mathsf{Con}_0(V)$, we can find a cut $J(x)$ with $J(x) \rightarrow \mathsf{Con}_x(V)$. We now consider the following non-standard proof predicate.

$$\Box^*_W\varphi := \exists\, x{\in}J\ \Box_{W,x}\varphi$$

We note that we have $\mathsf{Con}^*(V)$, where $\mathsf{Con}^*(V)$ of course denotes $\neg(\exists\, x{\in}J\ \Box_{V,x}\bot)$. As always, we extend the language on J by adding witnesses and define a series of theories in the usual way. That is, by adding more and more sentences (in J) to our theories while staying consistent (in our non-standard sense).

$$V = V_0 \subseteq V_1 \subseteq V_2 \subseteq \cdots \text{with } \mathsf{Con}^*(V_i) \tag{2.5}$$

We note that $\Box^*_{V_i}\varphi$ and $\Box^*_{V_i}\neg\varphi$ is not possible, and that for $\varphi \in J$ we can not have $\mathsf{Con}^*(\varphi \wedge \neg\varphi)$. These observations seem to be too trivial to make, but actually many a non-standard proof predicate encountered in the literature does prove the consistency of inconsistent theories.

As always, the sequence (2.5) defines a cut $I \subseteq J$, that induces a Henkin set W and we can relate our required interpretation k to this Henkin set as was, for example, done in [Vis91].

We now consider the case that for some fixed b we have $\mathsf{Con}_b(V) \wedge \Box_{V,b+1}\bot$. We note that we can see the uniqueness of this b without using any substantial induction. Basically, we shall now do the same construction as before only that we now possibly stop at b.

For example the cut $J(x)$ will now be replaced by $x \leq b$. Thus, we may end up with a truncated Henkin set W. But this set is complete with respect to relatively small formulas. Moreover, W is certainly closed under subformulas and substitution of witnesses. Thus, W is sufficiently large to define the required interpretation k.

In both cases we can perform the following reasoning.

$$\begin{array}{lll}
\Box_V\varphi & \rightarrow & \exists x\ \Box_{V,x}\varphi \\
& \rightarrow & \exists x\ \Box_U(\mathsf{Con}_x(V) \wedge \Box_{V,x}\varphi) \\
& \rightarrow & \Box_U\Box^*_V\varphi \\
& \rightarrow & \Box_U\varphi^k
\end{array}$$

The remarks from [Vis91] on the bounds of our proofs are still applicable and we thus obtain a smooth interpretation. ⊣

Lemma 2.1.2. *In the presence of* exp*, we can prove that for reflexive U, $U \rhd V \rightarrow \forall x\ \Box_U\mathsf{Con}_x(V)$.*

Proof. The informal argument is conceptually very clear and we have depicted it in Figure 2.2. The accompanying reasoning is as follows.

We assume $U \rhd V$, whence for some k we have $k : U \rhd V$. Thus, for axioms interpretability we find that $\forall u\,\exists p\ (\mathsf{Axiom}_V(u) \rightarrow \mathsf{Proof}_U(p, u^k))$. We are now to see that $\forall x\ U \vdash \mathsf{Con}_x(V)$. So, we fix some x. By our assumption we get that for some l, that

$$\forall\, u{\leq}x\,\exists p\ (\mathsf{Axiom}_V(u) \rightarrow \mathsf{Proof}_{U,l}(p, u^k)). \tag{2.6}$$

This formula is actually equivalent to the Σ_1-formula

$$\exists n\,\forall\, u{\leq}x\,\exists\, p{\leq}n\ (\mathsf{Axiom}_V(u) \rightarrow \mathsf{Proof}_{U,l}(p, u^k)) \tag{2.7}$$

In U:

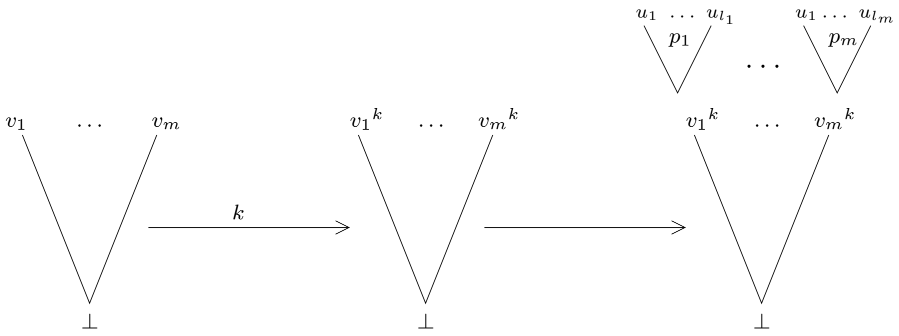
$u_1 \dots u_{l_1}$
p_1
$u_1 \dots u_{l_m}$
p_m
$\dots$
$v_1 \quad \dots \quad v_m$
k
${v_1}^k \quad \dots \quad {v_m}^k$
${v_1}^k \quad \dots \quad {v_m}^k$
$\bot$
$\bot$
$\bot$

Figure 2.2: Transformations on proofs

from which we may conclude by provable Σ_1-completeness,

$$U \vdash \exists n \, \forall \, u{\leq}x \, \exists \, p{\leq}n \; (\mathsf{Axiom}_V(u) \rightarrow \mathsf{Proof}_{U,l}(p, u^k)). \tag{2.8}$$

We now reason in U and suppose that there is some V, x-proof p of $\bot$. The assumptions in p are axioms $v_1 \dots v_m$ of V, with each $v_i \leq x$. Moreover, all the formulas ψ in p have $\rho(\psi) \leq x$. By Lemma 1.2.7, p transforms to a proof p^k of $\bot^k$ which is again $\bot$.

The assumptions in p^k are now among the ${v_1}^k \dots {v_m}^k$. By Remark 1.2.8 we get that for some n' depending on x and k, we have that all the axioms in p^k are $\leq n'$ and all the ψ occurring in p^k have $\rho(\psi) \leq n'$.

Now by (2.8), we have U, l-proofs $p_i \leq n$ of ${v_i}^k$. The assumptions in the p_i are axioms of U. Clearly all of these axioms are $\leq l$. We can now form a $U, l{+}n'$-proof p' of $\bot$ by substituting all the p_i for the $(v_i)^k$. Thus we have shown $\mathsf{Proof}_{U,l+n'}(p', \bot)$. But this clearly contradicts the reflexivity of U.

The informal argument is readily formalized to obtain $T \vdash U \rhd V \rightarrow \forall x \; \Box_U \mathsf{Con}(V, x)$. However there are some subtleties.

First of all, to conclude that (2.6) is equivalent to (2.7), a genuine application of $\mathrm{B}\Sigma_1$ is needed. If U lacks $\mathrm{B}\Sigma_1$, we have to switch to smooth interpretability to still have the implication valid. Smoothness then automatically also provides the l that we used in 2.6.

In addition we need that T proves the totality of exponentiation. For weaker theories, we only have provable $\exists\Sigma^b_1$-completeness. But if $\mathsf{Axiom}_V(u)$ is Δ^b_1, we can only guarantee that $\forall \, u{\leq}m \, \exists \, p{\leq}n \; (\mathsf{Axiom}_V(u) \rightarrow \mathsf{Proof}_U(p, u^k))$ is Π^b_2. As far as we know, exponentiation is needed to prove $\exists\Pi^b_2$-completeness.

All other transformations of objects in our proof only require the totality of $\omega_1(x)$. $\dashv$

The assumption that U is reflexive can in a sense not be dispensed with.

That is, if

$$\forall V\ (U \rhd V \to \forall x\ \Box_U \mathsf{Con}_x(V)), \tag{2.9}$$

then U is reflexive, as clearly $U \rhd U$. In a similar way we see that if

$$\forall U\ (U \rhd V \to \forall x\ \Box_U \mathsf{Con}_x(V)), \tag{2.10}$$

then V is reflexive. However, V being reflexive could never be a sufficient condition for (2.10) to hold, as we know from [Sha97] that interpreting reflexive theories in finitely many axioms is complete Σ_3.

Lemma 2.1.3. *In S^1_2 we can prove $\forall x\ \Box_U \mathsf{Con}_x(V) \to \forall^{\forall\Pi^b_1}\pi\ (\Box_V\pi \to \Box_U\pi)$.*

Proof. There are no conditions on U and V for this implication to hold. We shall directly give the formal proof as the informal proof does not give a clearer picture.

Thus, we reason in S^1_2 and assume $\forall x\ \Box_U \mathsf{Con}_x(V)$. Now we consider any $\pi \in \forall\Pi^b_1$ such that $\Box_V\pi$. Thus, for some x we have $\Box_{V,x}\pi$. We choose x large enough, so that we also have (see Remark 1.2.4)

$$\Box_U(\neg\pi \to \Box_{V,x}\neg\pi). \tag{2.11}$$

As $\Box_{V,x}\pi \to \Box_U\Box_{V,x}\pi$, we also have that

$$\Box_U\Box_{V,x}\pi. \tag{2.12}$$

Combining (2.11), (2.12) and the assumption that $\forall x\ \Box_U \mathsf{Con}_x(V)$, we see that indeed $\Box_U\pi$. ⊣

Lemma 2.1.4. *In S^1_2 we can prove that for reflexive V we have*

$$\forall^{\forall\Pi^b_1}\pi\ (\Box_V\pi \to \Box_U\pi) \to \forall x\ \Box_U \mathsf{Con}_x(V).$$

Proof. If V is reflexive and $\forall^{\forall\Pi^b_1}\pi\ (\Box_V\pi \to \Box_U\pi)$ then, as for every x, $\mathsf{Con}_{\overline{x}}(V)$ is a $\forall\Pi^b_1$-formula, also $\forall x\ \Box_U \mathsf{Con}_x(V)$. ⊣

It is obvious that

$$\forall U\ [\forall^{\forall\Pi^b_1}\pi\ (\Box_V\pi \to \Box_U\pi) \to \forall x\ \Box_U \mathsf{Con}_x(V)] \tag{2.13}$$

implies that V is reflexive. Likewise,

$$\forall V\ [\forall^{\forall\Pi^b_1}\pi\ (\Box_V\pi \to \Box_U\pi) \to \forall x\ \Box_U \mathsf{Con}_x(V)] \tag{2.14}$$

implies that U is reflexive. However, U being reflexive can never be a sufficient condition for (2.14) to hold. An easy counterexample is obtained by taking U to be PRA and V to be $\mathrm{I}\Sigma_1$. (See for example Chapter 11.)

Lemma 2.1.5. *(In S^1_2:) For reflexive V we have $\forall^{\forall\Pi^b_1}\pi\ (\Box_V\pi \to \Box_U\pi) \to U \rhd V$.*

Proof. We know of no direct proof of this implication. Also, all proofs in the literature go via Lemmata 2.1.4 and 2.1.1, and hence use reflexivity of V. $\dashv$

Again, by [Sha97] and Lemma 2.1.6 we see that U being reflexive can not be a sufficient condition for $\forall^{\forall\Pi_1^b}\pi\ (\Box_V\pi \rightarrow \Box_U\pi) \rightarrow U \rhd V$ to hold.

In our context, the reflexivity of V is not necessary, as $\forall U\ U \rhd \mathsf{S}^1_2$ and S^1_2 is not reflexive.

Lemma 2.1.6. *Let U be a reflexive and sequential theory. We have in S^1_2 that $U \rhd V \rightarrow \forall^{\forall\Pi_1^b}\pi\ (\Box_V\pi \rightarrow \Box_U\pi)$.*

If moreover $U \vdash \mathsf{exp}$ we also get $U \rhd V \rightarrow \forall^{\Pi_1}\pi\ (\Box_V\pi \rightarrow \Box_U\pi)$. If U is not reflexive, we still have that $U \rhd V \rightarrow \exists^{U\text{-}\mathsf{Cut}}J\,\forall^{\Pi_1}\pi\ (\Box_V\pi \rightarrow \Box_U\pi^J)$.

For these implications, it is actually sufficient to work with the notion of theorems interpretability.

Proof. The intuition for the formal proof comes from Pudlák's lemma, which in turn is tailored to compensate a lack of induction. We shall first give an informal proof sketch if U has full induction. Then we shall give the formal proof using Pudlák's lemma.

If U has full induction and $j : U \rhd V$, we may assume by Remark 1.3.9 assume that j maps identity to identity. By Theorem 1.3.10 we now see that $\mathcal{M} \preceq_{\mathsf{end}} \mathcal{M}^j$. If for some $\pi \in \Pi_1$, $\Box_V\pi$ then by soundness $\mathcal{M}^j \models \pi$, whence $\mathcal{M} \models \pi$. As $\mathcal{M}$ was arbitrary, we get by the completeness theorem that $\Box_U\pi$.

To transform this argument into a formal one, valid for weak theories, there are two major adaptations to be made. First, the use of the soundness and completeness theorem has to be avoided . This can be done by simply staying in the realm of provability. Secondly, we should get rid of the use of full induction. This is done by switching to a cut in Pudlák's lemma.

Thus, the formal argument runs as follows. Reason in T and assume $U \rhd V$.

We fix some $j : U \rhd V$. By Pudlák's lemma, Lemma 1.3.11, we now find[1] a definable U-cut J and a j, J-function h such that

$$\forall^{\Delta_0}\varphi\ \Box_U \forall\,\vec{x}{\in}J\ (\varphi^j(h(\vec{x})) \leftrightarrow \varphi(\vec{x})).$$

We shall see that for this cut J we have that

$$\forall^{\Pi_1}\pi\ (\Box_V\pi \rightarrow \Box_U\pi^J). \tag{2.15}$$

Therefore, we fix some $\pi \in \Pi_1$ and assume $\Box_V\pi$. Let $\varphi(x) \in \Delta_0$ be such that $\pi = \forall x\ \varphi(x)$. Thus we have $\Box_V\forall x\ \varphi(x)$, hence by theorems interpretability

$$\Box_U\forall x\ (\delta(x) \rightarrow \varphi^j(x)). \tag{2.16}$$

We are to see $\Box_U\forall x\ (J(x) \rightarrow \varphi(x))$. To see this, we reason in U and fix x such that $J(x)$. By definition of J, $h(x)$ is defined. By the definition of h, we have

[1] Remark 1.3.12 ensures us that we can find them also in the case of theorems interpretability.

$\delta(h(x))$, whence by (2.16), $\varphi^j(h(x))$. Pudlák's lemma now yields the desired $\varphi(x)$. As x was arbitrary, we have proved (2.16).

So far, we have not used the reflexivity of U. We shall now see that

$$\forall^{\forall\Pi_1^b}\pi\ (\Box_U\pi^J \to \Box_U\pi)$$

holds for any U-cut J whenever U is reflexive. For this purpose, we fix some $\pi \in \forall\Pi_1^b$, some U-cut J and assume $\Box_U\pi^J$. Thus, $\exists n\ \Box_{U,n}\pi^J$ and also $\exists n\ \Box_U\Box_{U,n}\pi^J$. If $\pi = \forall x\ \varphi(x)$ with $\varphi(x) \in \Pi_1^b$, we get $\exists n\ \Box_U\Box_{U,n}\forall x\ (x \in J \to \varphi(x))$, whence also

$$\exists n\ \Box_U\forall x\ \Box_{U,n}(x \in J \to \varphi(x)).$$

By Lemma 1.3.2 and Remark 1.3.3, for large enough n, this implies

$$\exists n\ \Box_U\forall x\ \Box_{U,n}\varphi(x)$$

and by Lemma 1.2.5 (only here we use that $\pi \in \forall\Pi_1^b$) we obtain the required $\Box_U\forall x\ \varphi(x)$. ⊣

U being reflexive and sequential is a sufficient condition for $U \rhd V \to \forall^{\forall\Pi_1^b}\pi\ (\Box_V\pi \to \Box_U\pi)$ to hold. For sequential (or even ℓ-reflexive, as defined in Subsection 2.2) theories, reflexivity is also a necessary condition. That is to say, that for such theories,

$$\forall V\ [U \rhd V \to \forall^{\forall\Pi_1^b}\pi\ (\Box_V\pi \to \Box_U\pi)], \tag{2.17}$$

implies that U is reflexive.[2] For, if U is sequential, we get by Lemma 2.2.2 that for every n, $U \rhd \mathsf{S}^1_2 + \mathsf{Con}_n(U)$. Thus, by (2.17) we get that $\forall n\ U \vdash \mathsf{Con}_n(U)$.

The sequentiality is essentially used here: we can exhibit a non-sequential non-reflexive U which satisfies (2.17).

By a result of Hanf [Han65] we can find a finitely axiomatized decidable theory T with $\mathrm{PA} + \mathsf{Con}(T) \vdash \mathsf{Con}(\mathrm{PA})$. We now let $U := \mathrm{PA} \boxplus T$ with the numberization corresponding to the PA-summand. Here, $\boxplus$ is the disjoint union as defined in Appendix A of [JV04a] and as studied in [MPS90]. We make two observations of U.

First, U satisfies (2.17). Suppose that $\mathrm{PA} \boxplus T \rhd V$ and $\Box_V\pi$. Then $V \rhd \mathsf{S}^1_2 + \pi$ whence $\mathrm{PA} \boxplus T \rhd \mathsf{S}^1_2 + \pi$. As we have pairing in $\mathsf{S}^1_2 + \pi$, we can apply Theorem A.5 from [JV04a] and obtain that $\mathrm{PA} \rhd \mathsf{S}^1_2 + \pi$ or $T \rhd \mathsf{S}^1_2 + \pi$. By the decidability of T and by the essentially undecidability of S^1_2, we see that $\mathrm{PA} \rhd \mathsf{S}^1_2 + \pi$. By Lemma 2.1.6 we conclude that $\mathrm{PA} \vdash \pi$, whence $\mathrm{PA} \boxplus T \vdash \pi$.

Second, we see that $\mathrm{PA} \boxplus T$ cannot be reflexive. Suppose, for a contradiction, that $\forall n\ \mathrm{PA} \boxplus T \vdash \mathsf{Con}_n((\mathrm{PA} \boxplus T))$. Then, for all n, $\mathrm{PA} \vdash \mathsf{Con}_n((\mathrm{PA} \boxplus T))$ and thus, for sufficiently large n, $\mathrm{PA} \vdash \mathsf{Con}_n(T)$. Since T is finitely axiomatizable and PA proves cut-elimination, $\mathrm{PA} \vdash \mathsf{Con}(T)$. But this would imply that $\mathrm{PA} \vdash \mathsf{Con}(\mathrm{PA})$ which is a contradiction.

[2] Note that the tempting fixed point $\varphi(\pi) \leftrightarrow (\mathsf{S}^1_2 + \forall^{\Pi_1}\pi\ \varphi(\pi) \rhd \mathsf{S}^1_2 + \pi \leftrightarrow \mathsf{True}_{\Pi_1}(\pi))$ also yields a reflexive (inconsistent) theory $\mathsf{S}^1_2 + \forall^{\Pi_1}\pi\ \varphi(\pi)$.

Again, by [Sha97] we note that V being reflexive can never be a sufficient condition for $\forall U\ [U \rhd V \rightarrow \forall^{\forall \Pi_1^b} \pi\ (\Box_V \pi \rightarrow \Box_U \pi)]$.

The main work on the Orey-Hájek characterization has now been done. We can easily extract some useful, mostly well-known corollaries.

Corollary 2.1.7. *If U is a reflexive theory, then*

$$T \vdash U \rhd V \leftrightarrow \forall x\ \Box_U \mathsf{Con}_x(V).$$

Here T contains exp *and $\rhd$ denotes smooth interpretability.*

Corollary 2.1.8. *(In S^1_2:) If V is a reflexive theory, then the following are equivalent.*

1. $U \rhd V$
2. $\exists^{U\text{-}\mathsf{Cut}} J\, \forall^{\Pi_1} \pi\ (\Box_V \pi \rightarrow \Box_U \pi^J)$
3. $\exists^{U\text{-}\mathsf{Cut}} J\, \forall x\ \Box_U \mathsf{Con}^J_x(V)$

Proof. This is part of Theorem 2.3 from [Sha97]. (1) ⇒ (2) is already proved in Lemma 2.1.6, (2) ⇒ (3) follows from the transitivity of V and (3) ⇒ (1) is a sharpening of Lemma 2.1.1. which closely follows Theorem 1.3.7. Note that $\rhd$ may denote denote smooth or theorems interpretability. ⊣

Corollary 2.1.9. *If V is reflexive, then*

$$\mathsf{S}^1_2 \vdash U \rhd_t V \leftrightarrow U \rhd_s V.$$

Proof. By Remark 1.3.12 and Corollary 2.1.8. ⊣

Corollary 2.1.10. *If U and V are both reflexive theories we have that the following are provably equivalent in S^1_2.*

1. $U \rhd V$
2. $\forall^{\forall \Pi_1^b} \pi\ (\Box_V \pi \rightarrow \Box_U \pi)$
3. $\forall x\ \Box_U \mathsf{Con}_x(V)$

Proof. If we go (1) ⇒ (2) ⇒ (3) ⇒ (1) we do not need the totality of exp that was needed for (1) ⇒ (3). ⊣

As an application we can, for example, see[3] that PA $\rhd$ PA + InCon(PA). It is well known that PA is essentially reflexive, so we use Corollary 2.1.10. Thus, it is sufficient to show that PA + InCon(PA) is Π_1-conservative over PA.

So, suppose that PA + InCon(PA) $\vdash \pi$ for some Π_1-sentence π. In other words PA $\vdash \Box\bot \rightarrow \pi$. We shall now see that PA $\vdash \Box\pi \rightarrow \pi$, which by Löb's Theorem gives us PA $\vdash \pi$.

Thus, in PA, assume $\Box\pi$. Suppose for a contradiction that $\neg\pi$. By Σ_1-completeness we also get $\Box\neg\pi$, which yields $\Box\bot$ with the assumption $\Box\pi$. But we have $\Box\bot \rightarrow \pi$ and we conclude π. A contradiction.

[3] By using **IL**W, we see that actually for all T we have $T \rhd T + \mathsf{InCon}(T)$.

2.2 Characterizations and functors

In this section, we rearrange the material to make it look more mathematical. We reformulate such notions as reflexivity in terms of functors between appropriate degree structures, viewed as pre-ordering categories. Theorems like the Orey-Hájek characterization receive a natural formulations in this framework. Also the precise relation between the Orey-Hájek and the Friedman characterization becomes fully perspicuous.

We shall work extensively with the notion of local interpretability. A theory U interprets a theory V locally if it interprets all of its finite subtheories. Again, in weak meta-theories there are various divergent notions possible.[4] In this section we will not worry about these subtle distinctions, assuming that our meta-theory is EA plus Σ_1-collection. Moreover, we shall always explicitly mention our numberizations.

Let THRY be the structure of theories ordered by $\subseteq$. Let DEG be the degrees of interpretability between theories. Let $\mathsf{DEG}^{\mathsf{loc}}$ be the degree structure of local interpretability. The ordering of DEG will be written as $U \lhd V$ or $U \longrightarrow V$. The ordering of $\mathsf{DEG}^{\mathsf{loc}}$ will be written as $U \lhd_{\mathsf{loc}} V$ or $U \xrightarrow{\mathsf{loc}} V$.

We will not divide out the preorders, treating the degree structures as preorder categories. If we want to restrict, e.g., DEG to a subclass of theories, we use a subscript to signal that. Consider a theory V. We define:

- $\mho_V := \mathsf{S}^1_2 + \{\mathsf{Con}_n(V) \mid n \in \omega\}$.
  (We pronounce this as 'mho of V', where 'mho' rhymes with 'Joe'.)

- $\mho^+_V := \mathsf{EA} + \{\mathsf{Con}_n(V) \mid n \in \omega\} = \mho_V + \mathsf{exp}$.

We will call the operation $V \mapsto \mho_V$: *reflexivization*. The name is motivated by Lemma 2.2.6, which says that $\mho_V$ is, in a sense, the smallest reflexive theory in which V is interpretable. We will later see that $\mho$ and $\mho^+$ give us functors between appropriate categories.

Lemma 2.1.1 expressed the following basic insight. Here, h_V denotes the 'Henkin interpretation' which is the syntactic variant of the Henkin model.

Theorem 2.2.1. *$h_V : \mho_V \rhd V$.*

In this subsection we distinguish three kinds of reflexivity.

- A numberized theory $\langle U, j\rangle$ is *reflexive* iff $j : U \rhd \mho_U$.

- A theory U is existentially reflexive, or *e-reflexive* iff, for some j, $j : U \rhd \mho_U$. (In other words, U is e-reflexive iff it has a reflexive numberization.)

- A theory V is locally reflexive, or *ℓ-reflexive* iff $V \rhd_{\mathsf{loc}} \mho_V$.

[4] Two such notions are $U \rhd_{\mathsf{loc},t} V :\iff \forall \phi\ (\Box_V \phi \to \exists k\ \Box_U \phi^k)$ and $U \rhd_{\mathsf{loc},s} V :\iff \forall x\, \exists y\, \forall\, \alpha{\in}\mathsf{axioms}_{V[x]}\ \exists p, k{<}y\ \mathsf{Proof}_U(p, \alpha^k)$.

By a sharpened version of the second incompleteness theorem, one can show that e-reflexive theories cannot be finitely axiomatizable. (This result is due to Pudlák, see [Pud85] or [Vis93].)

Sequential theories are important in our study. A useful feature of them is that they allow for truth predicates. Moreover, they are easily seen to interpret S^1_2, by taking for 0, e.g., the empty sequence, etc. Here is a basic insight concerning sequential theories.

Lemma 2.2.2. *Sequential theories are ℓ-reflexive. I.o.w., if V is sequential, then $V \rhd_{\mathsf{loc}} \mho_V$.*

Proof. Given some fixed $n \in \omega$, we are to interpret $\mathsf{S}^1_2 + \mathsf{Con}_n(V)$. Going from a V, n-proof p to a cut-free V, n-proof p' can cause a multi-exponential blow-up. However, the multi-exponent is linear in n (see [Ger03]).

By Solovay's techniques on shortening of cuts we can find a V-cut J for which this multi-exponent is always defined. Thus, for every proof p in J, there is a cut-free proof p'.

The idea is now to prove, by using the truth predicates, that at any step in p', a true formula is obtained. As always, we compensate a lack of induction in V by shortening J even further. ⊣

Note that if U is ℓ-reflexive, then so is $U \boxplus U$ (see Appendix A of [JV04a]). Since $U \boxplus U$ is not sequential, we see that there are non-sequential ℓ-reflexive theories.

2.2.1 Reflexivization as a Functor

In this subsection, we treat a basic insight (Theorem 2.2.3), from which, in combination with Theorem 2.2.1, many others will follow by simple semi-modal arguments. Theorem 2.2.3 also tells us that $\mho_{(\cdot)}$ can act as a functor between various categories.

Theorem 2.2.3. *Suppose $U \rhd_{\mathsf{loc}} V$. Then, $\mho_U \supseteq \mho_V$.*

Proof. Suppose $U \rhd_{\mathsf{loc}} V$. Consider any n. By assumption there is an interpretation j such that $j : U \rhd V{\restriction}n$, where $V{\restriction}n$ denotes the theory axiomatized by all axioms of V with Gödel numbers $\leq n$. So, $\forall\,\phi{\in}\alpha_{V,n}\,\exists r\ r{:}\Box_U\phi^j$. By Σ_1-collection, we can find a k, such that $\forall\,\phi{\in}\alpha_{V,n}\,\exists r{<}k\ r{:}\Box_U\phi^j$. Taking $m := |k|$, we find:[5] $\forall\,\phi{\in}\alpha_{V,n}\,\exists r{<}k\ r{:}\Box_{U,m}\phi^j$. Hence, by Σ_1-completeness, $\mho_U \vdash \forall\,\phi{\in}\alpha_{V,n}\,\exists r{<}k\ r{:}\Box_{U,m}\phi^j$.

Reason in $\mho_U$. Suppose $p : \Box_{V,n}\bot$. Using j, we can transform p into a proof $q : \Box_{U,m^\star}\bot$, where $m^\star$ is a sufficiently large standard number, depending only on n, j and k. Note that q exists because the proofs of the translations of axioms that we plug in are bounded by the standard number k. But $\mathsf{Con}_{m^\star}(U)$. Hence, $\mathsf{Con}_n(V)$. ⊣

[5] We see that a 'carefree meta-theory' for this argument should be something like EA plus Σ_1-collection. This is by a result of Slaman ([Sla04]) nothing but $\mathrm{I}\Delta_1$.

Theorem 2.2.3 tells us that reflexivization can be considered as a functor from $\mathsf{DEG}^{\mathsf{loc}}$ to THRY. It follows that also $\mho^+$ can also be considered as a functor from $\mathsf{DEG}^{\mathsf{loc}}$ to THRY.

It would be appropriate to call $\mho$ the Orey-Hájek functor and $\mho^+$ the Friedman functor, because of their connection (see below) to resp. the Orey-Hájek characterization and the Friedman characterization.

Note that Theorem 2.2.3 tells us a.o. that reflexivization can be considered as an operation that works on theories-as-sets-of-theorems. We may contrast this with an 'intensional' operation like[6] $U \mapsto \mathsf{S}^1_2 + \mathsf{Con}(U)$. We collect some immediate consequences of Theorem 2.2.3.

Lemma 2.2.4.
(i) e-Reflexiveness is preserved under mutual interpretability.
(ii) ℓ-Reflexiveness is preserved under mutual local interpretability.

Proof. Ad (i). Suppose U is e-reflexive and $U \equiv V$. Using Theorem 2.2.3, we find that: $V \equiv U \equiv \mho_U \equiv \mho_V$. Hence V is e-reflexive.

Ad (ii). Suppose U is ℓ-reflexive and $U \equiv_{\mathsf{loc}} V$. Using Theorem 2.2.1 and 2.2.3, we find that: $V \equiv_{\mathsf{loc}} U \equiv_{\mathsf{loc}} \mho_U \equiv \mho_V$. Hence V is ℓ-reflexive. ⊣

Lemma 2.2.5. *Suppose U is e-reflexive. Then, $U \rhd_{\mathsf{loc}} V$ iff $U \rhd V$.*

Proof. Suppose U is e-reflexive and $U \rhd_{\mathsf{loc}} V$. Then, $U \rhd \mho_U \rhd \mho_V \rhd V$. The other direction is (even more) trivial. ⊣

Lemma 2.2.6. *Suppose U is ℓ-reflexive. Then, $\langle \mho_U, \mathsf{id}\rangle$ is the smallest reflexive theory in DEG that interprets U.*

Proof. By Theorem 2.2.1 indeed $\mho_U \rhd U$. By the ℓ-reflexivity of U we get that $U \equiv_{\mathsf{loc}} \mho_U$, whence, by Theorem 2.2.3 $\mho_U = \mho_{\mho_U}$ and indeed $\langle \mho_U, \mathsf{id}\rangle$ is reflexive.

If for some reflexive T we have $T \rhd_{\mathsf{loc}} U$, we get by Lemma 2.2.5 that $T \rhd U$. By Theorem 2.2.3 we get $\mho_T \supseteq \mho_U$. But, using the reflexivity of T we now get $T \rhd \mho_T \supseteq \mho_U$ and we are done. ⊣

Note that the reflexivization of the theory of pure identity is S^1_2, which is finitely axiomatizable and, hence, not reflexive. Since $\mho_V$ is itself ℓ-reflexive, we always have that $\langle \mho_{\mho_V}, \mathsf{id}\rangle$ is reflexive.

Question 2.2.7. *Give an example of a U that is not ℓ-reflexive, but where $\langle \mho_U, \mathsf{id}\rangle$ is reflexive.*

Lemma 2.2.8. *Suppose U is ℓ-reflexive. Then, $U \rhd_{\mathsf{loc}} V$ iff $\mho_U \supseteq \mho_V$.*

[6] A similar observation is also made in [Sha97]. For example, let U be a theory in the language of pure identity. The non-logical axioms of U are given by α, where $\alpha(x)$ expresses: for some $n < |x|$, x is an an n-fold conjunction (associating to the right) of $\perp$ and n is the smallest ZF-proof of $\perp$. We see that U is, in a weak sense, finitely axiomatized, that U is co-extensional with the theory of pure identity, but that $\mathsf{S}^1_2 + \mathsf{Con}(U)$ is far stronger than S^1_2. Moreover, S^1_2 proves the consistency of the theory of pure identity.

Proof. Suppose $\mho_U \supseteq \mho_V$. Then, $U \rhd_{\mathsf{loc}} \mho_U \supseteq \mho_V \rhd V$. ⊣

Let $\mathsf{DEG}_{\mathsf{lr}}$ be the degrees of interpretability between ℓ-reflexive theories. Let $\mathsf{DEG}_{\mathsf{lr}}^{\mathsf{loc}}$ be the degree structure of local interpretability restricted to ℓ-reflexive theories.

Lemma 2.2.8 tells us that reflexivization is an embedding of $\mathsf{DEG}_{\mathsf{lr}}^{\mathsf{loc}}$ in THRY.

2.2.2 The Orey-Hájek Characterization

We can now reformulate the Orey-Hájek characterizations using local interpretability. All the characterizations are direct consequences of Theorem 2.2.3. Here is the first Orey-Hájek characterization.

Theorem 2.2.9 (Orey-Hájek 1).
Suppose $\langle U, j\rangle$ is reflexive. Then, $U \rhd V$ iff $j : U \rhd \mho_V$.

Note that the conclusion of this theorem, universally quantified over V, implies that $\langle U, j\rangle$ is reflexive. If V is e-reflexive, then $V \equiv \mho_V$. Thus, the following theorem is a triviality.

Theorem 2.2.10 (Orey-Hájek 2).
Suppose V is e-reflexive. Then, $U \rhd V$ iff $U \rhd \mho_V$.

Again, the conclusion of Orey-Hájek 2, universally quantified over U, is equivalent to the premise. It is by now folklore that we also have an Orey-Hájek characterization for local interpretability. It is contained in the following theorem.

Theorem 2.2.11. *For ℓ-reflexive U we have the following.*

$$\begin{array}{rcl} U \rhd_{\mathsf{loc}} V & \Leftrightarrow & \mho_U \supseteq \mho_V \\ & \Leftrightarrow & \mho_U \rhd V \\ & \Leftrightarrow & U \rhd_{\mathsf{loc}} \mho_V. \end{array}$$

As a corollary to Theorem 2.2.11, we shall now see that reflexivization can be viewed, modulo mutual relative interpretability, as the right adjoint of the embedding functor between the degrees of global interpretability of locally reflexive theories and the degrees of local interpretability of locally reflexive theories. Let us therefore single out one of the equivalences[7] of Theorem 2.2.11.

$$U \rhd_{\mathsf{loc}} V \Leftrightarrow \mho_U \rhd V \tag{2.18}$$

We reformulate this equivalence a bit, to make the adjunction fully explicit. We now treat, par abus de langage, $\mho$ as a functor from $\mathsf{DEG}_{\mathsf{lr}}^{\mathsf{loc}}$ to $\mathsf{DEG}_{\mathsf{lr}}$. Let emb

[7] Again we note that the conclusion of the equivalence, universally quantified over V, is equivalent with the premise, that is, the ℓ-reflexivity of U.

be the embedding functor of $\mathsf{DEG}_{\mathsf{lr}}$ in $\mathsf{DEG}^{\mathsf{loc}}_{\mathsf{lr}}$. Now (2.18) tells us that $\mho_{(\cdot)}$ is the right adjoint of emb. We may represent this fact in the following picture.

$$\frac{\mathsf{emb}(V) \xrightarrow{\mathsf{loc}} U}{V \xrightarrow[\mathsf{glob}]{} \mho_U}$$

It is immediate from general facts about adjoints that, for ℓ-reflexive U, $\mho_U \equiv \mho_{\mho_U}$. Note, however, that Lemma 2.2.6 is more informative.

2.2.3 Variants of $\mho$

As long as we are working modulo relative interpretability there are many interesting variants of $\mho$. In this subsection, we shall discuss three such variants.

Recall that $U[n]$ denotes the theory axiomatized by the first n axioms of U. Let $\Omega^\infty := \mathsf{S}^1_2 + \{\Omega_i \mid i \in \omega\}$, where Ω_i expresses the totality of the function $\omega_i(x)$ (see [HP93]). $\digamma_U$ is our first variant of $\mho_U$.

- $\digamma_U := \Omega^\infty + \{\mathsf{Cutfree\text{-}Con}(U[n]) \mid n \in \omega\}$.

We have the following lemma.

Lemma 2.2.12. $\digamma_U \equiv \mho_U$.

The essence of the proof is given in [Vis93], Subsection 3.2. Note that, for finitely axiomatized theories U, we have $\digamma_U = \Omega^\infty + \mathsf{Cutfree\text{-}Con}(U)$. Inspection of the results, gives $\Omega^\infty \equiv \mho_{\mathsf{S}^1_2}$. (See again [Vis93].) It follows that Ω^∞, being an e-reflexive theory, is not finitely axiomatizable.

We proceed to the next variant of $\mho_U$. Let U be sequential and suppose $j : U \rhd \mathsf{S}^1_2$. We define a new theory $\nabla_{U,j}$ as follows. We add a new unary predicate $\mathcal{I}$ to the language of U. The theory $\nabla_{U,j}$ is axiomatized by U plus the axiom $\mathsf{Cut}_j(\mathcal{I})$ plus all axioms of the form:

$$\mathsf{Cut}_j(A) \to \forall x\ (\mathcal{I}(x) \to A(x)),$$

where A is a U-formula having only x free. Clearly $U \equiv_{\mathsf{loc}} \nabla_{U,j}$.

Lemma 2.2.13. *Suppose $\langle U, j\rangle$ is a numberized theory and suppose that U is sequential. Then, $\nabla_{U,j} \equiv \mho_U$.*

Proof. We first see that $\nabla_{U,j} \rhd \mho_U$. By sequentiality of U, we can find for any n a $\langle U, j\rangle$-cut I such that $U \vdash \mathsf{Con}^I_n(U)$. We have $\nabla_{U,j} \vdash \mathcal{I} \subseteq I$. Hence, $\nabla_{U,j} \vdash \mathsf{Con}^{\mathcal{I}}_n(U)$. It follows that $\mathcal{I} : \nabla_{U,j} \rhd \mho_U$.

Conversely, by a simple compactness argument we find that $\mho_U \vdash \mho_{\nabla_{U,j}}$. Hence $\mho_U \rhd \nabla_{U,j}$. $\dashv$

Before we can give the third variant of $\mho_U$, we first have to agree on some notation. Let Γ be some set of sentences in the language of arithmetic. We define, for arbitrary U, and for $\langle V, j\rangle$ a numberized theory, the Γ-content of that theory as follows.

- $\mathsf{Cnt}_\Gamma(U) := \mathsf{S}^1_2 + \{\phi{\in}\Gamma \mid U \rhd (\mathsf{S}^1_2 + \phi)\}$
- $\mathsf{Cnt}_\Gamma(\langle V, j\rangle) := \mathsf{S}^1_2 + \{\phi{\in}\Gamma \mid j : V \rhd (\mathsf{S}^1_2 + \phi)\}$,

These definitions do not give us a bona fide theories with sufficiently simple axiomatizations. We can handle this problem by employing a variant of Craig's trick. (See for example Definition 4.2.6.)

Note that $\mathsf{Cnt}_\Gamma(U)$ might be inconsistent, where U is not. E.g., there is an Orey-sentence O which is of complexity Δ_2, such that $\mathsf{PA} \rhd (\mathsf{S}^1_2 + O)$ and $\mathsf{PA} \rhd (\mathsf{S}^1_2 + \neg O)$. So, $\mathsf{Cnt}_{\Sigma_2}(\mathsf{PA})$ is inconsistent.

Lemma 2.2.14. *If $U \rhd_{\mathsf{loc}} V$, then $\mathsf{Cnt}_\Gamma(U) \supseteq \mathsf{Cnt}_\Gamma(V)$.*

Lemma 2.2.14 tells us that $\mathsf{Cnt}_\Gamma(\cdot)$ is a functor from $\mathsf{DEG}^{\mathsf{loc}}$ to THRY.

Lemma 2.2.15.

1. $\mathsf{Cnt}_{\forall\Pi^b_1}(U) \subseteq \mho_U$.
2. *Suppose that U is ℓ-reflexive. Then,* $\mathsf{Cnt}_{\forall\Pi^b_1}(U) = \mho_U$.
3. $\mathsf{Cnt}_{\Pi_1}(U) \equiv \mathsf{Cnt}_{\forall\Pi^b_1}(U)$.

Proof. Ad (1). Suppose we have $j : U \rhd (\mathsf{S}^1_2 + \pi)$. We show that $\mho_U \vdash \pi$. We have, for sufficiently large n,

$$\begin{aligned} \mho_U \vdash \neg\pi &\to \Box_{U,n}\neg\pi^j \\ &\to \Box_{U,n}\bot \\ &\to \bot. \end{aligned}$$

So, indeed, $\mho_U \vdash \pi$.

Ad (2). Suppose U is ℓ-reflexive. Then, clearly, $\mho_U \subseteq \mathsf{Cnt}_{\forall\Pi^b_1}(U)$, since the $\mathsf{con}_n(U)$ are $\forall\Pi^b_1$.

Ad (3). We claim that there is a definable S^1_2-cut J, such that, for any Π_1-sentence π, there is a $\forall\Pi^b_1$-sentence $\pi^\star$, such that $\mathsf{S}^1_2 \vdash \pi \to \pi^\star$ and $\mathsf{S}^1_2 \vdash \pi^\star \to \pi^J$. Using this claim, we see that

$$\mathsf{id} : \mathsf{Cnt}_{\Pi_1}(U) \rhd \mathsf{Cnt}_{\forall\Pi^b_1}(U) \text{ and } J : \mathsf{Cnt}_{\forall\Pi^b_1}(U) \rhd \mathsf{Cnt}_{\Pi_1}(U).$$

The claim can be proved in a fancy way by invoking the formalization by Gaifman and Dimitracopoulos (see [GD82]) of Matijacevič's Theorem in EA aka $I\Delta_0 + \mathsf{exp}$.

A simpler argument is as follows. Suppose $\pi = \forall\vec{x}\ \pi_0\vec{x}$, where π_0 is Δ_0. Take $\pi^\star := \forall\vec{x}\ \pi_0(|\vec{x}|)$ and let J be some cut such that $\mathsf{S}^1_2 \vdash \forall z{\in}J\ 2^z{\downarrow}$. $\dashv$

2.2.4 The Friedman Functor

In this subsection we study the Friedman functor $\mho^+$.

Lemma 2.2.16. *Suppose U is ℓ-reflexive. Then, $\mho^+_U$ and $\mho_U$ prove the same $\forall\Pi^b_1$-sentences.*

Proof. Consider $\pi \in \forall\Pi^b_1$. Suppose $\mho^+_U \vdash \pi$, Then, for some n, $\mathsf{EA} + \mathsf{Con}_n(U) \vdash \pi$. By a results of Wilkie and Paris (see [WP87], or see [Vis90a] or [Vis92a]), we have, for some cut J, $\mathsf{S}^1_2 + \mathsf{con}_n(U) \vdash \pi^J$. Let $k : U \rhd (\mathsf{S}^1_2 + \mathsf{con}_n(U))$. We have, for sufficiently large m,

$$\begin{aligned}
\mho_U \vdash \neg\pi \quad &\to \quad \Box_{U,m}(\neg\,\pi^J)^k \\
&\to \quad \Box_{U,m}(\Box_{U,n}\bot)^k \\
&\to \quad \Box_{U,m}\bot \\
&\to \quad \bot.
\end{aligned}$$

Hence, $\mho_U \vdash \pi$. $\dashv$

Lemma 2.2.17. *Suppose U is ℓ-reflexive. Then,*

$$U \rhd_{\mathsf{loc}} V \Leftrightarrow \mho^+_U \supseteq \mho^+_V.$$

Proof. Suppose U is ℓ-reflexive. It is sufficient to show that, if $\mho^+_U \supseteq \mho^+_V$, then $\mho_U \supseteq \mho_V$. But this is immediate by Lemma 2.2.16. $\dashv$

By cut elimination (see [Ger03]) we can show that over EA we may replace the $\mathsf{Con}_n(U)$ by $\mathsf{Cutfree\text{-}Con}(U[n])$.

In case U is finitely axiomatized, we have the following simplifications.

- U is ℓ-reflexive iff $U \rhd (\mathsf{S}^1_2 + \mathsf{Cutfree\text{-}Con}(U))$.
- $\mho^+_U = \mathsf{EA} + \mathsf{Cutfree\text{-}Con}(U)$.

Putting things together, we get a version of the Friedman characterization.

Theorem 2.2.18. *Suppose U and V are finitely axiomatized and ℓ-reflexive, then:*

$$U \rhd V \Leftrightarrow \mathsf{EA} + \mathsf{Cutfree\text{-}Con}(U) \vdash \mathsf{Cutfree\text{-}Con}(V).$$

Wilkie and Paris show in [WP87] that $\mathsf{EA} \vdash \mathsf{Cutfree\text{-}Con}(\mathsf{S}^1_2)$. It follows that $\mathsf{EA} = \mho^+_{\mathsf{S}^1_2}$. Hence, $\mathsf{EA} \equiv_{\forall\Pi^b_1} \mho_{\mathsf{S}^1_2}$. (This is approximately Theorem 8.15 of [WP87].) Thus, if we 'measure' the complexity of theories using the Friedman functor, then S^1_2 is of the lowest complexity.

2.3 End-extensions

We have a model-theoretic characterization of interpretability between extensions of PA in the language of PA(see Theorem 1.3.10). It is simply that $U \rhd V$ iff every model $\mathcal{M}$ of U has an endextension $\mathcal{N}$ satisfying V. In this section we generalize this result as far as possible.

It seems to us that the rules of the game are to formulate the characterization as much as possible in terms of the structure of the models without mentioning syntax. In this respect our result is not quite perfect, since we have to mention a definable inner model.

Consider a model $\mathcal{M}$ of signature Σ and a model $\mathcal{N}$ of signature Θ. Suppose m is a relative interpretation such that $\mathcal{M}^m \models \mathsf{S}^1_2$. We say that $\mathcal{N}$ *is an m-end-extension of* $\mathcal{M}$, or $m : \mathcal{M} \preceq_{\mathsf{end}} \mathcal{N}$, iff, for all relative interpretations n with $\mathcal{N}^n \models \mathsf{S}^1_2$, there is an initial embedding of $\mathcal{M}^m$ in $\mathcal{N}^n$. We say that $\mathcal{N}$ is *an end-extension of* $\mathcal{M}$ or $\mathcal{M} \preceq_{\mathsf{end}} \mathcal{N}$ iff, for some m, $m : \mathcal{M} \preceq_{\mathsf{end}} \mathcal{N}$.

Here are some basic facts on m-end-extensions.

1. If $m : \mathcal{M} \preceq_{\mathsf{end}} \mathcal{N}$ and $n : \mathcal{N} \preceq_{\mathsf{end}} \mathcal{K}$, then $m : \mathcal{M} \preceq_{\mathsf{end}} \mathcal{K}$.

2. If $\mathcal{M}^m$ satisfies full induction in $\mathcal{M}$, then $m : \mathcal{M} \preceq_{\mathsf{end}} \mathcal{M}$. (We do not know whether the converse holds.) Moreover, any internal model $\mathcal{N}$ of $\mathcal{M}$ is an m-end-extension of $\mathcal{M}$.

Theorem 2.3.1. *Suppose U is sequential and e-reflexive. We can find an $m : U \rhd \mathsf{S}^1_2$ such that, for any ℓ-reflexive V, the following are equivalent:*

1. *$U \rhd V$;*

2. *for all $\mathcal{M} \in \mathsf{Mod}(U)$, there is an $\mathcal{N} \in \mathsf{Mod}(V)$ such that $m : \mathcal{M} \preceq_{\mathsf{end}} \mathcal{N}$;*

3. *there is an $l : U \rhd \mathsf{S}^1_2$ such that, for all $\mathcal{M} \in \mathsf{Mod}(U)$, there is an $\mathcal{N} \in \mathsf{Mod}(V)$ such that $l : \mathcal{M} \preceq_{\mathsf{end}} \mathcal{N}$.*

Proof. Suppose U is sequential and e-reflexive. We first find m. Pick any $l : U \rhd \mathsf{S}^1_2$. By Lemma 2.2.13, we can find $p : U \rhd \nabla_{U,l}$. Recall that $\nabla_{U,l}$ contains U. We take $m := p \circ \mathcal{I}$.

(1) ⇒ (2). Suppose $k : U \rhd V$. We can 'lift' k in the obvious way to $k^\star : \nabla_{U,l} \rhd V$. We consider $q := p \circ k^\star : U \rhd V$. Let $\mathcal{M} \in \mathsf{Mod}(U)$ be given. We take $\mathcal{N} := \mathcal{M}^q$. Suppose that for some interpretation n, $\mathcal{N}^n \models \mathsf{S}^1_2$. We want to show that there is an initial embedding from $\mathcal{M}^m$ to $\mathcal{N}^n$.

We will use Figure 2.3 to support our argument. Let us first resume the list of interpretations that will be used in our proof.

$$\begin{array}{rcl@{\qquad}rcl}
l & : & U \rhd \mathsf{S}^1_2 & p & : & U \rhd \nabla_{U,l} \\
\mathcal{I} & : & \nabla_{U,l} \rhd \mathsf{S}^1_2 & m := p \circ \mathcal{I} & : & U \rhd \mathsf{S}^1_2 \\
k & : & U \rhd V & k^\star & : & \nabla_{U,l} \rhd V \\
q := p \circ k^\star & : & U \rhd V & & &
\end{array}$$

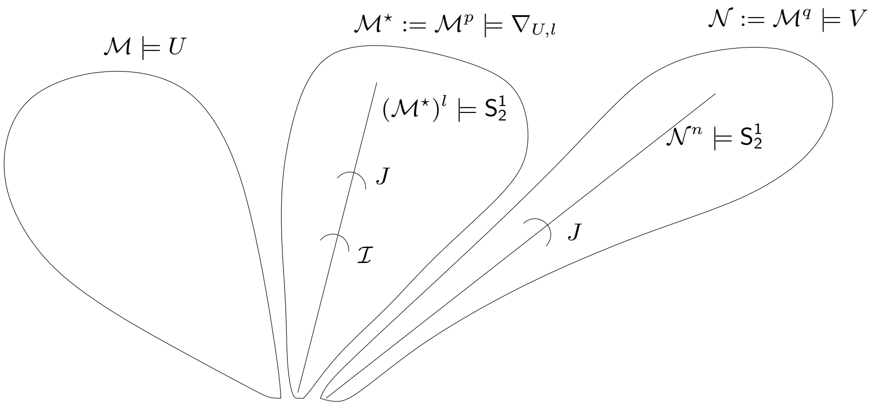
$\mathcal{M} \models U$
$\mathcal{M}^\star := \mathcal{M}^p \models \nabla_{U,l}$
$\mathcal{N} := \mathcal{M}^q \models V$
$(\mathcal{M}^\star)^l \models \mathsf{S}^1_2$
$\mathcal{N}^n \models \mathsf{S}^1_2$
J
$\mathcal{I}$
J

Figure 2.3: End extension theorem

Now we consider the internal model $\mathcal{M}^\star := \mathcal{M}^p$ of $\mathcal{M}$. We note that: $\mathcal{M}^\star \models \nabla_{U,l}$. Our main characters $\mathcal{N}$, $\mathcal{M}^m$ and $\mathcal{N}^n$ exist as internal models of $\mathcal{M}^\star$:

- $\mathcal{N} = \mathcal{M}^q = \mathcal{M}^{(p \circ k^\star)} = (\mathcal{M}^p)^{k^\star} = (\mathcal{M}^\star)^{k^\star}$,
- $\mathcal{M}^m = \mathcal{M}^{(p \circ \mathcal{I})} = (\mathcal{M}^p)^{\mathcal{I}} = (\mathcal{M}^\star)^{\mathcal{I}}$,
- $\mathcal{N}^n = ((\mathcal{M}^\star)^{k^\star})^n = (\mathcal{M}^\star)^{(k^\star \circ n)}$

So we only have to show that there is an initial embedding from $(\mathcal{M}^\star)^{\mathcal{I}}$ to $(\mathcal{M}^\star)^{(k^\star \circ n)}$.

Let us consider $(\mathcal{M}^\star)^l$ and $(\mathcal{M}^\star)^{(k^\star \circ n)}$. Since U is sequential and $\mathcal{M}^\star \models U$, we can, with Pudlák's lemma, find a definable cut J of $(\mathcal{M}^\star)^l$ isomorphic to a cut of $(\mathcal{M}^\star)^{(k^\star \circ n)}$. Both $k^\star$ and l only involve the language of U, so we find that J is given by a U-formula. Hence, by the definition of $\mathcal{I}$, we have that $(\mathcal{M}^\star)^{\mathcal{I}}$ is a cut of J. We may conclude that there is an initial embedding from $(\mathcal{M}^\star)^{\mathcal{I}}$ to $(\mathcal{M}^\star)^{(k^\star \circ n)}$.

(2) $\Rightarrow$ (3). This one is trivial.

(3) $\Rightarrow$ (1). Suppose that V is ℓ-reflexive, $l : U \rhd \mathsf{S}^1_2$, and, for all $\mathcal{M} \in \mathsf{Mod}(U)$, there is an $\mathcal{N} \in \mathsf{Mod}(V)$ such that $l : \mathcal{M} \preceq_{\mathsf{end}} \mathcal{N}$. For any $n \in \omega$, there is a $k : V \rhd \mathsf{S}^1_2$ such that $V \vdash \mathsf{Con}^k_n(V)$. Since $\mathcal{M}^l$ has an initial embedding in $\mathcal{N}^k$, it follows that $\mathcal{M}^l \models \mathsf{Con}_n(V)$. Since $\mathcal{M}$ was arbitrary, we have, by the completeness theorem, that $U \vdash \mathsf{Con}^l_n(V)$. Hence $l : U \rhd \mho_V$, and, thus $U \rhd V$. $\dashv$

Chapter 3

Interpretability logics

One possible way to study interpretability is by means of modal logics. With such an approach we can capture a large part of the structural behavior of interpretations.[1] Let us consider such a structural rule.

For any theories U, V and W we have that, if $U \rhd V$ and $V \rhd W$, then also $U \rhd W$. It is not hard to catch this in a modal logic. But modal logics talk about propositions and interpretability talks about theories.

It does not seem to be a good idea to directly translate propositional variables to theories. For what does the negation of a theory mean? And how to read implication? And how to translate modal statements involving iterated modalities?

The usual way to relate modal logics to interpretability is to translate propositional variables to arithmetical sentences that are added to some base theory T. Of course, the meta-theory should be strong enough to allow for arithmetization. As we shall see, by this approach, we get quite an expressive formalism in which the logic of provability is naturally embedded.

We shall work with a modal language containing two modalities, a unary modality $\Box$ and a binary modality $\rhd$. As always, we shall use $\Diamond A$ as short for $\neg\Box\neg A$. Apart from propositional variables we also have two constants $\top$ and $\bot$ in our language.

In this dissertation we thus use the same symbol $\rhd$ both for formalized interpretability and for our binary modal operator. The same holds for $\Box$. But the context will always decide on how to read the symbol.

Definition 3.0.2. An arithmetical T-realization is a map $*$ sending propositional variables p to arithmetical sentences p^*. The realization $*$ is extended to a map that is defined on all modal formulae as follows.

It is defined to commute with all boolean connectives. Moreover $(A \rhd B)^* = (T \cup \{A^*\}) \rhd (T \cup \{B^*\})$ (we shall write $A^* \rhd_T B^*$) and $(\Box A)^* = \Box_T A^*$. Here $\rhd_T$ and $\Box_T$ denote the formulas expressing formalized interpretability and formalized provability respectively, over T, as defined in Section 1.2.

[1] Some pioneering work on this, is in [Šve83] and [Háj81].

We shall reserve the symbol $*$ to range over T-realizations. Moreover, we will speak just of realizations if the T is clear from the context. In the literature realizations are also referred to as interpretations or translations. As these words are already reserved for other notions in our paper, we prefer to talk of realizations.

Definition 3.0.3. A modal formula A is an *interpretability principle* of a theory T, if $\forall * \, T \vdash A^*$. The *interpretability logic* of a theory T, we write $\mathbf{IL}(\mathrm{T})$, is the set of all the interpretability principles of T or a logic that generates it.

Likewise, we can talk of the set of all provability principles of a theory T, denoted by $\mathbf{PL}(\mathrm{T})$. Since the famous result by Solovay, $\mathbf{PL}(\mathrm{T})$ is known for a large class of theories T. (Below we will define **GL**.)

Theorem 3.0.4 (Solovay [Sol76]). $\mathbf{PL}(\mathrm{T}) = \mathbf{GL}$ *for any* Σ_1*-sound theory* T *containing* exp.

Definition 3.0.5. The interpretability logic of all reasonable arithmetical theories, we write $\mathbf{IL}(\mathrm{All})$, is the set of formulas φ such that $\forall T \, \forall * \;\; T \vdash \varphi^*$. Here the T ranges over all the reasonable arithmetical theories.

3.1 The logic IL

The logic **IL** that we shall present below, is a sort of core logic. It is contained in all other interpretability logics that we shall consider. We shall see that $\mathbf{IL} \subset \mathbf{IL}(\mathrm{T})$ for any reasonable T.

In writing formulas we shall omit brackets that are superfluous according to the following reading conventions. We say that the operators $\Diamond$, $\Box$ 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 $\rightarrow$ and $\leftrightarrow$. We shall also omit outer brackets. Thus, we shall write $A \rhd B \rightarrow A \wedge \Box C \rhd B \wedge \Box C$ instead of $((A \rhd B) \rightarrow ((A \wedge (\Box C)) \rhd (B \wedge (\Box C))))$.

A schema of interpretability logic is syntactically like a formula. They are used to generate formulae that have a specific form. We will not be specific about the syntax of schemata as this is similar to that of formulas. Below, one can think of A, B and C as place holders.

The rule of Modus Ponens allows one to conclude B from premises $A \rightarrow B$ and A. The rule of Necessitation allows one to conclude $\Box A$ from the premise A.

Definition 3.1.1. The logic **IL** is the smallest set of formulas being closed under the rules of Necessitation and of Modus Ponens, that contains all tautological formulas and all instantiations of the following axiom schemata.

L1 $\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B)$

L2 $\Box A \rightarrow \Box\Box A$

L3 $\Box(\Box A \to A) \to \Box A$

J1 $\Box(A \to B) \to A \rhd B$

J2 $(A \rhd B) \wedge (B \rhd C) \to A \rhd C$

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

J4 $A \rhd B \to (\Diamond A \to \Diamond B)$

J5 $\Diamond A \rhd A$

Sometimes we will write $\mathbf{IL} \vdash \varphi \to \psi \to \chi$ as short for $\mathbf{IL} \vdash \varphi \to \psi$ & $\mathbf{IL} \vdash \psi \to \chi$. Similarly for $\rhd$. We adhere to a similar convention when we employ binary relations. Thus, $xRyS_xz \Vdash B$ is short for xRy & yS_xz & $z \Vdash B$, and so on.

Sometimes we will consider the part of **IL** that does not contain the $\rhd$-modality. This is the well-known provability logic **GL**, whose axiom schemata are L1-L3. The axiom schema L3 is often referred to as Löb's axiom.

Some elementary reasoning in **IL** is captured in the following lemma.

Lemma 3.1.2.

1. $\mathbf{IL} \vdash \Box A \leftrightarrow \neg A \rhd \bot$

2. $\mathbf{IL} \vdash A \rhd A \wedge \Box\neg A$

3. $\mathbf{IL} \vdash A \vee \Diamond A \rhd A$

Proof. All of these statements have very easy proofs. We give an informal proof of the second statement. Reason in **IL**. It is easy to see $A \rhd (A \wedge \Box\neg A) \vee (A \wedge \Diamond A)$. By L3 we get $\Diamond A \to \Diamond(A \wedge \Box\neg A)$. Thus, $A \wedge \Diamond A \rhd \Diamond(A \wedge \Box\neg A)$ and by J5 we get $\Diamond(A \wedge \Box\neg A) \rhd A \wedge \Box\neg A$. As certainly $A \wedge \Box\neg A \rhd A \wedge \Box\neg A$ we have that $(A \wedge \Box\neg A) \vee (A \wedge \Diamond A) \rhd A \wedge \Box\neg A$ and the result follows from transitivity of $\rhd$. ⊣

We shall now briefly argue that all the axioms of **IL** are indeed sound. That is, we shall see that they are provable in any theory under any realization.

The principles L_1-L_3 are the familiar provability conditions. They are well known to hold (be sound) in S^1_2. The principle J_1 is easy to prove by taking the identity translation.

To see the soundness of J_2, we should describe how we can code the composition of two interpretations into a single interpretation. Let $k : U \rhd V$ and $j : V \rhd W$ with $k := \langle \delta_k, F_k \rangle$ and $j := \langle \delta_j, F_j \rangle$. We define $k \circ j$ to be $\langle \delta_{k \circ j}, F_{k \circ j} \rangle$ with

- $\delta_{k \circ j} := \delta_k \wedge (\delta_j)^k$,

- $F_{k \circ j}(R) := (F_j(R))^k$.

By an easy formula induction, we now see that $\mathsf{S}^1_2 \vdash (\varphi^j)^k \leftrightarrow \varphi^{k\circ j}$ and we are done.

To see the soundness of J_3, we reason as follows. We suppose that $j : \alpha \rhd_T \gamma$ and $k : \beta \rhd_T \gamma$. We need to construct an interpretation $j \vee k$ that uses the translation of j in case α and the translation of k otherwise. We thus define

- $\delta_{j\vee k} := (\delta_j \wedge \alpha) \vee (\delta_k \wedge \neg\alpha)$,
- $F_{j\vee k}(R) := (F_j(R) \wedge \alpha) \vee (F_k(R) \wedge \neg\alpha)$.

We note that $j \vee k$ can be very different from $k \vee j$. Again by easy formula induction we now see that $\mathsf{S}^1_2 \vdash \varphi^{j\vee k} \leftrightarrow (\alpha \wedge \varphi^j) \vee (\neg\alpha \wedge \varphi^k)$ and we are done.

J_4 is very easy. For, if $j : \alpha \rhd_T \beta$, we certainly have that $\Box_T(\alpha \rightarrow \beta^j)$. If now $\Box_T\neg\beta$ then $\Box_T(\alpha \rightarrow \neg\beta^j)$ and we get $\Box_T\neg\alpha$.

The only principle of **IL** that needs some serious argument is J_5. In proving the soundness of J_5, thinking about interpretability in terms of uniform model constructions yields the right heuristics. If we know the consistency of $T+\alpha$, we should be able to construct, in a uniform way, a model of $T + \alpha$. This uniform construction is just the Henkin construction. Theorem 1.3.6 tells us that indeed we have access to the Henkin construction.

3.2 More logics

The interpretability logic **IL** is a sort of basic interpretability logic. All other interpretability logics we consider shall be extensions of **IL** with further principles. Principles we shall consider in this paper are amongst the following.[2]

$$\begin{array}{lll}
\mathsf{F} & := & A \rhd \Diamond A \rightarrow \Box\neg A \\
\mathsf{W} & := & A \rhd B \rightarrow A \rhd B \wedge \Box\neg A \\
\mathsf{M_0} & := & A \rhd B \rightarrow \Diamond A \wedge \Box C \rhd B \wedge \Box C \\
\mathsf{W^*} & := & A \rhd B \rightarrow B \wedge \Box C \rhd B \wedge \Box C \wedge \Box\neg A \\
\mathsf{P_0} & := & A \rhd \Diamond B \rightarrow \Box(A \rhd B) \\
\mathsf{R} & := & A \rhd B \rightarrow \neg(A \rhd \neg C) \rhd B \wedge \Box C \\
\mathsf{R^*} & := & A \rhd B \rightarrow \neg(A \rhd \neg C) \rhd B \wedge \Box C \wedge \Box\neg A \\
\mathsf{M} & := & A \rhd B \rightarrow A \wedge \Box C \rhd B \wedge \Box C \\
\mathsf{P_R} & := & A \rhd B \rightarrow \Box(\Diamond A \rightarrow \Diamond B) \\
\mathsf{P} & := & A \rhd B \rightarrow \Box(A \rhd B)
\end{array}$$

If X is a set of axiom schemata we will denote by **IL**X the logic that arises by adding the axiom schemata in X to **IL**. Thus, **IL**X is the smallest set of formulas being closed under the rules of Modus Ponens and Necessitation and containing all tautologies and all instantiations of the axiom schemata of **IL** (L1-J5) and of the axiom schemata of X. For a schema Y, we write **IL**$\mathsf{X} \vdash \mathsf{Y}$ if **IL**X proves every instantiation of Y.

[2] In [GJ04] the modal principle $A \rhd B \rightarrow \neg(A \rhd \neg C) \wedge (D \rhd C) \rhd B \wedge \Box C$ was called R. This principle and the one called R here, are easily seen to be equivalent over **IL**.

A central theme in the study of formalized interpretability is to determine $\mathbf{IL}(\mathrm{T})$ for a specific T. For two classes of theories, $\mathbf{IL}(\mathrm{T})$ is known.

Definition 3.2.1. A theory T is essentially reflexive if all of its finite sentential extensions are reflexive.

Theorem 3.2.2 (Berarducci [Ber90], Shavrukov [Sha88]). *If T is an essentially reflexive theory, then $\mathbf{IL}(\mathrm{T}) = \mathbf{IL}\mathsf{M}$.*

Theorem 3.2.3 (Visser [Vis90a]). *If T is finitely axiomatizable and contains* supexp*, then $\mathbf{IL}(\mathrm{T}) = \mathbf{IL}\mathsf{P}$.*

Definition 3.2.4. The interpretability logic of all reasonable numberized theories –we write $\mathbf{IL}(\mathrm{All})$– is the set of formulas φ such that $\forall T\, \forall *\ T \vdash \varphi^*$. Here the T ranges over all numberized theories.

For sure $\mathbf{IL}(\mathrm{All})$ should be in the intersection of $\mathbf{IL}\mathsf{M}$ and $\mathbf{IL}\mathsf{P}$. Up to now, $\mathbf{IL}(\mathrm{All})$ is unknown. It is one of the major open problems in the field of interpretability logics, to characterize $\mathbf{IL}(\mathrm{All})$ in a modal way.

Definition 3.2.5. Let Γ be a set of formulas. We say that φ is provable from Γ in $\mathbf{IL}\mathsf{X}$ and write $\Gamma \vdash_{\mathbf{IL}\mathsf{X}} \varphi$, iff there is a finite sequence of formulae ending on φ, each being a theorem of $\mathbf{IL}\mathsf{X}$, a formula from Γ, or the result of applying Modus Ponens to formulas earlier in the sequence.

Clearly we have $\varnothing \vdash_{\mathbf{IL}\mathsf{X}} \varphi \Leftrightarrow \mathbf{IL}\mathsf{X} \vdash \varphi$. In the sequel we will often write just $\Gamma \vdash \varphi$ instead of $\Gamma \vdash_{\mathbf{IL}\mathsf{X}} \varphi$ if the context allows us to do so. It is well known that we have a deduction theorem for this notion of derivability.

Lemma 3.2.6 (Deduction Theorem). $\Gamma, A \vdash_{\mathbf{IL}\mathsf{X}} B \Leftrightarrow \Gamma \vdash_{\mathbf{IL}\mathsf{X}} A \to B$

Proof. "$\Leftarrow$" is obvious and "$\Rightarrow$" goes by induction on the length n of the $\mathbf{IL}\mathsf{X}$-proof σ of B from Γ, A.

If $n{>}1$, then $\sigma = \tau, B$, where B is obtained from some C and $C \to B$ occurring earlier in τ. Thus we can find subsequences τ' and τ'' of τ such that τ', C and $\tau'', C \to B$ are $\mathbf{IL}\mathsf{X}$-proofs from Γ, A. By the induction hypothesis we find $\mathbf{IL}\mathsf{X}$-proofs from Γ of the form $\sigma', A \to C$ and $\sigma'', A \to (C \to B)$. We now use the tautology $(A \to (C \to B)) \to ((A \to C) \to (A \to B))$ to get an $\mathbf{IL}\mathsf{X}$-proof of $A \to B$ from Γ. ⊣

Definition 3.2.7. A set Γ is $\mathbf{IL}\mathsf{X}$-consistent iff $\Gamma \nvdash_{\mathbf{IL}\mathsf{X}} \bot$. An $\mathbf{IL}\mathsf{X}$-consistent set is maximal $\mathbf{IL}\mathsf{X}$-consistent if for any φ, either $\varphi \in \Gamma$ or $\neg\varphi \in \Gamma$.

Lemma 3.2.8. *Every $\mathbf{IL}\mathsf{X}$-consistent set can be extended to a maximal $\mathbf{IL}\mathsf{X}$-consistent one.*

Proof. This is Lindebaum's lemma for $\mathbf{IL}\mathsf{X}$. We can just do the regular argument as we have the deduction theorem. Note that there are countably many different formulas. ⊣

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

Definition 3.2.9. Let Γ and Δ denote maximal **IL**X-consistent sets.

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

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

Lemma 3.2.10. *Let Γ and Δ denote maximal* **IL**X*-consistent sets. We have $\Gamma \prec \Delta$ iff $\Gamma \prec_\bot \Delta$.*

Proof. Above we have seen that $\Gamma \prec_A \Delta \Rightarrow \Gamma \prec \Delta$. For the other direction suppose now that $\Gamma \prec \Delta$. If $A \rhd \bot \in \Gamma$ then, by Lemma 3.1.2.1, $\Box\neg A \in \Gamma$ whence $\neg A, \Box\neg A \in \Delta$. $\dashv$

3.3 Semantics

Interpretability logics come with a Kripke-like semantics. As the signature of our language is countable, we shall only consider countable models.

Definition 3.3.1. An **IL**-frame is a triple $\langle W, R, S\rangle$. Here W is a non-empty countable universe, R is a binary relation on W and S is a set of binary relations on W, indexed by elements of W. The R and S satisfy the following requirements.

1. R is conversely well-founded[3]
2. xRy & $yRz \to xRz$
3. $yS_xz \to xRy$ & xRz
4. $xRy \to yS_xy$
5. $xRyRz \to yS_xz$
6. $uS_xvS_xw \to uS_xw$

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

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

Definition 3.3.2. An **IL**-model is a quadruple $\langle W, R, S, \Vdash\rangle$. Here $\langle W, R, S\rangle$ is an **IL**-frame and $\Vdash$ is a subset of $W \times \mathsf{Prop}$. We write $w \Vdash p$ for $\langle w, p\rangle \in \Vdash$. As usual, $\Vdash$ is extended to a subset $\widetilde{\Vdash}$ of $W \times \mathsf{Form}_{\mathbf{IL}}$ by demanding the following.

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

Note that $\widetilde{\Vdash}$ is completely determined by $\Vdash$. Thus we will denote $\widetilde{\Vdash}$ also by $\Vdash$. We call $\Vdash$ a forcing relation. The $\Vdash$-relation depends on the model M. If necessary, we will write $M, w \Vdash \varphi$, if not, we will just write $w \Vdash \varphi$. In this case we say that φ holds at w, or that φ is forced at w. We say that *p is in the range of* $\Vdash$ if $w \Vdash p$ for some w.

If $F = \langle W, R, S\rangle$ is an **IL**-frame, we will write $x \in F$ to denote $x \in W$ and similarly for **IL**-models. Attributes on F will be inherited by its constituent parts. For example $F_i = \langle W_i, R_i, S_i\rangle$. Often however we will write $F_i \models xRy$ instead of $F_i \models xR_iy$ and likewise for the S-relation. This notation is consistent with notation in first order logic where the symbol R is interpreted in the structure F_i as R_i.

If $M = \langle W, R, S, \Vdash\rangle$, we say that M is based on the frame $\langle W, R, S\rangle$ and we call $\langle W, R, S\rangle$ its underlying frame.

If Γ is a set of formulas, we will write $M, x \Vdash \Gamma$ as short for $\forall\, \gamma{\in}\Gamma\ M, x \Vdash \gamma$. We have similar reading conventions for frames and for validity.

Definition 3.3.3 (Generated submodel). Let $M = \langle W, R, S, \Vdash\rangle$ be an **IL**-model and let $m \in M$. We define $m{\upharpoonright}{*}$ to be the set $\{x \in W \mid x{=}m \vee mRx\}$. By $M{\upharpoonright}m$ we denote the submodel generated by m defined as follows.

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

Lemma 3.3.4 (Generated Submodel Lemma). *Let M be an* **IL***-model and let $m \in M$. For all formulas φ and all $x \in m{\upharpoonright}{*}$ we have that*

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

Proof. By an easy induction on the complexity of φ. ⊣

We say that an **IL**-model makes a formula φ true, and write $M \models \varphi$, if φ is forced in all the nodes of M. In a formula we write

$$M \models \varphi :\Leftrightarrow \forall\, w{\in}M \ w \Vdash \varphi.$$

If $F = \langle W, R, S\rangle$ is an **IL**-frame and $\Vdash$ a subset of $W \times \mathsf{Prop}$, we denote by $\langle W, \Vdash\rangle$ the **IL**-model that is based on F and has forcing relation $\Vdash$. We say that a frame F makes a formula φ true, and write $F \models \varphi$, if any model based on F makes φ true. In a second-order formula:

$$F \models \varphi :\Leftrightarrow \forall \Vdash \ \langle F, \Vdash\rangle \models \varphi$$

We say that an **IL**-model or frame makes a scheme true if it makes all its instantiations true. If we want to express this by a formula we should have a means to quantify over all instantiations. For example, we could regard an instantiation of a scheme X as a substitution σ carried out on X resulting in X^σ. We do not wish to be very precise here, as it is clear what is meant. Our definitions thus read

$$F \models \mathsf{X} \text{ iff } \forall \sigma \ F \models \mathsf{X}^\sigma$$

for frames F, and

$$M \models \mathsf{X} \text{ iff } \forall \sigma \ M \models \mathsf{X}^\sigma$$

for models M. Sometimes we will also write $F \models \mathbf{IL}\mathsf{X}$ for $F \models \mathsf{X}$.

It turns out that checking the validity of a scheme on a frame is fairly easy. If X is some scheme[4], let τ be some base substitution that sends different placeholders to different propositional variables.

Lemma 3.3.5. *Let* X *be a scheme, and* τ *be a corresponding base substitution as described above. Let* F *be an* **IL***-frame. We have*

$$F \models \mathsf{X}^\tau \Leftrightarrow \forall \sigma \ F \models \mathsf{X}^\sigma.$$

Proof. If $\forall \sigma \ F \models \mathsf{X}^\sigma$, then certainly $F \models \mathsf{X}^\tau$, thus we should concentrate on the other direction. Thus, assuming $F \models \mathsf{X}^\tau$ we fix some σ and $\Vdash$ and set out to prove $\langle F, \Vdash\rangle \models \mathsf{X}^\sigma$. We define another forcing relation $\Vdash'$ on F by saying that for any place holder A in X we have

$$w \Vdash' \tau(A) :\Leftrightarrow \langle F, \Vdash\rangle \models \sigma(A)$$

By induction on the complexity of a subscheme[5] Y of X we can now prove

$$\langle F, \Vdash'\rangle, w \Vdash' \mathsf{Y}^\tau \Leftrightarrow \langle F, \Vdash\rangle, w \Vdash \mathsf{Y}^\sigma.$$

By our assumption we get that $\langle F, \Vdash\rangle, w \Vdash \mathsf{X}^\sigma$. ⊣

[4] Or a set of schemata. All of our reasoning generalizes without problems to sets of schemata. We will therefore no longer mention the distinction.

[5] It is clear what this notion should be.

If χ is some formula in first, or higher, order predicate logic, we will evaluate $F \models \chi$ in the standard way. In this case F is considered as a structure of first or higher order predicate logic. We will not be too formal about these matters as the context will always dict us which reading to choose.

Definition 3.3.6. Let X be a scheme of interpretability logic. We say that a formula $\mathcal{C}$ in first or higher order predicate logic is a frame condition of X if

$$F \models \mathcal{C} \quad \text{iff } F \models \mathsf{X}.$$

The $\mathcal{C}$ in Definition 3.3.6 is also called the frame condition of the logic **IL**X. A frame satisfying the **IL**X frame condition is often called an **IL**X-frame. In case no such frame condition exists, an **IL**X-frame resp. model is just a frame resp. model, validating X.

The semantics for interpretability logics is good in the sense that we have the necessary soundness results.

Lemma 3.3.7 (Soundness). $\mathbf{IL} \vdash \varphi \Rightarrow \forall F \ F \models \varphi$

Proof. By induction on the length of an **IL**-proof of φ. The requirements on R and S in Definition 3.3.1 are precisely such that the axiom schemata hold. Note that all axiom schemata have their semantical counterpart except for the schema $(A \rhd C) \wedge (B \rhd C) \rightarrow A \vee B \rhd C$. ⊣

Lemma 3.3.8 (Soundness). *Let $\mathcal{C}$ be the frame condition of the logic* **IL**X*. We have that*

$$\mathbf{IL}\mathsf{X} \vdash \varphi \Rightarrow \forall F \ (F \models \mathcal{C} \Rightarrow F \models \varphi).$$

Proof. As that of Lemma 3.3.7, plugging in the definition of the frame condition at the right places. Note that we only need the direction $F \models \mathcal{C} \Rightarrow F \models X$ in the proof. ⊣

Corollary 3.3.9. *Let M be a model satisfying the* **IL**X *frame condition, and let $m \in M$. We have that $\Gamma := \{\varphi \mid M, m \Vdash \varphi\}$ is a maximal* **IL**X*-consistent set.*

Proof. Clearly $\bot \notin \Gamma$. Also $A \in \Gamma$ or $\neg A \in \Gamma$. By the soundness lemma, Lemma 3.3.8, we see that Γ is closed under **IL**X consequences. ⊣

Lemma 3.3.10. *Let M be a model such that $\forall w{\in}M \ \ w \Vdash$* **IL**$\mathsf{X}$ *then* **IL**$\mathsf{X} \vdash \varphi \Rightarrow M \models \varphi$*.*

Proof. By induction on the derivation of φ. ⊣

A modal logic **IL**X with frame condition $\mathcal{C}$ is called complete if we have the implication the other way round too. That is,

$$\forall F \ (F \models \mathcal{C} \Rightarrow F \models \varphi) \Rightarrow \mathbf{IL}\mathsf{X} \vdash \varphi.$$

A major concern of Part II of this thesis is the question whether a given modal logic **IL**X is complete.

Definition 3.3.11. $\Gamma \Vdash_{\mathbf{IL}\mathsf{X}} \varphi$ iff $\forall M \ M \models \mathbf{IL}\mathsf{X} \Rightarrow (\forall m{\in}M \ [M,m \Vdash \Gamma \Rightarrow M,m \Vdash \varphi])$

Lemma 3.3.12. *Let Γ be a finite set of formulas and let* $\mathbf{IL}\mathsf{X}$ *be a complete logic. We have that* $\Gamma \vdash_{\mathbf{IL}\mathsf{X}} \varphi$ *iff* $\Gamma \Vdash_{\mathbf{IL}\mathsf{X}} \varphi$.

Proof. Trivial. By the deduction theorem $\Gamma \vdash_{\mathbf{IL}\mathsf{X}} \varphi \Leftrightarrow \vdash_{\mathbf{IL}\mathsf{X}} \bigwedge \Gamma \to \varphi$. By our assumption on completeness we get the result. Note that the requirement that Γ be finite is necessary, as our modal logics are in general not compact (see also Section 5.1.1). $\dashv$

Often we shall need to compare different frames or models. If $F = \langle W, R, S \rangle$ and $F' = \langle W', R', S' \rangle$ are frames, we say that F is a subframe of F' and write $F \subseteq F'$, if $W \subseteq W'$, $R \subseteq R'$ and $S \subseteq S'$. Here $S \subseteq S'$ is short for $\forall w{\in}W \ (S_w \subseteq S'_w)$.

Chapter 4

Remarks on **IL**(All)

In this chapter we study modal formulas that are interpretability principles for any numberized theory T. In other words, we shall study **IL**(All). The problem to give a modal characterization of **IL**(All) still remains open. The best candidate for now would be $\mathbf{IL}\mathsf{R}^*$. In this chapter, we shall first make some basic observations on **IL**(All) and discuss the role of reflexivity. In Section 4.2 we will present two modal systems that generate principles in **IL**(All).

4.1 Basic observations

If $\varphi \in \mathbf{IL}(\mathrm{All})$, it should certainly be an interpretability principle of any essentially reflexive theory and of any finitely axiomatizable theory containing supexp. Thus, by Theorem 3.2.2 and 3.2.3 we see that $\mathbf{IL}(\mathrm{All}) \subseteq \mathbf{IL}\mathsf{M} \cap \mathbf{IL}\mathsf{P}$. And actually, we know that this is a strict inclusion. In [Vis97] it is shown that $A \rhd \Diamond B \rightarrow \Box(A \rhd \Diamond B) \in (\mathbf{IL}\mathsf{M} \cap \mathbf{IL}\mathsf{P}) \setminus \mathbf{IL}(\mathrm{All})$.

4.1.1 Modal considerations

We shall see in Section 4.2 that all of the principles $\mathsf{M_0}$, W, $\mathsf{P_0}$ and R are indeed interpretability principles for any numberized theory. As $\mathbf{IL}(\mathrm{All}) \subset \mathbf{IL}\mathsf{M} \cap \mathbf{IL}\mathsf{P}$, we know that any possible new principle should be found in this intersection. To search this intersection, the modal semantics has proved to be an excellent guideline.

There is a close connection between principles, c.q. modal formulas, and properties of frames, viz. frame conditions. For example, it is easy to calculate the frame conditions of $\mathsf{M_0}$, W, $\mathsf{P_0}$ and R and see that, indeed, they follow from both the frame condition of $\mathbf{IL}\mathsf{M}$ and the frame condition of $\mathbf{IL}\mathsf{P}$.

Thus, looking for good candidates for $\varphi \in \mathbf{IL}(\mathrm{All})$ is often done by looking for good frame conditions that hold on both $\mathbf{IL}\mathsf{M}$ and $\mathbf{IL}\mathsf{P}$ frames. But then again, a single frame condition can yield various different principles.

The principle $\mathsf{P_0}$ was found by Visser by strengthening the frame condition of $\mathsf{M_0}$ so that it still is in $\mathbf{IL}\mathsf{M} \cap \mathbf{IL}\mathsf{P}$. (See [Joo98].) In an attempt to prove the modal completeness of $\mathbf{IL}\mathsf{P_0W^*}$, the new principle R was discovered in [GJ04]. And as it turns out, $\mathsf{P_0}$ and R have the same frame condition.

By easy semantical arguments we saw that all the frame conditions for $\mathsf{M_0}$, $\mathsf{P_0}$ and R hold on all $\mathbf{IL}\mathsf{P}$-frames. These arguments all employed models of "height" three. It is also possible to find principles in $\mathbf{IL}\mathsf{M} \cap \mathbf{IL}\mathsf{P}$ that use for their justification only models up to height two. Two such examples are

$$(A \rhd \Diamond B) \wedge (B \rhd C) \rightarrow (\Diamond A \rightarrow \Diamond(A \wedge \Diamond B) \vee \Diamond(B \wedge \Diamond C)) \quad \text{and}$$
$$(\Diamond A \rhd \Diamond B) \wedge (A \rhd B) \rightarrow \Box(\Diamond A \rightarrow \Diamond B).$$

However, all these statements seem to be refutable in some numberized theories.

In [Vis88] **IL**(All) was conjectured[1] to be $\mathbf{IL}\mathsf{W}$. In [Vis91] this conjecture was falsified and strengthened to a new conjecture. It was conjectured that $\mathbf{IL}\mathsf{W}^*$, which is a proper extension of $\mathbf{IL}\mathsf{W}$ is **IL**(All).

In [Joo98] this conjecture was falsified. It was proved that the logic $\mathbf{IL}\mathsf{W^*P_0}$ is a proper extension of $\mathbf{IL}\mathsf{W}^*$, and that $\mathbf{IL}\mathsf{W^*P_0}$ is a subsystem of **IL**(All). In [GJ04] it is shown that $\mathbf{IL}\mathsf{RW}$ is a proper extension of $\mathbf{IL}\mathsf{W^*P_0}$ and that $\mathbf{IL}\mathsf{RW} \subseteq \mathbf{IL}(\text{All})$. In Lemma 8.1.4 it is shown that $\mathbf{IL}\mathsf{RW} = \mathbf{IL}\mathsf{R}^*$. The current best guess for **IL**(All) would thus be $\mathbf{IL}\mathsf{R}^*$.

If for some theory T we have that $\mathbf{IL}(\text{T}) = \mathbf{IL}\mathsf{R}^*$, then the question would be settled and $\mathbf{IL}(\text{All}) = \mathbf{IL}\mathsf{R}^*$. It is also possible that **IL**(All) is never attained. That is, for no T we have that $\mathbf{IL}(\text{T}) = \mathbf{IL}(\text{All})$.

If some theory T does attain **IL**(All), then this theory can certainly not be finitely axiomatized, as we know that $\mathsf{P} \notin \mathbf{IL}(\text{All})$. As we shall see, we also cannot have too much reflection in T, as this would yield new principles that essentially depend on this reflection.

4.1.2 Reflexive theories

We shall now see that if a theory T has too much reflection, then it can never be such that $\mathbf{IL}(\text{T}) = \mathbf{IL}(\text{All})$.

Lemma 4.1.1. *Let U be a reflexive theory containing* exp. *We have* $\mathbf{IL}(\text{U}) \vdash \top \rhd B \rightarrow \top \rhd B \wedge \Box\bot$.

Proof. By Lemma 2.1.2 we get that $U \vdash \top \rhd \beta \rightarrow \forall x\ \Box_U \mathsf{Con}_{U,x}(\beta)$. By Lemma 1.2.3 and Remark 1.2.4 we find n large enough so that $U \vdash \Box_U(\Box\bot \rightarrow \Box_{U,n}\Box\bot)$. We now reason in U and assume $\top \rhd \beta$. Thus, we get $\forall x\ \Box_U \mathsf{Con}_{U,x}(\beta)$ and $\forall x\ \Box_U(\Box\bot \rightarrow \mathsf{Con}_{U,x}(\beta \wedge \Box\bot))$. By Lemma 2.1.1 we obtain $\Box\bot \rhd \beta \wedge \Box\bot$. But, by Lemma 3.1.2, $\top \rhd \Box\bot$ and we obtain $\top \rhd \beta \wedge \Box\bot$. ⊣

By an easy semantical argument we see that $\mathbf{IL} \nvdash \top \rhd B \rightarrow \top \rhd B \wedge \Box\bot$. However, the principle is provable in $\mathbf{IL}\mathsf{W}$.

[1] For the modal language restricted to the *unary* connective $\top \rhd A$ in combination with $\Box$, the problem of the interpretability logic of all numberized theories has been solved by Maarten de Rijke, see his [dR92].

Lemma 4.1.2. *Let U be a theory, containing* exp*, such that any* Π_1*-extension is reflexive. Then,*

$$\mathbf{IL}(\mathrm{U}) \vdash \bigwedge_i \Diamond A_i \rhd B \rightarrow \bigwedge_i \Diamond A_i \wedge \Box C \rhd B \wedge \Box C.$$

Proof. As the proof of Lemma 4.1.1 noting that $\Diamond$-formulas always translate to a Π_1-formula. ⊣

It is clear that $\mathbf{IL}\mathsf{P} \nvdash \Diamond A \rhd B \rightarrow \Diamond A \wedge \Box C \rhd B \wedge \Box C$. Thus, theories U, satisfying the conditions of Lemma 4.1.2, (like PRA) can never be a candidate for $\mathbf{IL}(\mathrm{U}) = \mathbf{IL}(\mathrm{All})$. It is not excluded that for some reflexive[2] U that does not satisfy the conditions of Lemma 4.1.2, we have $\mathbf{IL}(\mathrm{U}) = \mathbf{IL}(\mathrm{All})$. An example of such a theory is $\mathrm{EA} + \mathsf{Con}(\mathrm{EA}) + \mathsf{Con}(\mathrm{EA} + \mathsf{Con}(\mathrm{EA})) + \cdots$.

4.1.3 Essentially reflexive theories

We know that $\mathbf{IL}\mathsf{M}$ is the interpretability logic of any essentially reflexive theory. But what sort of theories are these essentially reflexive theories? In this subsection we see that this depends on the notion of essential reflexivity. But, for most natural theories T that are essentially reflexive, we shall see that T has full induction.

In the definition of essentially reflexive, Definition 3.2.1, we stressed that we only considered sentential extensions of T. This is called *local essential reflexivity.* We can also consider extensions with formulas. This gives rise to the notion of *global essential reflexivity.* We can restate the definition as follows.

$$\forall \varphi \, \forall n \; T \vdash \varphi(x) \rightarrow \mathsf{Con}_n(T + \varphi(\dot{x}))$$

In this subsection we shall compare the two notions of essential reflexivity.

Lemma 4.1.3. *If T is an essentially globally reflexive theory extending*[3] EA*, then T satisfies full induction.*

Proof. We will show that T satisfies the full induction rule, from which the result follows. So, suppose that

$$T \vdash \varphi(0) \wedge \forall x \; (\varphi(x) \rightarrow \varphi(x+1)).$$

Then, for some m,

$$T \vdash \Box_{T,m}(\varphi(0) \wedge \forall x \; (\varphi(x) \rightarrow \varphi(x+1))).$$

Thus, also

$$T \vdash \forall x \; \Box_{T,m}(\varphi(\overline{x}) \rightarrow \varphi(\overline{x+1}))$$

[2] It does not seem to impose any obstacle either if any Σ_1-extension is reflexive.

[3] It is not hard to extend the argument to S^1_2, by using cuts, efficient numerals and different induction principles.

can be obtained uniformly in x.

All these proofs can be glued together to obtain $T \vdash \forall x\ \Box_{T,m}\varphi(x)$, whence by essential reflexivity we get $T \vdash \forall x\ \varphi(x)$. $\dashv$

Note that the same argument only yields that T is closed under the Π_1-induction rule if T is just reflexive. But this is a really weak closure condition.

The use of global reflexivity was really needed. It is known that Lemma 4.1.3 does not hold for essentially locally reflexive theories. Here follows a short argument that is attributed to Feferman.

If T is any theory in the language of arithmetic, then $U := T \cup \{\varphi \to \mathsf{Con}(\varphi) \mid \varphi \text{ a sentence } \}$ has two nice properties, as is readily verified. First, U is essentially locally reflexive, and secondly, $T + \mathsf{True}_{\Pi_1} \supseteq U$. Here True_{Π_1} denotes the set of all true (in the standard model) Π_1-sentences.

To see that U is essentially locally reflexive, we see that for any sentence ψ, and for any number n we have $U \vdash \psi \to \mathsf{Con}_n(U + \psi)$. For this, it is sufficient to show that $U \vdash \psi \to \mathsf{Con}(\psi \wedge U[n])$ where $U[n]$ denotes the conjunction of the first n axioms. By definition $\psi \wedge U[n] \to \mathsf{Con}(\psi \wedge U[n])$ is an axiom of U, whence $U \vdash \psi \to \mathsf{Con}(\psi \wedge U[n])$.

To see that U is included in $T + \mathsf{True}_{\Pi_1}$, we need to see that any axiom of the form $\varphi \to \mathsf{Con}(\varphi)$ is. But, either $\mathsf{Con}(\varphi) \in \mathsf{True}_{\Pi_1}$ and $T + \mathsf{True}_{\Pi_1} \vdash \varphi \to \mathsf{Con}(\varphi)$, or $\mathsf{Con}(\varphi)$ is not true. In that case we have $\vdash \neg\varphi$, and consequently $\vdash \varphi \to \mathsf{Con}(\varphi)$.

Thus, for example, $\mathrm{EA} + \{\varphi \to \mathsf{Con}(\varphi) \mid \varphi \text{ a sentence } \} \subseteq \mathrm{EA} + \mathsf{True}_{\Pi_1}$. It is well known that no Σ_3-axiomatized theory can prove $\mathrm{I}\Sigma_1$ (see for example Fact 2.3 from [Joo03a]). But $\mathrm{EA} + \mathsf{True}_{\Pi_1}$ has a Π_2-axiomatization, thus $\mathrm{EA} + \{\varphi \to \mathsf{Con}(\varphi) \mid \varphi \text{ a sentence } \} \nvdash \mathrm{I}\Sigma_1$.

Admittedly, theories like the U above are a bit artificial. All natural theories that are essentially reflexive are globally so and hence by Lemma 4.1.3 satisfy full induction.

4.2 Arithmetical soundness proofs

In this section we shall give arithmetical soundness proofs for interpretability principles that hold in all reasonable arithmetical theories. These principles should thus certainly hold in any finitely axiomatizable and in any essentially reflexive theory. This means that the principles should be provable both in **IL**P and **IL**M.

We shall see that the two modal proofs give rise to two different arithmetical soundness proofs. The M-style proofs use definable cuts and find place in some sort of modal system as described in Subsection 4.2.1. The P-style proofs are based on quasi-finite approximations of theories. This behavior is captured also in a modal-like system as we shall see in Subsection 4.2.4.

The modal systems that we present are not completely formal. In [JV04b] a more formal treatment of the systems is given.

4.2.1 Cuts and interpretability logics

All our knowledge about cuts and interpretations can be collected in modal principles. These principles will contain variables I and J denoting (possibly non-standard) cuts. In almost all principles, the cut variables can be universally quantified. That is to say, the arithmetical translations of these principles are provable for any possible choice of the cuts I and J. The sole exception is the principle M^J.

We will only list principles that contain cut variables. In our reasoning we shall freely use all regular principles from **IL**. The rules that we use are as always just Modus Ponens and Necessitation.

$(\rightarrow)^\mathsf{J}$	$\Box(\Box^J A \rightarrow \Box A)$	$(\forall J)$
L_1^J	$\Box(\Box^J(A \rightarrow B) \rightarrow (\Box^J A \rightarrow \Box^J B))$	$(\forall J)$
$\mathsf{L}_{2a}^\mathsf{J}$	$\Box A \rightarrow \Box\Box^J A$	$(\forall J)$
$\mathsf{L}_{2b}^\mathsf{J}$	$\Box(\Box^I A \rightarrow \Box^I \Box^J A)$	$(\forall I \forall J)$
$\mathsf{L}_{3a}^\mathsf{J}$	$\Box(\Box^J A \rightarrow A) \rightarrow \Box A$	$(\forall J)$
$\mathsf{L}_{3b}^\mathsf{J}$	$\Box(\Box^J(\Box^I A \rightarrow A) \rightarrow \Box^J A)$	$(\forall I \forall J)$
J_5^J	$\Diamond^J A \rhd A$	$(\forall J)$
M^J	$A \rhd B \rightarrow A \wedge \Box^J C \rhd B \wedge \Box C$	$(\exists J)$

It might be desirable to add some simple operations on cuts to the modal language like $I \subseteq J$, $I \cap J$ and $I \cup J$. Like this, we get for example the following principle.

$$\Box^J(A \rightarrow B) \rightarrow (\Box^I A \rightarrow \Box^{I \cup J} B) \quad (\forall I \forall J)$$

It is not hard to see that all the principles mentioned above indeed hold in all numberized theories. The principle $(\rightarrow)^\mathsf{J}$ is a triviality; L_1^J reflects that concatenation of proofs in a cut remains within this cut as concatenation is approximately multiplication; $\mathsf{L}_{2a}^\mathsf{J}$ is a special case of Lemma 4.2.1 and $\mathsf{L}_{2b}^\mathsf{J}$ expresses the formalization of this lemma; $\mathsf{L}_{3a}^\mathsf{J}$ is Löb's theorem with cuts, as proved in Lemma 4.2.2 and $\mathsf{L}_{3b}^\mathsf{J}$ is the formalization of this lemma; J_5^J follows from Theorem 1.3.7; M^J is Lemma 4.2.3.

Lemma 4.2.1. *For any U-cuts I and J we have that $T \vdash \Box_U^I \alpha \rightarrow \Box_U^I \Box_U^J \alpha$.*

Proof. Reason in T and assume that $\mathsf{Proof}_U(p, \alpha)$ for some $p \in I$. As $\mathsf{Proof}_U(p, \alpha)$ $\in \exists\Sigma_1^b$, by Lemma 1.2.3 we get for some p' that $\mathsf{Proof}_U(p', \mathsf{Proof}_U(p, \alpha))$. As I is closed under ω_1, we see that actually $p' \in I$, whence $\Box_U^I \mathsf{Proof}_U(p, \alpha)$. Lemma 1.3.2 now gives us the desired $\Box_U^I \Box_U^J \alpha$. ⊣

Lemma 4.2.2. *For any U-cuts I and J we have that $T \vdash \Box_U^J(\Box_U^I \alpha \rightarrow \alpha) \rightarrow \Box_U^J \alpha$.*

Proof. The lemma really just boils down to copying the standard proof of Löb's theorem making some some minor adaptations. In the proof we shall omit the subscript U to the boxes.

Thus, let F be a fixed point of the equation $F \leftrightarrow (\Box^I F \to \alpha)$. By applying twice $\mathsf{Nec}^{\mathsf{J}}$ on the interesting side of the bi-implication, we arrive at $\Box^J\Box^I(F \to (\Box^I F \to \alpha))$. We now reason within T using our assumption $\mathbb{A}:\ \Box^J(\Box^I\alpha \to \alpha)$ as follows.

$$\begin{array}{lcll}
\Box^J\Box^I(F \to (\Box^I F \to \alpha)) & \to & \Box^J(\Box^I F \to (\Box^I\Box^I F \to \Box^I\alpha)) & \text{by } \mathsf{L}_2^{\mathsf{J}} \\
 & \to & \Box^J(\Box^I F \to \Box^I\alpha) & \text{by } \mathbb{A} \\
 & \to & \Box^J(\Box^I F \to \alpha) & (*) \\
 & \to & \Box^J F & \text{by } \mathsf{L}_2^{\mathsf{J}} \\
 & \to & \Box^J\Box^I F & \text{by } (*) \\
 & \to & \Box^J\alpha &
\end{array}$$

$\dashv$

Lemma 4.2.3. *For any α, β and γ we have that $T \vdash \alpha \rhd \beta \to \exists J\ (\alpha \wedge \Box^J\gamma \rhd \beta \wedge \Box\gamma)$.*

Proof. This is a direct consequence of Pudlák's lemma. So, we suppose $j : \alpha \rhd \beta$ and consider the corresponding $(T+\alpha)$-cut J and the j, J-function h that are given by Lemma 1.3.11. Now, as $\mathsf{Proof}_T(p,\gamma) \in \Delta_0$, we get that $\Box_{T+\alpha}\forall p{\in}J\ (\mathsf{Proof}_T(p,\gamma) \leftrightarrow (\mathsf{Proof}_T(h(p),\gamma))^j)$ and thus certainly

$$\Box_{T+\alpha}(\Box^J\gamma \to (\Box\gamma)^j). \tag{4.1}$$

It is now easy to see that $j : T+\alpha \rhd T+\beta$. For, if $\Box_{T+\beta+\Box\gamma}\varphi$, we get $\Box_{T+\beta}\Box\gamma \to \varphi$, whence by our assumption $\Box_{T+\alpha}(\Box\gamma \to \varphi)^j$, i.e., $\Box_{T+\alpha}((\Box\gamma)^j \to \varphi^j)$. By (4.1) we now get the required $\Box_{T+\alpha+\Box^J\gamma}\varphi^j$. $\dashv$

With the modal principles we have given here, many interesting facts can be derived. With $A \equiv B$ we shall denote that A and B are equi-interpretable. That is, $(A \rhd B)$ & $(B \rhd A)$.

Lemma 4.2.4. *For any I and J, we have $A \equiv A \wedge \Box^I\neg A \equiv A \vee \Diamond^J A$.*

Proof. Just copy the proofs from **IL**, replacing some regular principles with the new principles relativized to a cut. In the derivation of $A \rhd A \wedge \Box^I\neg A$ we use $\mathsf{L}_{3b}^{\mathsf{J}}$ to obtain $\Box(\Diamond^I A \to \Diamond^I(A \wedge \Box^I\neg A))$. $\dashv$

Lemma 4.2.5. *For any J we have $\neg(A \rhd \neg C) \to \Diamond(A \wedge \Box^J C)$.*

Proof. By contraposition we get that (sloppy notation) $\Box(A \to \Diamond^J\neg C) \to A \rhd \Diamond^J\neg C \rhd \neg C$. $\dashv$

4.2.2 Approximations of Theories

It is a triviality that for finitely axiomatized theories we have $\alpha \rhd \beta \to \Box(\alpha \rhd \beta)$. For, $\alpha \rhd \beta$ is nothing but a Σ_1-sentence and we get $\Box(\alpha \rhd \beta)$.

Thus, if we want to mimic the P behavior for a general theory T, we should make the $\alpha \rhd_T \beta$ a simple enough statement so that we get $\Box(\alpha \rhd \beta)$. Clearly, for $\alpha \rhd_T \beta$ in general this is not possible, but in some situations we are also satisfied with $\Box(\alpha \rhd_{T'} \beta)$ where T' is some approximation of T.

There are two choices of T' that can be made. First, we could take for T' a finite subtheory of T, and note that[4] $T + \alpha \rhd T' + \beta$. Second, we could define a theory T' that is extensionally the same as T, but for which $T + \alpha \rhd T' + \beta$ is so simple that we actually get $\Box(T + \alpha \rhd T' + \beta)$. We shall work out the second variant, albeit some of our arguments can also be carried out using the first approach.

A first idea would be to take for the axioms of T' just the axioms of T that are in translated form provable in $T + \alpha$. This almost works, but we want to be sure that T' contains verifiably enough arithmetic to do for example a Henkin construction.

Thus, the second idea would be to just add S^1_2 to our first approach. This turns out to only work in the presence of Σ_1-collection and exp. The exp is then needed to get provable Σ_1-completeness whence L_2 for $\Box_{T'}$.

We shall use a use a variation of Craig's trick so that our theory T' will stay $\exists\Sigma^b_1$-definable. The same trick makes the use of $\mathrm{B}\Sigma_1$ superfluous.

Let s^1_2 be the sentence axiomatizing S^1_2.

Definition 4.2.6. If $k : T + \alpha \rhd T + \beta$, we define T^k as follows.

$$\begin{array}{ll}\mathsf{Axiom}_{T^k}(x) := & (x = s^1_2) \ \vee \\ & \exists p\ (x = \ulcorner \varphi \wedge (\underline{p} = \underline{p})\urcorner \wedge \mathsf{Axiom}_T(\varphi) \wedge \mathsf{Proof}_{T+\alpha}(p, \varphi^k))\end{array}$$

It is clear that $\mathsf{Axiom}_{T^k}(x)$ is in poly-time decidable if $\mathsf{Axiom}_T(x)$ is so. Note that we work with efficient numerals $\underline{p}$.

Lemma 4.2.7. *(In S^1_2) If $k : \alpha \rhd_T \beta$, then $\Box_T\varphi \leftrightarrow \Box_{T^k}\varphi$ and consequently $T^k + \alpha \equiv T + \alpha \rhd T + \beta \equiv T^k + \beta$.*

Proof. $\Box_{T^k}\varphi \to \Box_T\varphi$ is clear, as we can replace every axiom $\varphi \wedge (\underline{p} = \underline{p})$ of T^k by a proof of $\varphi \wedge (\underline{p} = \underline{p})$ from the single T-axiom φ. Note that we have these proofs available as we used efficient numerals.

On the other hand, if $\Box_T\varphi$, we have a proof p of φ from the axioms, say $\tau_0, \ldots, \tau_n$. Now, by the assumption that $k : \alpha \rhd_T \beta$ (smoothness gives the appropriate bounds) we obtain $(T + \alpha)$-proofs of p_i of ${\tau_i}^k$. We can now replace every axiom occurrence of τ_i in p by

$$\frac{\tau_i \wedge (\underline{p_i} = \underline{p_i})}{\tau_i} \wedge E, l$$

and obtain a T^k-proof of φ. ⊣

Note that, although we do have $\Box_{\mathsf{S}^1_2}(\Box_{T^k}\varphi \to \Box_T\varphi)$ we shall in general not have $\Box_{\mathsf{S}^1_2}(\Box_T\varphi \to \Box_{T^k}\varphi)$.

[4] It would have been even nicer to get $T' + \alpha \rhd T' + \beta$, but it is not clear if this can always be established.

Lemma 4.2.8. $\mathsf{S}^1_2 \vdash k : T+\alpha \rhd T+\beta \to \Box_T(T+\alpha \rhd T^k+\beta)$

Proof. As we shall need bounds on proofs of statements of the form $\underline{p}=\underline{p}$ we consider some function f that is monotone in x, such that for any x, the k-translation of the canonical proof of $\underline{x}=\underline{x}$ is bounded by $f(x,k)$. Clearly, in S^1_2 we can define such a function f and prove its totality.

Now, we reason in S^1_2 and assume $k : T+\alpha \rhd T+\beta$. Thus certainly $\Box_{T+\alpha}\beta^k$ and also

$$\Box_T \Box_{T+\alpha}\beta^k. \tag{4.2}$$

Likewise we get $\Box_{T+\alpha}(s^1_2)^k$ and also $\Box_T\Box_{T+\alpha}(s^1_2)^k$. Let b be such that $\mathsf{Proof}_{T+\alpha}(b,\beta^k)$ and let s be such that $\mathsf{Proof}_{T+\alpha}(s,(s^1_2)^k)$.

Now, we reason in T. We are going to show that $k : T+\alpha \rhd_s T^k+\beta$. So, let us consider some arbitrary x. Let now $y := \max\{s,b,x\cdot f(x,k)\}$. We shall see that for any $\tau \leq x$ with $\mathsf{Axiom}_{T^k+\beta}(\tau)$, there is a proof $p' \leq y$ with $\mathsf{Proof}_{T+\alpha}(p',\tau^k)$.

If $\mathsf{Axiom}_{T^k+\beta}(\tau)$, either $\tau=\beta$ and we are done by (4.2), or we have that $\mathsf{Axiom}_{T^k}(\tau)$. Let us consider the latter case. Again, if $\tau = s^1_2$ we are done. So, we may assume that τ is of the form $\varphi \wedge (\underline{p}=\underline{p})$, with $\mathsf{Proof}_{T+\alpha}(p,\varphi^k)$. Clearly, $p\leq\tau\leq x$. We can now easily obtain a $(T+\alpha)$-proof p' of $\varphi^k \wedge (\underline{p}=\underline{p})^k$. As p' is obtained by concatenating a proof of $(\underline{p}=\underline{p})^k$ to p, it is, by our assumptions on f, surely bounded by $x\cdot f(x,k)$. ⊣

We note that we may replace $\rhd_s$ in the antecedent of Lemma 4.2.8 by $\rhd_t$.

4.2.3 Approximations and modal logics

Just as in Subsection 4.2.1, we can make some sort of modal system in which facts about approximations and interpretability are reflected. As we shall see, the situation is a slightly more complicated than in the case of cuts and modal logics. This is due to the fact that we seem to lose necessitation.

Let us first introduce some notation. With $\alpha\rhd^k\beta$ we shall denote $T+\alpha \rhd T^k+\beta$, and with $\Box^k\alpha$ we shall denote $\Box_{T^k}\alpha$.

In Lemma 4.2.7 we have seen that $\Box_T\alpha \to \Box_{T^k}\alpha$. However, in general we do not have $\Box_T(\Box_T\alpha \to \Box_{T^k}\alpha)$. It is thus unlikely that our modal system should reflect necessitation. However, there is an easy way to handle this.

Definition 4.2.9. With $\mathbf{IL}\mathsf{X}^{\Box}$ we denote the modal logic, whose axioms are all the axioms of $\mathbf{IL}\mathsf{X}$ preceded by some number (possibly zero) of boxes. The only rule of $\mathbf{IL}\mathsf{X}^{\Box}$ is modus ponens. If Y is some set of axiom schemata, we denote by $\mathbf{IL}\mathsf{X}^{\Box}\mathsf{Y}$, the system with axioms all axioms (or equivalently, all theorems) of $\mathbf{IL}\mathsf{X}^{\Box}$ and all instantiations of schemata from Y. The sole rule of inference is modus ponens.

Lemma 4.2.10. $\mathbf{IL}\mathsf{X} = \mathbf{IL}\mathsf{X}^{\Box}$

Proof. Both $\mathbf{IL}\mathsf{X} \subseteq \mathbf{IL}\mathsf{X}^{\Box}$ and $\mathbf{IL}\mathsf{X}^{\Box} \subseteq \mathbf{IL}\mathsf{X}$ go by an easy induction on the length of proofs. We only use L_1 for one direction and necessitation for the other. ⊣

Before we give a list with principles we make one more convention. We say that $\Box^{\mathsf{id}}A$ resp. $A \rhd^{\mathsf{id}} B$ is on the syntactic level the same as $\Box A$ resp. $A \rhd B$. The quantifiers are to be understood to range over interpretations $k : \top \rhd_T \top$.

$(\mathsf{E}\Box)^{\mathsf{k}}$	$\Box^k A \leftrightarrow \Box A$	$(\forall k)$
$(\mathsf{E}\rhd)^{\mathsf{k}}$	$A \rhd^k B \leftrightarrow A \rhd B$	$(\forall k)$

$(\rightarrow \Box)^{\mathsf{k}}$	$\Box^k A \rightarrow \Box A$	$(\forall k)$
$(\rightarrow \rhd)^{\mathsf{k}}$	$A \rhd B \rightarrow A \rhd^k B$	$(\forall k)$
$\mathsf{L}_1^{\mathsf{k}}$	$\Box^k(A \rightarrow B) \rightarrow (\Box^k A \rightarrow \Box^k B)$	$(\forall k)$
$\mathsf{L}_2^{\mathsf{k}}$	$\Box^l A \rightarrow \Box^k \Box^l A$	$(\forall k \forall l)$
$\mathsf{L}_3^{\mathsf{k}}$	$\Box^k(\Box^k A \rightarrow A) \rightarrow \Box^k A$	$(\forall k)$
$\mathsf{J}_1^{\mathsf{k}}$	$\Box^k(A \rightarrow B) \rightarrow A \rhd^k B$	$(\forall k)$
$\mathsf{J}_{2\mathsf{a}}^{\mathsf{k}}$	$(A \rhd B) \wedge (B \rhd^k C) \rightarrow A \rhd^k C$	$(\forall k)$
$\mathsf{J}_{2\mathsf{b}}^{\mathsf{k}}$	$(A \rhd^k B) \wedge \Box^k(B \rightarrow C) \rightarrow A \rhd^k C$	$(\forall k)$
$\mathsf{J}_3^{\mathsf{k}}$	$(A \rhd^k C) \wedge (B \rhd^k C) \rightarrow A \vee B \rhd^k C$	$(\forall k)$
$\mathsf{J}_4^{\mathsf{k}}$	$A \rhd^k B \rightarrow (\Diamond A \rightarrow \Diamond^k B)$	$(\forall k)$
$\mathsf{J}_5^{\mathsf{k}}$	$A \rhd^l \Diamond^k B \rightarrow A \rhd^k B$	$(\forall k \forall l)$
P^{k}	$A \rhd B \rightarrow \Box(A \rhd^k B)$	$(\exists k)$

The modal reasoning we will perform using these principles will look like $\mathbf{IL}\mathsf{X}^{\Box}\mathsf{Y}$, where X is L_1-J_5 together with $(\rightarrow \Box)^{\mathsf{k}}$-$\mathsf{P}^{\mathsf{k}}$, and $\mathsf{Y} = \{(\mathsf{E}\Box)^{\mathsf{k}}, (\mathsf{E}\rhd)^{\mathsf{k}}\}$. We call the latter axioms *extensionality axioms*. Of course, we should somehow take the nature of the quantifiers along in our reasoning.

It is not hard to see that all principles are arithmetically valid. As T^k contains S_2^1, many arguments like $\mathsf{L}_1^{\mathsf{k}}$-$\mathsf{L}_3^{\mathsf{k}}$ and $\mathsf{J}_5^{\mathsf{k}}$ go[5] as always. $\mathsf{J}_1^{\mathsf{k}}$ follows easily from $(\rightarrow \Box)^{\mathsf{k}}$. But, $(\rightarrow \Box)^{\mathsf{k}}$ together with $(\mathsf{E}\Box)^{\mathsf{k}}$ is just Lemma 4.2.7, and $(\mathsf{E}\rhd)^{\mathsf{k}}$ is a direct consequence of it. Finally, P^{k} is Lemma 4.2.8.

We make no claims on the completeness of our modal system. Neither do we say anything about efficiency. For example, if φ is derivable in the system and φ only contains standard modalities, then it is desirable that also $\Box\varphi$ is derivable.

4.2.4 Arithmetical soundness results

We now come to the actual soundness proofs of the principles W, M_0, W^*, P_0, and R. As M_0 and P_0 both follow from R and as W^* follows from M_0 and W, it would be sufficient to just prove the soundness[6] of R and W. However, we have decided to give short proofs for all principles. Like this, the close match

[5] We note that $A \rhd \Diamond B \rightarrow A \rhd B$ is over J_1 and J_2 equivalent to $\Diamond A \rhd A$.

[6] In [GJ04] a principle is given that is precisely W and R together. See also Lemma 8.1.4.

between the modal systems comes better to the fore. Per principle we shall give a proof in **IL**P and in **IL**M. These proofs can then be copied almost literally to yield arithmetical soundness proofs.

The principle W

Lemma 4.2.11. **IL**$\mathsf{P} \vdash \mathsf{W}$ *and* **IL**$\mathsf{P_R} \vdash \mathsf{W}$

Proof.

$$\begin{array}{rcll} A \rhd B & \to & \Box(A \rhd B) & \\ & \to & \Box(\Diamond A \to \Diamond B) & (*) \\ & \to & \Box(\Box\neg B \to \Box\neg A) & (**) \end{array}$$

Evidently $A \rhd B \to A \rhd (B \wedge \Box\neg A) \vee (B \wedge \Diamond A)$. As clearly $B \wedge \Box\neg A \rhd B \wedge \Box\neg A$, we have shown $A \rhd B \to A \rhd B \wedge \Box\neg A$ once we have proven $B \wedge \Diamond A \rhd B \wedge \Box\neg A$. But, by $(*)$,

$$\begin{array}{rcll} B \wedge \Diamond A & \rhd & B \wedge \Diamond B & \text{by } \mathsf{L}_3 \\ & \rhd & B \wedge \Diamond(B \wedge \Box\neg B) & \\ & \rhd & B \wedge \Box\neg B & \text{by } (**) \\ & \rhd & B \wedge \Box\neg A. & \end{array}$$

⊣

Lemma 4.2.12. **IL**$\mathsf{M} \vdash \mathsf{W}$

Proof. By M, $A \rhd B \to A \wedge \Box\neg A \rhd B \wedge \Box\neg A$. But $A \rhd A \wedge \Box\neg A$, whence $A \rhd B \to A \rhd B \wedge \Box\neg A$. ⊣

P-style soundness proof of W We just follow the modal proof of W in **IL**P. At some places, axioms are replaced by there counterparts that deal with finite approximations.

By $\mathsf{P^k}$ we have that for some k,

$$\begin{array}{rcll} \alpha \rhd \beta & \to & \Box(\alpha \rhd^k \beta) & \text{by } \mathsf{J}_4^\mathsf{k} \\ & \to & \Box(\Diamond\alpha \to \Diamond^k\beta) & (*) \\ & \to & \Box(\Box^k\neg\beta \to \Box\neg\alpha). & (**) \end{array}$$

Now $\alpha \rhd \beta \to (\beta \wedge \Box\neg\alpha) \vee (\beta \wedge \Diamond\alpha)$. Starting from the last disjunct we obtain by $(*)$

$$\begin{array}{rcll} \beta \wedge \Diamond\alpha & \rhd & \beta \wedge \Diamond^k\beta & \text{by } \mathsf{L}_3^\mathsf{k} \\ & \rhd & \Diamond^k(\beta \wedge \Box^k\neg\beta) & \text{by } \mathsf{J}_5^\mathsf{k} \text{ and } (\mathsf{E}\rhd)^\mathsf{k} \\ & \rhd & \beta \wedge \Box^k\neg\beta & \text{by } (**) \\ & \rhd & \beta \wedge \Box\neg\alpha. & \end{array}$$

M-style soundness proof of W We assume $j : \alpha \rhd \beta$ and fix the corresponding Pudlák cut J. By Lemma 4.2.4, $\alpha \rhd \alpha \wedge \Box^J\neg\alpha$, whence by $\mathsf{M^J}$ and J_2, $\alpha \rhd \beta \wedge \Box\neg\alpha$.

The principle $\mathsf{M_0}$

Lemma 4.2.13. $\mathbf{IL}\mathsf{P} \vdash \mathsf{M_0}$ *and* $\mathbf{IL}\mathsf{P_R} \vdash \mathsf{M_0}$

Proof.

$$\begin{array}{rcl}
A \rhd B & \to & \Box(A \rhd B) \\
 & \to & \Box(\Diamond A \to \Diamond B) \\
 & \to & \Box(\Diamond A \wedge \Box C \to \Diamond B \wedge \Box C) \\
 & \to & \Diamond A \wedge \Box C \rhd \Diamond B \wedge \Box C \\
 & \to & \Diamond A \wedge \Box C \rhd \Diamond(B \wedge \Box C) \\
 & \to & \Diamond A \wedge \Box C \rhd B \wedge \Box C
\end{array}$$

⊣

Lemma 4.2.14. $\mathbf{IL}\mathsf{M} \vdash \mathsf{M_0}$

Proof. $A \rhd B \to A \wedge \Box C \rhd B \wedge \Box C$. But, $\Diamond A \wedge \Box C \rhd \Diamond(A \wedge \Box C) \rhd A \wedge \Box C$, whence $A \rhd B \to \Diamond A \wedge \Box C \rhd B \wedge \Box C$. ⊣

P-style soundness proof of $\mathsf{M_0}$ Starting with an application from $\mathsf{P^k}$, for some k we obtain the following reasoning.

$$\begin{array}{rcll}
\alpha \rhd \beta & \to & \Box(\alpha \rhd^k \beta) & \mathsf{J^k_4} \\
 & \to & \Box(\Diamond\alpha \to \Diamond^k\beta) & \\
 & \to & \Box(\Diamond\alpha \wedge \Box\gamma \to \Diamond^k\beta \wedge \Box\gamma) & \\
 & \to & \Diamond\alpha \wedge \Box\gamma \rhd \Diamond^k\beta \wedge \Box\gamma & \text{a.o. by } \mathsf{L^k_2} \\
 & \to & \Diamond\alpha \wedge \Box\gamma \rhd \Diamond^k(\beta \wedge \Box\gamma) & \text{by } \mathsf{J^k_5} \text{ and } (\mathsf{E}\rhd)^{\mathsf{k}} \\
 & \to & \Diamond\alpha \wedge \Box\gamma \rhd \beta \wedge \Box\gamma &
\end{array}$$

M-style soundness proof of $\mathsf{M_0}$ $\alpha \rhd \beta \to \alpha \wedge \Box^J\gamma \rhd \beta \wedge \Box\gamma$ for some cut J. By $\mathsf{L^J_{2a}}$ for this particular J we get $\Diamond\alpha \wedge \Box\gamma \to \Diamond\alpha \wedge \Box\Box^J\gamma$, whence

$$\begin{array}{rcll}
\Diamond\alpha \wedge \Box\gamma & \rhd & \Diamond\alpha \wedge \Box\Box^J\gamma & \\
 & \rhd & \Diamond(\alpha \wedge \Box^J\gamma) & \\
 & \rhd & \alpha \wedge \Box^J\gamma & \text{by } \mathsf{M^J} \\
 & \rhd & \beta \wedge \Box\gamma. &
\end{array}$$

The principle $\mathsf{W^*}$

Lemma 4.2.15. $\mathbf{IL}\mathsf{P} \vdash \mathsf{W^*}$ *and* $\mathbf{IL}\mathsf{P_R} \vdash \mathsf{W^*}$

Proof. In $\mathbf{IL}\mathsf{P}$ (resp. $\mathbf{IL}\mathsf{P_R}$): if $A \rhd B$, then

$$\Box(\Box\neg B \to \Box\neg A) \tag{4.3}$$

and

$$\Box(\Diamond A \wedge \Box C \to \Diamond B \wedge \Box C). \tag{4.4}$$

Thus, $B \wedge \Box C \rhd (B \wedge \Box C \wedge \Box\neg A) \vee (B \wedge \Box C \wedge \Diamond A)$. Again, in the first case we would be done. In the second case we get the following reasoning.

$$\begin{array}{lcll}
B \wedge \Box C \wedge \Diamond A & \rhd & \Diamond A \wedge \Box C & \text{by (4.4)} \\
 & \rhd & \Diamond B \wedge \Box C & \text{by } \mathsf{L}_3 \\
 & \rhd & \Diamond (B \wedge \Box\neg B) \wedge \Box C & \text{by } \mathsf{L}_2 \\
 & \rhd & \Diamond (B \wedge \Box C \wedge \Box\neg B) & \text{by } \mathsf{J}_5 \\
 & \rhd & B \wedge \Box C \wedge \Box\neg B & \text{by (4.3)} \\
 & \rhd & B \wedge \Box C \wedge \Box\neg A &
\end{array}$$

$\dashv$

Lemma 4.2.16. **IL**$\mathsf{M} \vdash \mathsf{W}^*$

Proof. So, in **IL**M, assume $A \rhd B$. By an application of M we get $B \wedge \Box C \rhd (B \wedge \Box C \wedge \Box\neg A) \vee (B \wedge \Box C \wedge \Diamond A)$. Again, in the first case we would be done. In the second case we get the following reasoning.

$$\begin{array}{lcll}
B \wedge \Box C \wedge \Diamond A & \rhd & \Diamond A \wedge \Box C & \text{by } \mathsf{L}_3 \\
 & \rhd & \Diamond (A \wedge \Box\neg A) \wedge \Box C & \text{by } \mathsf{L}_2 \\
 & \rhd & \Diamond (A \wedge \Box C \wedge \Box\neg A) & \text{by } \mathsf{J}_5 \\
 & \rhd & A \wedge \Box C \wedge \Box\neg A & \text{by } \mathsf{M} \text{ and } A \rhd B \\
 & \rhd & B \wedge \Box C \wedge \Box\neg A &
\end{array}$$

$\dashv$

P-style soundness proof of W^* For some k we get starting with an application of P^{k} the following reasoning.

$$\begin{array}{lcll}
\alpha \rhd \beta & \rightarrow & \Box(\alpha \rhd^k \beta) & \text{by } \mathsf{J}_4^{\mathsf{k}} \\
 & \rightarrow & \Box(\Diamond\alpha \rightarrow \Diamond^k \beta) & \\
 & \rightarrow & \Box(\Box^k \neg\beta \rightarrow \Box\neg\alpha) & (*) \\
 & \rightarrow & \Box(\Diamond\alpha \wedge \Box\gamma \rightarrow \Diamond^k \beta \wedge \Box\gamma) & (**)
\end{array}$$

We follow the modal proof.

$$\begin{array}{lcll}
\beta \wedge \Box\gamma \wedge \Diamond\alpha & \rhd & \Diamond\alpha \wedge \Box\gamma & \text{by } (**) \\
 & \rhd & \Diamond^k \beta \wedge \Box\gamma & \text{by } \mathsf{L}_3^{\mathsf{k}} \\
 & \rhd & \Diamond^k (\beta \wedge \Box^k \neg\beta) \wedge \Box\gamma & \text{by } \mathsf{L}_2^{\mathsf{k}} \\
 & \rhd & \Diamond^k (\beta \wedge \Box\gamma \wedge \Box^k \neg\beta) & \text{by } \mathsf{J}_5^{\mathsf{k}} \text{ and } (\mathsf{E}\rhd)^{\mathsf{k}} \\
 & \rhd & \beta \wedge \Box\gamma \wedge \Box^k \neg\beta & \text{by } (*) \\
 & \rhd & \beta \wedge \Box\gamma \wedge \Box\neg\alpha &
\end{array}$$

M-style soundness proof of W^* Also following the modal proof. Let J be the Pudlák cut of $j : \alpha \rhd \beta$. We get the following reasoning.

$$\begin{array}{lcll}
\beta \wedge \Box\gamma \wedge \Diamond\alpha & \rhd & \Diamond\alpha \wedge \Box\gamma & \text{by } \mathsf{L}_{\mathsf{3a}}^{\mathsf{J}} \\
 & \rhd & \Diamond(\alpha \wedge \Box^J \neg\alpha) \wedge \Box\gamma & \text{by } \mathsf{L}_{\mathsf{2a}}^{\mathsf{J}} \\
 & \rhd & \Diamond(\alpha \wedge \Box^J \gamma \wedge \Box^J \neg\alpha) & \text{by } \mathsf{J}_5 \\
 & \rhd & \alpha \wedge \Box^J \gamma \wedge \Box^J \neg\alpha & \text{by } \mathsf{M}^{\mathsf{J}} \text{ and } j : \alpha \rhd \beta \\
 & \rhd & \beta \wedge \Box\gamma \wedge \Box\neg\alpha &
\end{array}$$

The principle $\mathsf{P_0}$

Lemma 4.2.17. $\mathbf{IL}\mathsf{P} \vdash \mathsf{P_0}$

Proof. Within $\mathbf{IL}\mathsf{P}$: $A \rhd \Diamond B \to \Box(A \rhd \Diamond B) \to \Box(A \rhd B)$. ⊣

Lemma 4.2.18. $\mathbf{IL}\mathsf{P_R} \nvdash \mathsf{P_0}$

Proof. It is easy to see that frames satisfying $uRxRyS_uz \to xRz$ are sound for $\mathbf{IL}\mathsf{P_R}$. And it is equally easy to provide such a model on which $\mathsf{P_0}$ does not hold. ⊣

Lemma 4.2.18 nicely reflects that the frame condition for $\mathsf{P_0}$ essentially involves new S-transitions.

Lemma 4.2.19. $\mathbf{IL}\mathsf{M} \vdash \mathsf{P_0}$

Proof.

$$\begin{array}{rcl} A \rhd \Diamond B & \to & A \wedge \Box\neg B \rhd \bot \\ & \to & \Box(A \to \Diamond B) \\ & \to & \Box\Box(A \to \Diamond B) \\ & \to & \Box(A \rhd \Diamond B) \\ & \to & \Box(A \rhd B) \end{array}$$

⊣

P-style soundness proof of $\mathsf{P_0}$ The proof goes conform the modal proof. Thus, for some k, $\alpha \rhd \Diamond\beta \to \Box(\alpha \rhd^k \Diamond\beta)$. Hence, by $\mathsf{J_5^k}$ we get $\alpha \rhd \Diamond\beta \to \Box(\alpha \rhd \beta)$.

M-style soundness proof of $\mathsf{P_0}$ Again, we follow the modal proof. Thus, for some cut J we get the following.

$$\begin{array}{rcl} A \rhd \Diamond B & \to & A \wedge \Box^J\neg B \rhd \bot \\ & \to & \Box(A \to \Diamond^J B) \\ & \to & \Box\Box(A \to \Diamond^J B) \\ & \to & \Box(A \rhd \Diamond^J B) \\ & \to & \Box(A \rhd B) \end{array}$$

Note: the principle $A \rhd \Diamond B \to \Box(A \rhd \Diamond B)$ is also provable in both $\mathbf{IL}\mathsf{M}$ and $\mathbf{IL}\mathsf{P}$. In [Vis97] it is shown that this principle is not valid in PRA. It is nice to see where proof-attempts of this principle in our systems fail.

The principle R

Before we see that $\mathbf{IL}\mathsf{P} \vdash \mathsf{R}$, we first proof an auxiliary lemma.

Lemma 4.2.20. $\mathbf{IL} \vdash \neg(A \rhd \neg C) \wedge (A \rhd B) \to \Diamond(B \wedge \Box C)$

Proof. We prove the logical equivalent $(A \rhd B) \wedge \Box(B \to \Diamond\neg C) \to A \rhd \neg C$ in **IL**. But this is clear, as $(A \rhd B) \wedge \Box(B \to \Diamond\neg C) \to A \rhd B \wedge \Diamond\neg C$ and $\Diamond\neg C \rhd \neg C$. ⊣

Lemma 4.2.21. **IL**$\mathsf{P} \vdash \mathsf{P_0}$

Proof. $A \rhd B \to \Box(A \rhd B)$. Using this together with Lemma 4.2.20 we get that under the assumption $A \rhd B$, we have

$$\begin{array}{lll} \neg(A \rhd \neg C) & \rhd & \neg(A \rhd \neg C) \wedge (A \rhd B) \\ & \rhd & \Diamond(B \wedge \Box C) \\ & \rhd & B \wedge \Box C. \end{array}$$

⊣

Lemma 4.2.22. **IL**$\mathsf{P_R} \nvdash \mathsf{R}$

Proof. By exposing a countermodel as in the proof of Lemma 4.2.18. ⊣

Lemma 4.2.23. **IL**$\mathsf{M} \vdash \mathsf{R}$

Proof. In **IL** it is easy to see that $\neg(A \rhd \neg C) \to \Diamond(A \wedge \Box C)$. Thus, if $A \rhd B$ then

$$\begin{array}{lll} \neg(A \rhd \neg C) & \rhd & \Diamond(A \wedge \Box C) \\ & \rhd & A \wedge \Box C \\ & \rhd & B \wedge \Box C. \end{array}$$

⊣

P-style soundness proof of R Conform the modal proof, we first see that $(\alpha \rhd^k \beta) \wedge \neg(\alpha \rhd \neg\gamma) \to \Diamond^k(\beta \wedge \Box\gamma)$. For, suppose that $\alpha \rhd^k \beta$ and $\Box^k(\beta \to \Diamond\neg\gamma)$. Then, by $\mathsf{J}^{\mathsf{k}}_{\mathsf{2b}}$, $\alpha \rhd^k \Diamond\neg\gamma$. Thus, by $\mathsf{J}^{\mathsf{k}}_{\mathsf{5}}$, we get $\alpha \rhd \neg\gamma$. We have not used any extensionality axioms, thus also

$$\Box((\alpha \rhd^k \beta) \wedge \neg(\alpha \rhd \neg\gamma) \to \Diamond^k(\beta \wedge \Box\gamma)). \tag{4.5}$$

We now turn to the main proof. So, suppose $k : \alpha \rhd \beta$, then $\Box(\alpha \rhd^k \beta)$ and thus

$$\begin{array}{llll} \neg(\alpha \rhd \neg\gamma) & \rhd & \neg(\alpha \rhd \neg\gamma) \wedge (\alpha \rhd^k \beta) & \text{by (4.5)} \\ & \rhd & \Diamond^k(\beta \wedge \Box\gamma) & \text{by } \mathsf{J}^{\mathsf{k}}_{\mathsf{5}} \text{ and } (\mathsf{E}\rhd)^{\mathsf{k}} \\ & \rhd & \beta \wedge \Box\gamma. & \end{array}$$

M-style soundness proof of R Again following the modal poof. So, suppose that $j : \alpha \rhd \beta$ and let J be the corresponding Pudlák cut. By Lemma 4.2.5 we get that for this cut $\neg(\alpha \rhd \neg\gamma) \to \Diamond(\alpha \wedge \Box^J\gamma)$. Thus, if $j : \alpha \rhd \beta$ then

$$\begin{array}{lll} \neg(\alpha \rhd \neg\gamma) & \rhd & \Diamond(\alpha \wedge \Box^J\gamma) \\ & \rhd & \alpha \wedge \Box^J\gamma \\ & \rhd & \beta \wedge \Box\gamma. \end{array}$$

Mixing proof styles

Sometimes, mixing P and M-style proofs can be fruitful. The next lemma provides an example.

Lemma 4.2.24. *In any reasonable arithmetical theory we have that* $\alpha \rhd \Diamond\beta \to \Box(\neg(\alpha \rhd \neg\gamma) \to \Diamond(\beta \wedge \gamma))$.

Proof. Suppose $k : \alpha \rhd \Diamond\beta$ and let K be the corresponding Pudlák cut. Then, by M^{J} we get

$$\begin{array}{lll} \alpha \rhd \Diamond\beta & \to & \alpha \wedge \Box^K\gamma \rhd \Diamond\beta \wedge \Box\gamma \\ & \to & \alpha \wedge \Box^K\gamma \rhd \Diamond(\beta \wedge \gamma) \qquad \text{by } \mathsf{P}^{\mathsf{k}} \\ & \to & \Box(\alpha \wedge \Box^K\gamma \rhd^k \Diamond(\beta \wedge \gamma)) \qquad \text{by } \mathsf{J}_4^{\mathsf{k}} \\ & \to & \Box(\Diamond(\alpha \wedge \Box^K\gamma) \to \Diamond^k\Diamond(\beta \wedge \gamma)) \qquad \text{by } \mathsf{L}_2^{\mathsf{k}} \\ & \to & \Box(\Diamond(\alpha \wedge \Box^K\gamma) \to \Diamond(\beta \wedge \gamma)). \end{array}$$

But, by Lemma 4.2.5, we get $\Box(\neg(\alpha \rhd \neg\gamma) \to \Diamond(\alpha \wedge \Box^K\gamma))$ and we are done. ⊣

It is not hard to see that the above principle is already provable in **IL**R.

Lemma 4.2.25. **IL**$\mathsf{R} \vdash A \rhd B \to \neg(A \rhd \neg C) \wedge \Box D \rhd B \wedge \Box(C \wedge D)$

Proof. One easily sees that **IL** $\vdash \neg(A \rhd \neg C) \wedge \Box D \to \neg(A \rhd \neg(C \wedge D))$. One application of R now gives the desired result. ⊣

Lemma 4.2.26. **IL**$\mathsf{R} \vdash A \rhd \Diamond B \to \Box(\neg(A \rhd \neg C) \to \Diamond(B \wedge C))$

Proof. In **IL**R we get

$$\begin{array}{lll} A \rhd \Diamond B & \to & \neg(A \rhd \neg C) \rhd \Diamond B \wedge \Box C \\ & \to & \neg(A \rhd \neg C) \rhd \Diamond(B \wedge C) \\ & \to & \neg(A \rhd \neg C) \wedge \Box\neg(B \wedge C) \rhd \bot \\ & \to & \Box(\neg(A \rhd \neg C) \to \Diamond(B \wedge C)). \end{array}$$

⊣

It is also not hard to see that $A \rhd \Diamond B \to \Box(\neg(A \rhd \neg C) \to \Diamond(B \wedge C))$ follows semantically from the frame condition of R.

Part II

Modal matters in interpretability logics

Chapter 5

The construction method

In Part II of the thesis we shall study the modal semantics for interpretability logics. This semantics has proven to be quite a good one. However, there still seem to be some draw-backs. That is, elementary modal properties can easily lead to tremendous technical complications. Why is it, that modal completeness proofs for $\mathbf{IL}\mathsf{W}$ and $\mathbf{IL}\mathsf{M}_0$ are so difficult?

In part, we think, this is because the right technical aparatus has not yet been fully developed. A step in the good direction is made with the introduction of the so-called *full labels* in Section 8.2. However, we did not use these full labels in earlier sections. This is due to the simple reason that we had not yet invented/discovered them at the time of writing. Large part of those sections can thus, so we are convinced, be simplified.

Historical evidence makes that we indeed believe in the semantics. Two times, a new arithmetical principle was found on the basis of modal considerations only: the principle P_0 from [Joo98] and the new principle R from [GJ04].

In this chapter we describe and discuss the standard construction method (step-by-step method) to obtain, amongst others, modal completeness results. We conclude the chapter with a modal completeness proof of $\mathbf{IL}$.

5.1 General exposition of the construction method

Most of the applications of the construction method deal with modal completeness of a certain logic $\mathbf{IL}\mathsf{X}$. More precisely, showing that a logic $\mathbf{IL}\mathsf{X}$ is modally complete amounts to constructing, or finding, whenever $\mathbf{IL}\mathsf{X} \nvdash \varphi$, a model M and an $x \in M$ such that $M, x \Vdash \neg\varphi$. We will employ our construction method for this particular model construction.

In this section, we will not always give precise definitions of the notions we work with. All the definitions can be found in Section 5.2.

5.1.1 The main ingredients of the construction method

As we mentioned above, a modal completeness proof of a logic **IL**X amounts to a uniform model construction to obtain $M, x \Vdash \neg\varphi$ for **IL**X $\nvdash \varphi$. If **IL**X $\nvdash \varphi$, then $\{\neg\varphi\}$ is an **IL**X-consistent set and thus, by a version of Lindenbaum's Lemma (Lemma 3.2.8), it is extendible to a maximal **IL**X-consistent set. On the other hand, once we have an **IL**X-model $M, x \Vdash \neg\varphi$, we can find, by Corollary 3.3.9 a maximal **IL**X-consistent set Γ with $\neg\varphi \in \Gamma$. This Γ can simply be defined as the set of all formulas that hold at x.

To go from a maximal **IL**X-consistent set to a model is always the hard part. This part is carried out in our construction method. In this method, the maximal consistent set is somehow partly unfolded to a model.

Often in these sort of model constructions, the worlds in the model are MCS's. For propositional variables one then defines $x \Vdash p$ iff $p \in x$. In the setting of interpretability logics it is sometimes inevitable to use the same MCS in different places in the model.[1] Therefore we find it convenient not to identify a world x with a MCS, but rather label it with a MCS $\nu(x)$. However, we will still write sometimes $\varphi \in x$ instead of $\varphi \in \nu(x)$.

One complication in unfolding a MCS to a model lies in the incompactness of the modal logics we consider. This, in turn, is due to the fact that some frame conditions are not expressible in first order logic. As an example we can consider the following set.[2]

$$\Gamma := \{\Diamond p_0\} \cup \{\Box(p_i \rightarrow \Diamond p_{i+1}) \mid i \in \omega\}$$

Clearly, Γ is a **GL**-consistent set, and any finite part of it is satisfiable in some world in some model. However, it is not hard to see that in no **IL**-model all of Γ can hold simultaneously in some world in it.

If M is an **IL**X-model and $x \in M$, then $\{\varphi \mid M, x \Vdash \varphi\}$ is a MCS. By definition (and abuse of notation) we see that

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

We call this equivalence a truth lemma. (See for example Definition 5.2.5 for a more precise formulation.) In all completeness proofs a model is defined or constructed in which some form of a truth lemma holds. Now, by the observed incompactness phenomenon, we can not expect that for every MCS, say Γ, we can find a model "containing" Γ for which a truth lemma holds in full generality. There are various ways to circumvent this complication. Often one considers truncated parts of maximal consistent sets which are finite. In choosing how to truncate, one is driven by two opposite forces.

[1] As the truth definition of $A \rhd B$ has a $\forall\exists$ character, the corresponding notion of bisimulation is rather involved. As a consequence there is in general no obvious notion of a minimal bisimular model, contrary to the case of provability logics. This causes the necessity of several occurrences of MCS's.

[2] This example comes from Fine and Rautenberg and is treated in Chapter 7 of [Boo93].

On the one hand this truncated part should be small. It should be at least finite so that the incompactness phenomenon is blocked. The finiteness is also a desideratum if one is interested in the decidability of a logic.

On the other hand, the truncated part should be large. It should be large enough to admit inductive reasoning to prove a truth lemma. For this, often closure under subformulas and single negation suffices. Also, the truncated part should be large enough so that MCS's contain enough information to do the required calculation. For this, being closed under subformulas and single negations does not, in general, suffice. Examples of these sort of calculation are Lemma 6.1.7 and Lemma 7.1.16.

In our approach we take the best of both opposites. That is, we do not truncate at all. Like this, calculation becomes uniform, smooth and relatively easy. However, we demand a truth lemma to hold only for finitely many formulas.

The question is now, how to unfold the MCS containing $\neg\varphi$ to a model where $\neg\varphi$ holds in some world. We would have such a model if a truth lemma holds w.r.t. a finite set $\mathcal{D}$ containing $\neg\varphi$.

Proving that a truth lemma holds is usually done by induction on the complexity of formulas. As such, this is a typical "bottom up" or "inside out" activity. On the other hand, unfolding, or reading off, the truth value of a formula is a typical "top down" or "outside in" activity.

Yet, we do want to gradually build up a model so that we get closer and closer to a truth lemma. But, how could we possibly measure that we come closer to a truth lemma? Either everything is in place and a truth lemma holds, or a truth lemma does not hold, in which case it seems unclear how to measure to what extend it does not hold.

The gradually building up a model will take place by consecutively adding bits and pieces to the MCS we started out with. Thus somehow, we do want to measure that we come closer to a truth lemma by doing so. Therefore, we switch to an alternative forcing relation $\|\!\!\sim$ that follows the "outside in" direction that is so characteristic to the evaluation of $x \Vdash \varphi$, but at the same time incorporates the necessary elements of a truth lemma.

$$\begin{array}{lll l}
x\|\!\!\sim p & \text{iff} & p \in x & \text{for propositional variables } p \\
x\|\!\!\sim \varphi \wedge \psi & \text{iff} & x\|\!\!\sim\varphi \ \&\ x\|\!\!\sim\psi & \text{and likewise for} \\
 & & & \text{other boolean connectives} \\
x\|\!\!\sim \varphi \rhd \psi & \text{iff} & \multicolumn{2}{l}{\forall y\ [xRy \wedge \varphi \in x \to \exists z\ (yS_x z \wedge \psi \in z)]}
\end{array}$$

If $\mathcal{D}$ is a set of sentences that is closed under subformulas and single negations, then it is not hard to see that (see Lemma 5.2.9)

$$\forall x \forall\, \varphi{\in}\mathcal{D}\ [x\|\!\!\sim\varphi \text{ iff } \varphi \in x] \quad (*)$$

is equivalent to

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

Thus, if we want to obtain a truth lemma for a finite set $\mathcal{D}$ that is closed under single negations and subformulas, we are done if we can obtain $(*)$. But now it is clear how we can at each step measure that we come closer to a truth lemma. This brings us to the definition of problems and deficiencies.

A problem is some formula $\neg(\varphi \rhd \psi) \in x \cap \mathcal{D}$ such that $x \not\Vdash \neg(\varphi \rhd \psi)$. We define a deficiency to be a configuration such that $\varphi \rhd \psi \in x \cap \mathcal{D}$ but $x \not\Vdash \varphi \rhd \psi$. It now becomes clear how we can successively eliminate problems and deficiencies.

A deficiency $\varphi \rhd \psi \in x \cap \mathcal{D}$ is a deficiency because there is some y (or maybe more of them) with xRy, and $\varphi \in y$, but for no z with yS_xz, we have $\psi \in z$. This can simply be eliminated by adding a z with yS_xz and $\psi \in z$.

A problem $\neg(\varphi \rhd \psi) \in x \cap \mathcal{D}$ can be eliminated by adding a completely isolated y to the model with xRy and $\varphi, \neg\psi \in y$. As y is completely isolated, $yS_xz \Rightarrow z = y$ and thus indeed, it is not possible to reach a world where ψ holds. Now here is one complication.

We want that a problem or a deficiency, once eliminated, can never re-occur. For deficiencies this complication is not so severe, as the quantifier complexity is $\forall\exists$. Thus, any time "a deficiency becomes active", we can immediately deal with it.

With the elimination of a problem, things are more subtle. When we introduced $y \ni \varphi, \neg\psi$ to eliminate a problem $\neg(\varphi \rhd \psi) \in x \cap \mathcal{D}$, we did indeed eliminate it, as for no z with yS_xz we have $\psi \in z$. However, this should hold for any future expansion of the model too. Thus, any time we eliminate a problem $\neg(\varphi \rhd \psi) \in x \cap \mathcal{D}$, we introduce a world y with a promise that in no future time we will be able to go to a world z containing ψ via an S_x-transition. Somehow we should keep track of all these promises throughout the construction and make sure that all the promises are indeed kept. This is taken care of by our so called ψ-critical cones (see for example also [dJJ98]). As ψ is certainly not allowed to hold in R-successors of y, it is reasonable to demand that $\Box\neg\psi \in y$. (Where y was introduced to eliminate the problem $\neg(\varphi \rhd \psi) \in x \cap \mathcal{D}$.)

Note that problems have quantifier complexity $\exists\forall$. We have chosen to call them problems due to their prominent existential nature.

5.1.2 Some methods to obtain completeness

For modal logics in general, quite an arsenal of methods to obtain completeness is available. For instance the standard operations on canonical models like path–coding (unraveling), filtrations and bulldozing (see [BRV01]). Or one can mention uniform methods like the use of Shalqvist formulas or the David Lewis theorem [Boo93]. A very secure method is to construct counter models piece by piece. A nice example can be found in [Boo93], Chapter 10. In [HMV01] and in [HH02] a step-by-step method is exposed in the setting of universal algebras. New approximations of the model are given by moves in an (infinite) game.

For interpretability logics the available methods are rather limited in number. In the case of the basic logic **IL** a relatively simple unraveling works. Although **IL**M does allow a same treatment, the proof is already much less clear. (For both proofs, see [dJJ98]). However, for logics that contain **IL**M_0 but not **IL**M it is completely unclear how to obtain completeness via an unraveling and we are forced into more secure methods like the above mentioned building of models piece by piece. And this is precisely what we do in this paper.

Decidability and the finite model property are two related issues that more or less seem to divide the landscape of interpretability logics into the same classes. That is, the proof that **IL** has the finite model property is relatively easy. The same can be said about **IL**M. For logics like **IL**M_0 the issue seems much more involved and a proper proof of the finite model property, if one exists at all, has not been given yet. Alternatively, one could resort to other methods for showing decidability like the Mosaic method [BRV01].

5.2 The construction method

5.2.1 Preparing the construction

An **IL**X-labeled frame is just a Veltman frame in which every node is labeled by a maximal **IL**X-consistent set and some R-transitions are labeled by a formula. R-transitions labeled by a formula C indicate that some C-criticallity is essentially present at this place.

Definition 5.2.1. An **IL**X-*labeled frame* is a quadruple $\langle W, R, S, \nu\rangle$. Here $\langle W, R, S\rangle$ is an **IL**-frame and ν is a labeling function. The function ν assigns to each $x \in W$ a maximal **IL**X-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 there is no chance of confusion we will just speak of labeled frames or even just of frames rather than **IL**X-labeled frames. Labeled frames inherit all the terminology and notation from normal frames. Note that an **IL**X-labeled frame need not be, and shall in general not be, an **IL**X-frame. If we speak about a labeled **IL**X-frame we always mean an **IL**X-labeled **IL**X-frame. To indicate that $\nu(\langle x, y\rangle) = A$ we will sometimes write xR^Ay or $\nu(x, y) = A$.

Formally, given $F = \langle W, R, S, \nu\rangle$, one can see ν as a subset of $(W \cup (W \times W)) \times (\mathsf{Form}_{\mathbf{IL}} \cup \{\Gamma \mid \Gamma$ is a maximal **IL**X consistent set$\})$ such that the following properties hold.

- $\forall\, x{\in}W\ (\langle x, y\rangle \in \nu \Rightarrow y$ is a MCS)
- $\forall\, \langle x, y\rangle{\in}W \times W\ (\langle\langle x, y\rangle, z\rangle \in \nu \Rightarrow z$ is a formula)
- $\forall\, x{\in}W \exists y\ \langle x, y\rangle \in \nu$
- $\forall x, y, y'(\langle x, y\rangle \in \nu \wedge \langle x, y'\rangle \in \nu \rightarrow y = y')$

We will often regard ν as a partial function on $W \cup (W \times W)$ which is total on W and which has its values in $\mathsf{Form}_{\mathbf{IL}} \cup \{\Gamma \mid \Gamma$ is a maximal **IL**X consistent set$\}$

Remark 5.2.2. Every **IL**X-labeled frame $F = \langle W, R, S, \nu\rangle$ can be transformed to an **IL**-model $\overline{F}$ in a uniform way by defining for propositional variables p the valuation as $\overline{F}, x \Vdash p$ iff $p \in \nu(x)$. By Corollary 3.3.9 we can also regard any model M satisfying the **IL**X frame condition[3] as an **IL**X-labeled frame $\overline{M}$ by defining $\nu(m) := \{\varphi \mid M, m \Vdash \varphi\}$.

We sometimes refer to $\overline{F}$ as the model induced by the frame F. Alternatively we will speak about the model corresponding to F. Note that for **IL**X-models M, we have $\overline{\overline{M}} = M$, but in general $\overline{\overline{F}} \neq F$ for **IL**X-labeled frames F.

Definition 5.2.3. Let x be a world in some **IL**X-labeled frame $\langle W, R, S, \nu\rangle$. The *$C$-critical cone above x*, we write $\mathcal{C}_x^C$, is defined inductively as

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

Definition 5.2.4. Let x be a world in some **IL**X-labeled frame $\langle W, R, S, \nu\rangle$. The *generalized C-cone above x*, we write $\mathcal{G}_x^C$, is defined inductively as

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

It follows directly from the definition that the C-critical cone above x is part of the generalized C-cone above x. So, if $\mathcal{G}_x^B \cap \mathcal{G}_x^C = \varnothing$, then certainly $\mathcal{C}_x^B \cap \mathcal{C}_x^C = \varnothing$.

We also note that there is some redundancy in Definitions 5.2.3 and 5.2.4. The last clause in the inductive definitions demands closure of the cone under R-successors. But from Definition 3.3.1.5 closure of the cone under R follows from closure of the cone under S_x. We have chosen to explicitly adopt the closure under the R. In doing so, we obtain a notion that serves us also in the environment of so-called quasi frames (see Definition 5.3.1) in which not necessarily $(x{\upharpoonright})^2 \cap R \subseteq S_x$.

Definition 5.2.5. Let $F = \langle W, R, S, \nu\rangle$ be a labeled frame and let $\overline{F}$ be the induced **IL**-model. Furthermore, let $\mathcal{D}$ be some set of sentences. We say that *a truth lemma holds in F with respect to $\mathcal{D}$* if $\forall\, A{\in}\mathcal{D}\ \forall\, x{\in}\overline{F}$

$$\overline{F}, x \Vdash A \Leftrightarrow A \in \nu(x).$$

If there is no chance of confusion we will omit some parameters and just say "a truth lemma holds at F" or even "a truth lemma holds". The following definitions give us a means to measure how far we are away from a truth lemma.

[3] We could even say, any **IL**X-model.

Definition 5.2.6 (Temporary definition). [4] Let $\mathcal{D}$ be some set of sentences and let $F = \langle W, R, S, \nu \rangle$ be an **IL**X-labeled frame. A *$\mathcal{D}$-problem* is a pair $\langle x, \neg(A \rhd B) \rangle$ such that $\neg(A \rhd B) \in \nu(x) \cap \mathcal{D}$ and for every y with xRy we have $[A \in \nu(y) \Rightarrow \exists z\ (yS_x z \wedge B \in \nu(z))]$.

Definition 5.2.7 (Deficiencies). Let $\mathcal{D}$ be some set of sentences and let $F = \langle W, R, S, \nu \rangle$ be an **IL**X-labeled frame. A *$\mathcal{D}$-deficiency* is a triple $\langle x, y, C \rhd D \rangle$ with xRy, $C \rhd D \in \nu(x) \cap \mathcal{D}$, and $C \in \nu(y)$, but for no z with $yS_x z$ we have $D \in \nu(z)$.

If the set $\mathcal{D}$ is clear or fixed, we will just speak about problems and deficiencies.

Definition 5.2.8. Let A be a formula. We define the *single negation* of A, we write $\sim\! A$, as follows. If A is of the form $\neg B$ we define $\sim\! A$ to be B. If A is not a negated formula we set $\sim\! A := \neg A$.

The next lemma shows that a truth lemma w.r.t. $\mathcal{D}$ can be reformulated in the combinatoric terms of deficiencies and problems. (See also the equivalence of $(*)$ and $(**)$ in Section 5.1.)

Lemma 5.2.9. *Let $F = \langle W, R, S, \nu \rangle$ be a labeled frame, and let $\mathcal{D}$ be a set of sentences closed under single negation and subformulas. A truth lemma holds in F w.r.t. $\mathcal{D}$ iff there are no $\mathcal{D}$-problems nor $\mathcal{D}$-deficiencies.*

Proof. The proof is really very simple and precisely shows the interplay between all the ingredients. ⊣

The labeled frames we will construct are always supposed to satisfy some minimal reasonable requirements. We summarize these in the notion of adequacy.

Definition 5.2.10 (Adequate frames). A frame is called *adequate* if the following conditions are satisfied.

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

If no confusion is possible we will just speak of frames instead of adequate labeled frames. As a matter of fact, all the labeled frames we will see from now on will be adequate. In the light of adequacy it seems reasonable to work with a slightly more elegant definition of a $\mathcal{D}$-problem.

Definition 5.2.11 (Problems). Let $\mathcal{D}$ be some set of sentences. A *$\mathcal{D}$-problem* is a pair $\langle x, \neg(A \rhd B) \rangle$ such that $\neg(A \rhd B) \in \nu(x) \cap \mathcal{D}$ and for no $y \in \mathcal{C}_x^B$ we have $A \in \nu(y)$.

[4] We will eventually work with Definition 5.2.11.

From now on, this will be our working definition. Clearly, on adequate labeled frames, if $\langle x, \neg(A \rhd B)\rangle$ is not a problem in the new sense, it is not a problem in the old sense.

Remark 5.2.12. It is also easy to see that the we still have the interesting half of Lemma 5.2.9. Thus, we still have, that a truth lemma holds if there are no deficiencies nor problems.

To get a truth lemma we have to somehow get rid of problems and deficiencies. This will be done by adding bits and pieces to the original labeled frame. Thus the notion of an extension comes into play.

Definition 5.2.13 (Extension). Let $F = \langle W, R, S, \nu\rangle$ be a labeled frame. We say that $F' = \langle W', R', S', \nu'\rangle$ is an *extension* of F, we write $F \subseteq F'$, if $W \subseteq W'$ and the relations in F' restricted to F yield the corresponding relations in F.

More formally, the requirements on the restrictions in the above definition amount to saying that for $x, y, z \in F$ we have the following.

- $xR'y$ iff xRy
- $yS'_x z$ iff $yS_x z$
- $\nu'(x) = \nu(x)$
- $\nu'(\langle x, y\rangle)$ is defined iff $\nu(\langle x, y\rangle)$ is defined, and in this case $\nu'(\langle x, y\rangle) = \nu(\langle x, y\rangle)$.

A problem in F is said to be *eliminated* by the extension F' if it is no longer a problem in F'. Likewise we can speak about elimination of deficiencies.

Definition 5.2.14 (Depth). The *depth* of a finite frame F, we will write $\mathsf{depth}(F)$ is the maximal length of sequences of the form $x_0 R \ldots R x_n$. (For convenience we define $\max(\varnothing) = 0$.)

The depth of a point is just the depth of the subframe generated by that point.

Definition 5.2.15 (Union of bounded chains). An indexed set $\{F_i\}_{i\in\omega}$ of labeled frames is called a *chain* if for all i, $F_i \subseteq F_{i+1}$. It is called a *bounded chain* if for some number n, $\mathsf{depth}(F_i) \leq n$ for all $i \in \omega$. The *union* of a bounded chain $\{F_i\}_{i\in\omega}$ of labeled frames F_i is defined as follows.

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

It is clear why we really need the boundedness condition. We want the union to be an **IL**-frame. So, certainly R should be conversely well-founded. This can only be the case if our chain is bounded.

5.2.2 The main lemma

We now come to the main motor behind many results. It is formulated in rather general terms so that it has a wide range of applicability. As a draw-back, we get that any application still requires quite some work.

Lemma 5.2.16 (Main lemma). *Let* $\mathbf{IL}\mathsf{X}$ *be an interpretability logic and let* $\mathcal{C}$ *be a (first or higher order) frame condition such that for any* $\mathbf{IL}$*-frame* F *we have*

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

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

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

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

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

Proof. By subsequently eliminating problems and deficiencies by means of extensions. These elimination processes have to be robust in the sense that every problem or deficiency that has been dealt with, should not possibly re-emerge. But, the requirements of the lemma almost immediately imply this.

For the second part of the Main Lemma, we suppose that for any finite set $\mathcal{D}$ closed under subformulas and single negations, we can find a corresponding set of invariants $\mathcal{I}$. If now, for any such $\mathcal{D}$, all the corresponding invariants $\mathcal{I}$ hold on any one-point labeled frame, we are to see that $\mathbf{IL}\mathsf{X}$ is a complete logic, that is, $\mathbf{IL}\mathsf{X} \nvdash A \Rightarrow \exists M\ (M \models X\ \&\ M \models \neg A)$.

But this just follows from the above. If $\mathbf{IL}\mathsf{X} \nvdash A$, we can find a maximal $\mathbf{IL}\mathsf{X}$-consistent set Γ with $\neg A \in \Gamma$. Let $\mathcal{D}$ be the smallest set that contains $\neg A$ and is closed under subformulas and single negations and consider the invariants corresponding to $\mathcal{D}$. The labeled frame $F := \langle \{x\}, \varnothing, \varnothing, \langle x, \Gamma \rangle \rangle$ can thus be extended to a labeled adequate **IL**X-frame $\hat{F}$ on which a truth lemma with respect to $\mathcal{D}$ holds. Thus certainly $\overline{\hat{F}}, x \Vdash \neg A$, that is, A is not valid on the model induced by $\hat{F}$. $\dashv$

The construction method can also be used to obtain decidability via the finite model property. In such a case, one should re-use worlds that were introduced earlier in the construction.

The following two lemmata indicate how good labels can be found for the elimination of problems and deficiencies.

Lemma 5.2.17. *Let Γ be a maximal* **IL**X*-consistent set such that $\neg(A \rhd B) \in \Gamma$. Then there exists a maximal* **IL**X*-consistent set Δ such that $\Gamma \prec_B \Delta \ni A, \Box\neg A$.*

Proof. So, consider $\neg(A \rhd B) \in \Gamma$, and suppose that no required Δ exists. We can then find a[5] formula C for which $C \rhd B \in \Gamma$ such that

$$\neg C, \Box\neg C, A, \Box\neg A \vdash_{\mathbf{IL}\mathsf{X}} \bot.$$

Consequently

$$\vdash_{\mathbf{IL}\mathsf{X}} A \wedge \Box\neg A \rightarrow C \vee \Diamond C$$

and thus, by Lemma 3.1.2, also $\vdash_{\mathbf{IL}\mathsf{X}} A \rhd C$. But as $C \rhd B \in \Gamma$, also $A \rhd B \in \Gamma$. This clearly contradicts the consistency of Γ. $\dashv$

For deficiencies there is a similar lemma.

Lemma 5.2.18. *Consider $C \rhd D \in \Gamma \prec_B \Delta \ni C$. There exists Δ' with $\Gamma \prec_B \Delta' \ni D, \Box\neg D$.*

Proof. Suppose for a contradiction that $C \rhd D \in \Gamma \prec_B \Delta \ni C$ and there does not exist a Δ' with $\Gamma \prec_B \Delta' \ni D, \Box\neg D$. Taking the contraposition of Lemma 5.2.17 we get that $\neg(D \rhd B) \notin \Gamma$, whence $D \rhd B \in \Gamma$ and also $C \rhd B \in \Gamma$. This clearly contradicts the consistency of Δ as $\Gamma \prec_B \Delta \ni C$. $\dashv$

5.2.3 Completeness and the main lemma

The main lemma provides a powerful method for proving modal completeness. In several cases it is actually the only known method available.

Remark 5.2.19. A modal completeness proof for an interpretability logic **IL**X is by the main lemma reduced to the following four ingredients.

[5] Writing out the definition and by compactness, we get a finite number of formulas $C_1, \ldots, C_n$ with $C_i \rhd B \in \Gamma$, such that $\neg C_1, \ldots, \neg C_n, \Box\neg C_1, \ldots, \Box\neg C_n, A, \Box\neg A \vdash_{\mathbf{IL}\mathsf{X}} \bot$. We can now take $C := C_1 \vee \ldots \vee C_n$. Clearly, as all the $C_i \rhd B \in \Gamma$, also $C \rhd B \in \Gamma$.

- **Frame Condition** Providing a frame condition $\mathcal{C}$ and a proof that

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

- **Invariants** Given a finite set of sentences $\mathcal{D}$ (closed under subformulas and single negations), providing invariants $\mathcal{I}$ that hold for any one-point labeled frame. Certainly $\mathcal{I}$ should contain $xRy \to \exists\, A \in (\nu(y) \setminus \nu(x)) \cap \{\Box D \mid D \in \mathcal{D}\}$.

- **elimination**

  - **Problems** Providing a procedure of elimination by extension for problems in labeled frames that satisfy all the invariants. This procedure should come with a proof that it preserves all the invariants.

  - **Deficiencies** Providing a procedure of elimination by extension for deficiencies in labeled frames that satisfy all the invariants. Also this procedure should come with a proof that it preserves all the invariants.

- **Rounding up** A proof that for any bounded chain of labeled frames that satisfy the invariants, automatically, the union satisfies the frame condition $\mathcal{C}$ of the logic.

The completeness proofs that we will present will all have the same structure, also in their preparations. As we will see, eliminating problems is more elementary than eliminating deficiencies.

As we already pointed out, we eliminate a problem by adding some new world plus an adequate label to the model we had. Like this, we get a structure that need not even be an **IL**-model. For example, in general, the R relation is not transitive. To come back to at least an **IL**-model, we should close off the new structure under transitivity of R and S et cetera. This closing off is in its self an easy and elementary process but we do want that the invariants are preserved under this process. Therefore we should have started already with a structure that admitted a closure. Actually in this paper we will always want to obtain a model that satisfies the frame condition of the logic.

The preparations to a completeness proof in Chapters 6 and 7 thus have the following structure.

- Determining a frame condition for **IL**X and a corresponding notion of an **IL**X-frame.

- Defining a notion of a quasi **IL**X-frame.

- Defining some notions that remain constant throughout the closing of quasi **IL**X-frames, but somehow capture the dynamic features of this process.

- Proving that a quasi **IL**X-frame can be closed off to an adequate labeled **IL**X-frame.

- Preparing the elimination of deficiencies.

The most difficult job in a the completeness proofs we present in this paper, was in finding correct invariants and in preparing the elimination of deficiencies. Once this is fixed, the rest follows in a rather mechanical way. Especially the closure of quasi **IL**X-frames to adequate **IL**X-frames is a very laborious enterprise.

5.3 The logic IL

The modal logic **IL** has been proved to be modally complete in [dJV90]. We shall reprove the completeness here using the main lemma.

The completeness proof of **IL** can be seen as the mother of all our completeness proofs in interpretability logics. Not only does it reflect the general structure of applications of the main lemma clearly, also it so that we can use large parts of the preparations to the completeness proof of **IL** in other proofs too. Especially closability proofs are cumulative. Thus, we can use the lemma that any quasi-frame is closable to an adequate frame, in any other completeness proof.

5.3.1 Preparations

Definition 5.3.1. A *quasi-frame* G is a quadruple $\langle W, R, S, \nu\rangle$. Here W is a non-empty set of worlds, and R a binary relation on W. S is a set of binary relations on W indexed by elements of W. The ν is a labeling as defined on labeled frames. Critical cones and generalized cones are defined just in the same way as in the case of labeled frames. G should posess the following properties.

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

Clearly, adequate labeled frames are special cases of quasi frames. Quasi-frames inherit all the notations from labeled frames. In particular we can thus speak of chains and the like.

Lemma 5.3.2 (IL-closure). *Let $G = \langle W, R, S, \nu\rangle$ be a quasi-frame. There is an adequate* **IL***-frame F extending G. That is, $F = \langle W, R', S', \nu\rangle$ with $R \subseteq R'$ and $S \subseteq S'$.*

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

(i) $\gamma = \langle 0, a, b, c\rangle$ with $F_n \models aRbRc$ but $F_n \not\models aRc$

(ii) $\gamma = \langle 1, a, b\rangle$ with $F_n \models aRb$ but $F_n \not\models bS_ab$

(iii) $\gamma = \langle 2, a, b, c, d\rangle$ with $F_n \models bS_acS_ad$ but not $F_n \models bS_ad$

(iv) $\gamma = \langle 3, a, b, c\rangle$ with $F_n \models aRbRc$ but $F_n \not\models bS_ac$

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

Let $<_0$ be the well-ordering on

$$C := (\{0\} \times W^3) \cup (\{1\} \times W^2) \cup (\{2\} \times W^4) \cup (\{3\} \times W^3)$$

induced by the occurrence order in some fixed enumeration of C. We define our chain to start with

$F_0 := G$. To go from F_n to F_{n+1} we proceed as follows. Let γ be the $<_0$-minimal imperfection on F_n. In case no such γ exists we set $F_{n+1} := F_n$. If such a γ does exist, F_{n+1} is as dicted by the case distinctions.

(i) $F_{n+1} := \langle W_n, R_n \cup \{\langle a, c\rangle\}, S_n, \nu_n\rangle$

(ii) $F_{n+1} := \langle W_n, R_n, S_n \cup \{\langle a, b, b\rangle\}, \nu_n\rangle$

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

(iv) $F_{n+1} := \langle W_n, R_n \cup \{\langle a, c\rangle\}, S_n \cup \{\langle a, b, c\rangle\}, \nu_n\rangle$

By an easy but elaborate induction, we can see that each F_n is a quasi-frame. The induction boils down to checking for each case (i)-(iv) that all the properties (1)-(5) from Definition 5.3.1 remain valid.

Instead of proving (4) and (5), it is better to prove something stronger, that is, that the critical and generalized cones remain unchanged.

4'. $\forall n\ [F_{n+1} \models y{\in}\mathcal{G}^A_x \Leftrightarrow F_n \models y{\in}\mathcal{G}^A_x]$

5'. $\forall n\ [F_{n+1} \models y{\in}\mathcal{C}^A_x \Leftrightarrow F_n \models y{\in}\mathcal{C}^A_x]$

Next, it is not hard to prove that $F := \cup_{i\in\omega} F_i$ is the required adequate **IL**-frame. To this extent, the following properties have to be checked. All properties have easy proofs.

$(a.)$ W is the domain of F
$(b.)$ $R_0 \subseteq \cup_{i\in\omega} R_i$
$(c.)$ $S_0 \subseteq \cup_{i\in\omega} S_i$
$(d.)$ R is conv. wellfounded on F
$(e.)$ $F \models xRyRz \to xRz$
$(f.)$ $F \models yS_xz \to xRy$ & xRz
$(g.)$ $F \models xRy \to yS_xy$
$(h.)$ $F \models xRyRz \to yS_xz$
$(i.)$ $F \models uS_xvS_xw \to uS_xw$
$(j.)$ $F \models xRy \Rightarrow \nu(x) \prec \nu(y)$
$(k.)$ $A \neq B \Rightarrow F \models \mathcal{G}^A_x \cap \mathcal{G}^B_x = \varnothing$
$(l.)$ $F \models y{\in}\mathcal{C}^A_x \Rightarrow \nu(x) \prec_A \nu(y)$

$\dashv$

We note that the **IL**-frame $F \supseteq G$ from above is actually the minimal one extending G. If in the sequel, if we refer to the closure given by the lemma, we shall mean this minimal one. Also do we note that the proof is independent on the enumeration of C and hence the order $<_0$ on C. The lemma can also be applied to non-labeled structures. If we drop all the requirements on the labels in Definition 5.3.1 and in Lemma 5.3.2 we end up with a true statement about just the old **IL**-frames.

Lemma 5.3.2 also allows a very short proof running as follows. Any intersection of adequate **IL**-frames with the same domain is again an adequate **IL**-frame. There is an adequate **IL**-frame extending G. Thus by taking intersections we find a minimal one. We have chosen to present our explicit proof as they allow us, now and in the sequel, to see which properties remain invariant.

Corollary 5.3.3. *Let $\mathcal{D}$ be a finite set of sentences, closed under subformulas and single negations. Let $G = \langle W, R, S, \nu\rangle$ be a quasi-frame on which*

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

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

Proof. We can just take the property along in the proof of Lemma 5.3.2. In Case (i) and (iv) we note that $aRbRc \to \nu(a) \subseteq_\Box \nu(c)$. Thus, if $A {\in} ((\nu(c) \setminus \nu(b)) \cap \{\Box D \mid D \in \mathcal{D}\})$, then certainly $A \notin \nu(a)$. ⊣

We have now done all the preparations for the completeness proof. Normally, also a lemma is needed to deal with deficiencies. But in the case of **IL**, Lemma 5.2.18 suffices.

5.3.2 Modal completeness

Theorem 5.3.4. **IL** *is a complete logic*

Proof. We specify the four ingredients mentioned in Remark 5.2.19.

Frame Condition For **IL**, the frame condition is empty, that is, every frame is an **IL** frame.

Invariants Given a finite set of sentences $\mathcal{D}$ closed under subformulas and single negation, the only invariant is $xRy \to \exists\, A {\in} (\nu(y) \setminus \nu(x)) \cap \{\Box D \mid D \in \mathcal{D}\}$. Clearly this invariant holds on any one-point labeled frame.

Elimination So, let $F := \langle W, R, S, \nu\rangle$ be a labeled frame satisfying the invariant. We will see how to eliminate both problems and deficiencies while conserving the invariant.

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

1. With Lemma 5.2.17 we find Δ with $\nu(a) \prec_B \Delta \ni A, \Box\neg A$. We fix some $b \notin W$. We now define

$$G' := \langle W \cup \{b\}, R \cup \{\langle a, b\rangle\}, S, \nu \cup \{\langle b, \Delta\rangle, \langle\langle a, b\rangle, B\rangle\}\rangle.$$

It is easy to see that G' is actually a quasi-frame. Note that if $G' \models xRb$, then x must be a and whence $\nu(x) \prec \nu(b)$ by definition of $\nu(b)$. Also it is not hard to see that if $b \in \mathcal{C}_x^C$ for $x{\neq}a$, that then $\nu(x) \prec_C \nu(b)$. For, $b \in \mathcal{C}_x^C$ implies $a \in \mathcal{C}_x^C$ whence $\nu(x) \prec_C \nu(a)$. By $\nu(a) \prec \nu(b)$ we get that $\nu(x) \prec_C \nu(b)$. In case $x{=}a$ we see that by definition $b \in \mathcal{C}_a^B$. But, we have chosen Δ so that $\nu(a) \prec_B \nu(b)$. We also see that G' satisfies the invariant as $\Box\neg A \in \nu(b) \setminus \nu(a)$ and $\sim A \in \mathcal{D}$.

2. With Lemma 5.3.2 we extend G' to an adequate labeled **IL**-frame G. Corollary 5.3.3 tells us that the invariant indeed holds at G. Clearly $\langle a, \neg(A \rhd B)\rangle$ is no longer a problem in G.

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

1. We first define B to be the formula such that $b \in \mathcal{C}_a^B$. If such a B does not exist, we take B to be $\bot$. Note that if such a B does exist, it must be unique by Property 4 of Definition 5.3.1. By Lemma 5.2.18 we can now find a Δ' such that $\nu(a) \prec_B \Delta' \ni D, \Box\neg D$. We fix some $c \notin W$ and define

   $$G' := \langle W, R \cup \{a, c\}, S \cup \{a, b, c\}, \nu \cup \{c, \Delta'\}\rangle.$$

   Again it is not hard to see that G' is a quasi-frame that satisfies the invariant. Clearly R is conversely well-founded. The only new S in G' is bS_ac, but we also defined aRc. We have chosen Δ' such that $\nu(a) \prec_B \nu(c)$. Clearly $\Box\neg D \notin \nu(a)$.

2. Again, G' is closed off under the frame conditions with Lemma 5.3.2. Again we note that the invariant is preserved in this process. Clearly $\langle a, b, C \rhd D\rangle$ is not a deficiency in G.

   Rounding up Clearly the union of a bounded chain of **IL**-frames is again an **IL**-frame.

$\dashv$

It is well known that **IL** has the finite model property and whence is decidable. With some more effort however we could have obtained the finite model property using the main lemma. We have chosen not to do so, as for our purposes the completeness via the construction method is sufficient.

Also, to obtain the finite model property, one has to re-use worlds during the construction method. The constraints on which worlds can be re-used is per logic differently. One aim of this section was to prove some results on a construction that is present in all other completeness proofs too. Therefore we needed some uniformity and did not want to consider re-using of worlds.

Chapter 6

Completeness and applications

In this chapter we prove the modal completeness of **IL**M via the construction method. In Section 6.2 this proof is applied to classify the modal interpretability formulas that are under any translation provably Σ_1 in any essentially reflexive theory. In Subsection 6.2.3 we make some remarks on Σ_1-sentences and self provers.

6.1 The Logic ILM

The modal completeness of **IL**M was proved by de Jongh and Veltman in [dJV90]. In this section we will reprove the modal completeness of the logic **IL**M via the main lemma. The general approach is not much different from the completeness proof for **IL**.

The novelty consists of incorporating the **IL**M frame condition, that is, whenever yS_xzRu holds, we should also have yRu. In this case, adequacy imposes $\nu(y) \prec \nu(u)$.

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

6.1.1 Preparations

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

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

Definition 6.1.1. An **IL**M-frame is a frame such that $yS_xzRu \to yRu$ holds on it. A(n adequate) labeled **IL**M-frame is an adequate labeled **IL**M-frame on which $yS_xz \rightarrow \nu(y) \subseteq_\Box \nu(z)$ holds. We call $yS_xzRu \rightarrow yRu$ the frame condition of **IL**M.

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

Lemma 6.1.2. $F \models \forall x, y, u, v\ (yS_xuRv \rightarrow yRv) \Leftrightarrow F \models$ **IL**M

We will now introduce a notion of a quasi-**IL**M-frame and a corresponding closure lemma. In order to get an **IL**M-closure lemma in analogy with Lemma 5.3.2 we need to introduce a technicality.

Definition 6.1.3. The A-critical $\mathcal{M}$-cone of x, we write $\mathcal{M}_x^A$, is defined inductively as follows.

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

Definition 6.1.4. A quasi-frame is a quasi-**IL**M-frame if[1] the following properties hold.

- $R^{\mathsf{tr}}; S^{\mathsf{tr}}$ is conversely well-founded[2]
- $yS_xz \rightarrow \nu(y) \subseteq_\Box \nu(z)$
- $y \in \mathcal{M}_x^A \Rightarrow \nu(x) \prec_A \nu(y)$

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

Lemma 6.1.5 (ILM-closure). *Let $G = \langle W, R, S, \nu\rangle$ be a quasi-***IL**M*-frame. There is an adequate* **IL**M*-frame F extending G. That is, $F = \langle W, R', S', \nu\rangle$ with $R \subseteq R'$ and $S \subseteq S'$.*

Proof. The proof is very similar to that of Lemma 5.3.2. As a matter of fact, we will use large parts of the latter proof in here. For quasi-**IL**M-frames we also define the notion of an imperfection. An *imperfection* on a quasi-**IL**M-frame F_n

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

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

is a tuple γ that is either an imperfection on the quasi-frame F_n, or it is a tuple of the form

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

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

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

We now see by an easy but elaborate induction that every F_n is a quasi-**IL**M-frame. Again, this boils down to checking that at each of (i)-(v), all the eight properties from Definition 6.1.4 are preserved.

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

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

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

To see that $8'$ holds, we reason as follows. Suppose $F_{n+1} \models y \in \mathcal{M}^A_x$. Thus $\exists z_1, \ldots, z_l$ $(0 \leq l)$ with[3] $F_{n+1} \models xR^Az_1(S_x \cup R \cup (S^{\mathsf{tr}}; R))z_2, \ldots, z_l(S_x \cup R \cup (S^{\mathsf{tr}}; R))y$. We transform the sequence $z_1, \ldots, z_l$ into a sequence $u_1, \ldots, u_m$ $(0 \leq m)$ in the following way. Every occurrence of aRc in $z_1, \ldots, z_l$ is replaced by $aRbRc$. In case that for some $n < l$ we have $z_nS^{\mathsf{tr}}aRc = z_{n+1}$, we replace z_n, z_{n+1} by z_n, b, c and thus $z_n(S^{\mathsf{tr}}; R)bRc$. We leave the rest of the sequence $z_1, \ldots, z_l$ unchanged. Clearly $F_n \models xR^Au_1(S_x \cup R \cup (S^{\mathsf{tr}}; R))u_2, \ldots, u_m(S_x \cup R \cup (S^{\mathsf{tr}}; R))y$, whence $F_n \models y \in \mathcal{M}^A_x$.

We shall include one more example for Case (v): We have eliminated an imperfection concerning the **IL**M frame-condition and $F_{n+1} :=$ $\langle W_n, R_n \cup \{\langle b, d\rangle\}, S_n, \nu_n\rangle$. To see the conversely well-foundedness of R, we reason as follows. Suppose for a contradiction that there is an infinite sequence such that $F_{n+1} \models x_1Rx_2R\ldots$. We now get an infinite sequence $y_1, y_2, \ldots$ by replacing every occurrence of bRd in $x_1, x_2, \ldots$ by bS_acRd and leaving the rest unchanged. If there are infinitely many S_a-transitions in the sequence $y_1, y_2, \ldots$ (note that there are certainly infinitely many R-transitions in $y_1, y_2, \ldots$), we get a contradiction with our assumption that $R^{\mathsf{tr}}; S^{\mathsf{tr}}$ is conversely well-founded on F_n. In the other case we get a contradiction with the conversely well-foundedness of R on F_n.

Once we have seen that indeed, every F_n is a quasi-**IL**M-frame, it is not hard to see that $F := \cup_{i\in\omega}F_i$ is the required adequate **IL**M-frame. To this extend

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

we have to check a list of properties $(a.)$-$(n.)$. The properties $(a.)$-$(l.)$ are as in the proof of Lemma 5.3.2.

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

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

Again do we note that the closure obtained in Lemma 6.1.5 is unique. Thus we can refer to the **IL**M-closure of a quasi-**IL**M-frame. All the information about the labels can be dropped in Definition 6.1.4 and Lemma 6.1.5 to obtain a lemma about regular **IL**M-frames.

Corollary 6.1.6. *Let $\mathcal{D}$ be a finite set of sentences, closed under subformulas and single negations. Let $G = \langle W, R, S, \nu\rangle$ be a quasi-**IL**M-frame on which*

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

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

Proof. The proof is as the proof of Corollary 5.3.3. We only need to remark on Case (v): If bS_acRd, we have $\nu(b) \subseteq_\Box \nu(c)$. Thus, $A \in ((\nu(d) \setminus \nu(c)) \cap \{\Box D \mid D \in \mathcal{D}\})$ implies $A \notin \nu(b)$. ⊣

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

Lemma 6.1.7. *Let Γ and Δ be maximal **IL**M-consistent sets. Consider $C \rhd D \in \Gamma \prec_B \Delta \ni C$. There exists a maximal **IL**M-consistent set Δ' with $\Gamma \prec_B \Delta' \ni D, \Box\neg D$ and $\Delta \subseteq_\Box \Delta'$.*

Proof. By compactness and by commutation of boxes and conjunctions, it is sufficient to show that for any formula $\Box E \in \Delta$ there is a Δ'' with $\Gamma \prec_B \Delta'' \ni D \wedge \Box E \wedge \Box\neg D$. As $C \rhd D$ is in the maximal **IL**M-consistent set Γ, also $C \wedge \Box E \rhd D \wedge \Box E \in \Gamma$. Clearly $C \wedge \Box E \in \Delta$, whence, by Lemma 5.2.18 we find a Δ'' with $\Gamma \prec_B \Delta'' \ni D \wedge \Box E \wedge \Box(\neg D \vee \neg\Box E)$. As **IL**M $\vdash \Box E \wedge \Box(\neg D \vee \neg \Box E) \rightarrow \Box\neg D$, we see that also $D \wedge \Box E \wedge \Box\neg D \in \Delta''$. ⊣

6.1.2 Completeness of ILM

Theorem 6.1.8. **IL**M *is a complete logic.*

Proof. **Frame Condition** In the case of **IL**M the frame condition is easy and well known, as expressed in Lemma 6.1.2.

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

$$\mathcal{I} := \left\{ \begin{array}{l} xRy \rightarrow \exists\, A \in ((\nu(y) \setminus \nu(x)) \cap \{\Box D \mid D \in \mathcal{D}\}) \\ uS_xvRw \rightarrow uRw \end{array} \right.$$

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

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

1. Using Lemma 5.2.17 we can find a MCS Δ with $\nu(a) \prec_B \Delta \ni A, \Box\neg A$. We fix some $b \notin W$ and define

   $$G' := \langle W \cup \{b\}, R \cup \{\langle a, b\rangle\}, S, \nu \cup \{\langle b, \Delta\rangle, \langle\langle a, b\rangle, B\rangle\}\rangle.$$

   We now see that G' is a quasi-**IL**M-frame. Thus, we need to check the eight points from Definitions 6.1.4 and 5.3.1. We will comment on some of these points.

   To see, for example, Point 4, $C \neq D \rightarrow \mathcal{G}_x^C \cap \mathcal{G}_x^D = \varnothing$, we reason as follows. First, we notice that $\forall\, x, y{\in}W\ [G' \models y \in \mathcal{G}_x^C$ iff $F \models y \in \mathcal{G}_x^C]$ holds for any C. Suppose $G' \models \mathcal{G}_x^C \cap \mathcal{G}_x^D \neq \varnothing$. If $G' \models b \notin \mathcal{G}_x^C \cap \mathcal{G}_x^D$, then also $F \models \mathcal{G}_x^C \cap \mathcal{G}_x^D \neq \varnothing$. As F is an **IL**M-frame, it is certainly a quasi-**IL**M-frame, whence $C = D$. If now $G' \models b \in \mathcal{G}_x^C \cap \mathcal{G}_x^D$, necessarily $G' \models a \in \mathcal{G}_x^C \cap \mathcal{G}_x^D$, whence $F \models a \in \mathcal{G}_x^C \cap \mathcal{G}_x^D$ and $C = D$.

   To see Requirement 8, $y \in \mathcal{M}_x^E \rightarrow \nu(x) \prec_E \nu(y)$, we reason as follows. Again, we first note that $\forall\, x, y{\in}W\ [G' \models y \in \mathcal{M}_x^C$ iff $F \models y \in \mathcal{M}_x^C]$ holds for any C. We only need to consider the new element, that is, $b \in \mathcal{M}_x^E$. If $x = a$ and $E = B$, we get the property by choice of $\nu(b)$.

   For $x \neq a$, we consider two cases. Either $a \in \mathcal{M}_x^E$ or $a \notin \mathcal{M}_x^E$. In the first case, we get by the fact that F is a labeled **IL**M-frame $\nu(x) \prec_E \nu(a)$. But $\nu(a) \prec \nu(b)$, whence $\nu(x) \prec_E \nu(b)$. In the second, necessarily for some $a' \in \mathcal{M}_x^E$ we have $a' S^{\mathsf{tr}} a$. But now $\nu(a') \subseteq_\Box \nu(a)$. Clearly $\nu(x) \prec_E \nu(a') \subseteq_\Box \nu(a) \prec \nu(b) \rightarrow \nu(x) \prec_E \nu(b)$.

2. With Lemma 6.1.5 we extend G' to an adequate labeled **IL**M-frame G. It is now obvious that both of the invariants hold on G. The first one holds due to Corollary 6.1.6. The other is just included in the definition of **IL**M-frames. Obviously, $\langle a, \neg(A \rhd B)\rangle$ is not a problem any more in G.

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

1. We first define B to be the formula such that $b \in \mathcal{C}_a^B$. If such a B does not exist, we take B to be $\bot$. Note that if such a B does exist, it must be unique by Property 4 of Definition 5.3.1. By Lemma 3.2.10, or just by the fact that F is an **IL**M-frame, we have that $\nu(a) \prec_B \nu(b)$.

   By Lemma 6.1.7 we can now find a Δ' such that $\nu(a) \prec_B \Delta' \ni D, \Box\neg D$ and $\nu(b) \subseteq_\Box \Delta'$. We fix some $c \notin W$ and define

   $$G' := \langle W, R \cup \{\langle a, c\rangle\}, S \cup \{\langle a, b, c\rangle\}, \nu \cup \{\langle c, \Delta'\rangle\}\rangle.$$

To see that G' is indeed a quasi-**IL**M-frame, again eight properties should be checked. But all of these are fairly routine.

For Property 4 it is good to remark that, if $c \in \mathcal{G}^A_x$, then necessarily $b \in \mathcal{G}^A_x$ or $a \in \mathcal{G}^A_x$.

To see Property 8, we reason as follows. We only need to consider $c \in \mathcal{M}^A_x$. This is possible if $x = a$ and $b \in \mathcal{M}^A_a$, or if for some $y \in \mathcal{M}^A_x$ we have $yS^{\mathsf{tr}}a$, or if $a \in \mathcal{M}^A_x$. In the first case, we get that $b \in \mathcal{M}^A_a$, and thus also $b \in \mathcal{C}^A_a$ as F is an **IL**M-frame. Thus, by Property 4, we see that $A = B$. But Δ' was chosen such that $\nu(a) \prec_B \Delta'$. In the second case we see that $\nu(x) \prec_A \nu(y) \subseteq_\Box \nu(a) \prec \nu(c)$ whence $\nu(x) \prec_A \nu(c)$. In the third case we have $\nu(x) \prec_A \nu(a) \prec \nu(c)$, whence $\nu(x) \prec_A \nu(c)$.

2. Again, G' is closed off under the frame conditions with Lemma 6.1.5. Clearly, $\langle a, b, C \rhd D\rangle$ is not a deficiency on G.

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

6.1.3 Admissible rules

With the completeness at hand, a lot of reasoning about **IL**M gets easier. This holds in particular for derived/admissible rules of **IL**M.

Lemma 6.1.9.

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

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

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

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

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

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

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

Proof. (i). $\mathbf{IL}\mathsf{M} \vdash A \Rightarrow \mathbf{IL}\mathsf{M} \vdash \Box A$ by necessitation. Now suppose $\mathbf{IL}\mathsf{M} \vdash \Box A$. We want to see $\mathbf{IL}\mathsf{M} \vdash A$. Thus, we take an arbitrary model $M = \langle W, R, S, \Vdash\rangle$ and world $m \in M$. If there is an m_0 with $M \models m_0Rm$, then $M, m_0 \Vdash \Box A$, whence $M, m \Vdash A$. If there is no such m_0, we define (we may assume $m_0 \notin W$)

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

Clearly, M' is an **IL**M-model too (the **IL**M frame conditions in the new cases follows from the transitivity of R), whence $M', m_0 \Vdash \Box A$ and thus $M', m \Vdash A$. By the construction of M' and by Lemma 3.3.4 we also get $M, m \Vdash A$.

(ii).”$\Leftarrow$” is easy. For the other direction we assume **IL**M $\not\vdash \Box A$ and **IL**M $\not\vdash \Box B$ and set out to prove **IL**M $\not\vdash \Box A \vee \Box B$. By our assumption and by completeness, we find $M_0, m_0 \Vdash \Diamond\neg A$ and $M_1, m_1 \Vdash \Diamond\neg B$. We define (for some $r \notin W_0 \cup W_1$)

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

Now, M is an **IL**M-model and $M, r \Vdash \Diamond\neg A \wedge \Diamond\neg B$ as is easily seen by Lemma 3.3.4. By soundness we get **IL**M $\not\vdash \Box A \vee \Box B$.

(iii).”$\Leftarrow$” goes as follows. $\vdash A \to B \vee \Diamond B \Rightarrow \vdash \Box(A \to B \vee \Diamond B) \Rightarrow \vdash A \rhd B \vee \Diamond B \Rightarrow \vdash A \rhd B$. For the other direction, suppose that $\not\vdash A \to B \vee \Diamond B$. Thus, we can find a model $M = \langle W, R, S, \Vdash\rangle$ and $m \in M$ with $M, m \Vdash A \wedge \neg B \wedge \Box\neg B$. We now define (with $r \notin W$)

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

It is easy to see that M' is an **IL**M-model. By Lemma 3.3.4 we see that $M', x \Vdash \varphi$ iff $M, x \Vdash \varphi$ for $x \in W$. It is also not hard to see that $M', r \Vdash \neg(A \rhd B)$. For, we have $rRm \Vdash A$. By definition, $mS_ry \to (m{=}y \vee mRy)$ whence $y \nVdash B$.

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

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

It is easy to see that M' is an **IL**M-model. Lemma 3.3.4 and general knowledge about **IL**M tells us that the generated submodel from l is a witness to the fact that **IL**M $\nvdash \Diamond A \to \Diamond B$.[4]

(v). The ”$\Leftarrow$” direction is easy. For the other direction we reason as follows.[5]

We assume that $\not\vdash A \rhd B$ and set out to prove $\not\vdash \bigwedge \Diamond A_i \to A \rhd B$. As $\not\vdash A \rhd B$, we can find $M, r \Vdash \neg(A \rhd B)$. By Lemma 3.3.4 we may assume that r is a root of M. For all i, we assumed $\not\vdash \neg A_i$, whence we can find rooted models $M_i, r_i \Vdash A_i$. As in the other cases, we define a model $\tilde{M}$ that arises by gluing r under all the r_i. Clearly we now see that $\tilde{M}, r \Vdash \bigwedge \Diamond A_i \wedge \neg(A \rhd B)$.

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

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

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

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

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

$\dashv$

Note that, as $\mathbf{IL}\mathsf{M}$ is conservative over $\mathbf{GL}$, all of the above statements not involving $\rhd$ also hold for $\mathbf{GL}$. The same holds for derived statements. For example, from Lemma 6.1.9 we can combine (*iii*) and (*iv*) to obtain $\mathbf{IL}\mathsf{M} \vdash A \to B \vee \Diamond B \Leftrightarrow \mathbf{IL}\mathsf{M} \vdash \Diamond A \to \Diamond B$. Consequently, the same holds true for $\mathbf{GL}$.

6.1.4 Decidability

It is well known that $\mathbf{IL}\mathsf{M}$ has the finite model property. It is not hard to re-use worlds in the presented construction method so that we would end up with a finite counter model. Actually, this is precisely what has been done in [Joo98]. In that paper, one of the invariants was "there are no deficiencies". We have chosen not to include this invariant in our presentation, as this omission simplifies the presentation. Moreover, for our purposes the completeness without the finite model property obtained via our construction method suffices.

Our purpose to include a new proof of the well known completeness of $\mathbf{IL}\mathsf{M}$ is twofold. On the one hand the new proof serves well to expose the construction method. On the other hand, it is an indispensable ingredient in proving Theorem 6.2.5.

6.2 Σ_1-sentences

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

This question has been solved for provability logics by Visser in [Vis95]. In [dJP96], de Jongh and Pianigiani gave an alternative solution by using the logic $\mathbf{IL}\mathsf{M}$. Our proof shall use their proof method.

We will perform our argument fully in $\mathbf{IL}\mathsf{M}$. It is very tempting to think that our result would be an immediate corollary from for example [Gor03], [Jap94] or [Ign93b]. This would be the case, if a construction method were worked out for the logics from these respective papers. In [Gor03] a sort construction method is indeed worked out. This construction method should however be a bit sharpened to suit our purposes. Moreover that sharpening would essentially reduce to the solution we present here.

6.2.1 Model construction

Throughout this subsection, unless mentioned otherwise, T will be an essentially reflexive recursively enumerable arithmetical theory. By Theorem 3.2.2 we thus know that $\mathbf{IL}(\mathrm{T}) = \mathbf{IL}\mathsf{M}$. Let us first say more precisely what we mean by an essentially Σ_1-sentence.

Definition 6.2.1. A modal sentence φ is called an essentially Σ_1-sentence, if $\forall * \ \varphi^* \in \Sigma_1(T)$. Likewise, a formula φ is essentially Δ_1 if $\forall * \ \varphi^* \in \Delta_1(T)$

If φ is an essentially Σ_1-formula we will also write $\varphi \in \Sigma_1(T)$. Analogously for $\Delta_1(T)$.

Theorem 6.2.2. *Modulo modal logical equivalence, there exist just two essentially Δ_1-formulas. That is, $\Delta_1(T) = \{\top, \bot\}$.*

Proof. Let φ be a modal formula. If $\varphi \in \Delta_1(T)$, then, by provably Σ_1-completeness, both $\forall * \ T \vdash \delta^* \to \Box\delta^*$ and $\forall * \ T \vdash \neg\delta^* \to \Box\neg\delta^*$. Consequently $\forall * \ T \vdash \Box\delta^* \vee \Box\neg\delta^*$. Thus, $\forall * \ T \vdash (\Box\delta \vee \Box\neg\delta)^*$ whence $\mathbf{IL}\mathsf{M} \vdash \Box\delta \vee \Box\neg\delta$. By Lemma 6.1.9 we see that $\mathbf{IL}\mathsf{M} \vdash \delta$ or $\mathbf{IL}\mathsf{M} \vdash \neg\delta$. ⊣

We proved Theorem 6.2.2 for the interpretability logic of essentially reflexive theories. It is not hard to see that the theorem also holds for finitely axiomatizable theories. The only ingredients that we need to prove this are [$\mathbf{IL}\mathsf{P} \vdash \Box A \vee \Box B$ iff $\mathbf{IL}\mathsf{P} \vdash \Box A$ or $\mathbf{IL}\mathsf{P} \vdash \Box B$] and [$\mathbf{IL}\mathsf{P} \vdash \Box A$ iff $\mathbf{IL}\mathsf{P} \vdash A$]. As these two admissible rules also hold for $\mathbf{GL}$, we see that Theorem 6.2.2 also holds for $\mathbf{GL}$.

Lemma 6.2.3. *If $\varphi \in \Sigma_1(T)$, then, for any p and q, we have $\mathbf{IL}\mathsf{M} \vdash p \rhd q \to p \wedge \varphi \rhd q \wedge \varphi$.*

Proof. This is a direct consequence of Pudlák's lemma, Lemma 1.3.11. ⊣

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

Lemma 6.2.4. *Let Δ_0 and Δ_1 be maximal $\mathbf{IL}\mathsf{M}$-consistent sets. There is a maximal $\mathbf{IL}\mathsf{M}$-consistent set Γ such that $\Gamma \prec \Delta_0, \Delta_1$.*

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

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

or equivalently

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

By Lemma 6.1.9 we see that then either $\vdash \neg A_i$ for some i, or $\vdash \neg B_j$ for some j. This contradicts the consistency of Δ_0 and Δ_1. ⊣

Theorem 6.2.5. *$\varphi \in \Sigma_1(T) \Leftrightarrow \mathbf{IL}\mathsf{M} \vdash \varphi \leftrightarrow \bigvee_{i\in I} \Box C_i$ for some $\{C_i\}_{i\in I}$.*

Proof. Let φ be a formula that is not equivalent to a disjunction of $\Box$-formulas. According to Lemma 6.2.7 we can find MCS's Δ_0 and Δ_1 with $\varphi \in \Delta_0 \subseteq_\Box \Delta_1 \ni \neg\varphi$. By Lemma 6.2.4 we find a $\Gamma \prec \Delta_0, \Delta_1$. We define:

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

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

$$\mathcal{I} := \left\{ \begin{array}{l} xRy \wedge x \neq m_0 \rightarrow \exists\, A{\in}((\nu(y) \setminus \nu(x)) \cap \{\Box D \mid D \in \mathcal{D}\}) \\ uS_x vRw \rightarrow uRw \end{array} \right.$$

In the proof of Theorem 6.1.8 we have seen that we can eliminate both problems and deficiencies while conserving the invariants. The main lemma now gives us an **IL**M-model M with $M, l \Vdash \varphi$, $M, r \Vdash \neg\varphi$ and $lS_{m_0}r$. We now pick two fresh variables p and q. We define p to be true only at l and q only at r. Clearly $m_0 \Vdash \neg(p \rhd q \rightarrow p \wedge \varphi \rhd q \wedge \varphi)$, whence by Lemma 6.2.3 we get $\varphi \notin \Sigma_1(T)$. ⊣

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

Theorem 6.2.6. *Let T be a finitely axiomatized theory. For all arithmetical formulae α, β there exists a formula ρ with*

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

Proof. The proof is a direct corollary of the so-called FGH-theorem. (See [Vis02] for an exposition of the FGH-theorem.) We take ρ satisfying the following fixed point equation.

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

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

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

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

6.2.2 The Σ-lemma

We can say that the proof of Theorem 6.2.5 contained three main ingredients; Firstly, the main lemma; Secondly the modal completeness theorem for **IL**M via the construction method and; Thirdly the Σ-lemma. In this subsection we will prove the Σ-lemma and remark that it is in a sense optimal.

Lemma 6.2.7. *If φ is a formula not equivalent to a disjunction of $\Box$-formulas. Then there exist maximal* **IL**X*-consistent sets Δ_0, Δ_1 such that $\varphi \in \Delta_0 \subseteq_\Box \Delta_1 \ni \neg\varphi$.*

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

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

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

$$\bigvee_{i=0}^{n-1} \sigma_i \in Y \Rightarrow \exists\, i{<}n \;\sigma_i \in Y. \tag{6.1}$$

To see this, let $Y \in \Box_{\mathrm{con}}$ and $\bigvee_{i=0}^{n-1} \sigma_i \in Y$. Then for each $i{<}n$ we have $\sigma_i \in \Box_\vee$ and for some $i{<}n$ we must have σ_i consistent with Y (otherwise $\{\neg\varphi\}+Y$ would prove $\bigwedge_{i=0}^{n-1} \neg\sigma_i$ and be inconsistent). And thus, by the maximality of Y, we must have that some σ_i is in Y. This establishes (6.1).

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

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

is consistent.

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

$$\{\varphi\} + \{\neg\sigma \mid \sigma \in Y^{\mathrm{fin}}\} \tag{6.2}$$

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

$$\vdash \varphi \to \bigvee_{\sigma\in Y^{\mathrm{fin}}} \sigma. \tag{6.3}$$

Now we will show that:

$$\{\neg\varphi\} + \{\bigvee_{\sigma\in Y^{\mathrm{fin}}} \sigma \mid Y \in \Box_{\mathrm{con}}\} \text{ is inconsistent.} \tag{6.4}$$

For, suppose (6.4) were not the case. Then for some $S \in \Box_{\mathrm{con}}$

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

In particular we have $\bigvee_{\sigma \in S^{\mathrm{fin}}} \sigma \in S$. But for all $\sigma \in S^{\mathrm{fin}}$ we have $\sigma \notin S$. Now by (6.1) we obtain a contradiction and thus we have shown (6.4).

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

$$\vdash (\bigwedge_{Y \in \Box^{\mathrm{fin}}_{\mathrm{con}}} \bigvee_{\sigma \in Y^{\mathrm{fin}}} \sigma) \to \varphi. \tag{6.5}$$

By (6.3) we also have

$$\vdash \varphi \to \bigwedge_{Y \in \Box^{\mathrm{fin}}_{\mathrm{con}}} \bigvee_{\sigma \in Y^{\mathrm{fin}}} \sigma. \tag{6.6}$$

Combining (6.5) with (6.6) we get

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

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

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

$$\{\varphi\} + \{\neg\sigma \mid \sigma \in \Box_{\vee} - Y\} \tag{6.7}$$

$$\{\neg\varphi\} + Y \tag{6.8}$$

are consistent. The lemma follows by taking Δ_0 and Δ_1 extending (6.7) and (6.8) respectively. ⊣

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

It is obvious that that $p \wedge \Box p$ is not equivalent to a disjunction of $\Box$-formulas. Clearly $p \wedge \Box p \in \Delta_0 \prec \Delta_1 \ni \neg p \vee \Diamond\neg p$ is impossible. In a sense, this reflects the fact that there exist non trivial self-provers, as was shown by Kent ([Ken73]), Guaspari ([Gua83]) and Beklemishev ([Bek93]). Thus, provable Σ_1-completeness, that is $T \vdash \sigma \to \Box\sigma$ for $\sigma \in \Sigma_1(T)$, can not substitute Lemma 6.2.3.

6.2.3 Self provers and Σ_1-sentences

A self prover is a sentence φ that implies its own provability. That is, a sentence for which $\vdash \varphi \to \Box\varphi$, or equivalently, $\vdash \varphi \leftrightarrow \varphi \wedge \Box\varphi$. Self provers have been studied intensively amongst others by Kent ([Ken73]), Guaspari ([Gua83]), de Jongh and Pianigiani ([dJP96]). It is easy to see that any $\Sigma_1(T)$-sentence is indeed a self prover. We shall call such a self prover a *trivial self prover*.

In [Gua83], Guaspari has shown that there are many non-trivial self provers around. The most prominent example is probably $p \wedge \Box p$. But actually, any formula φ will generate a self prover $\varphi \wedge \Box\varphi$, as clearly $\varphi \wedge \Box\varphi \to \Box(\varphi \wedge \Box\varphi)$.

Definition 6.2.8. A formula φ is called a trivial self prover generator, we shall write t.s.g., if $\varphi \wedge \Box\varphi$ is a trivial self prover. That is, if $\varphi \wedge \Box\varphi \in \Sigma_1(T)$.

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

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

Theorem 6.2.9. *We have that $\Sigma(\varphi \wedge \Box\varphi)$ in* **GL** *if and only if the following condition is satisfied.*

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

1. $\vdash \varphi \leftrightarrow \bigvee\!\!\!\bigvee_l (\varphi_l \wedge \Box A_l) \vee \bigvee\!\!\!\bigvee_m \Box C_m$

2. $\nvdash \Box A_l \to \varphi$ *for all l*

3. *φ_l is a non-empty conjunction of literals and $\Diamond$-formulas*

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

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

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

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

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

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

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

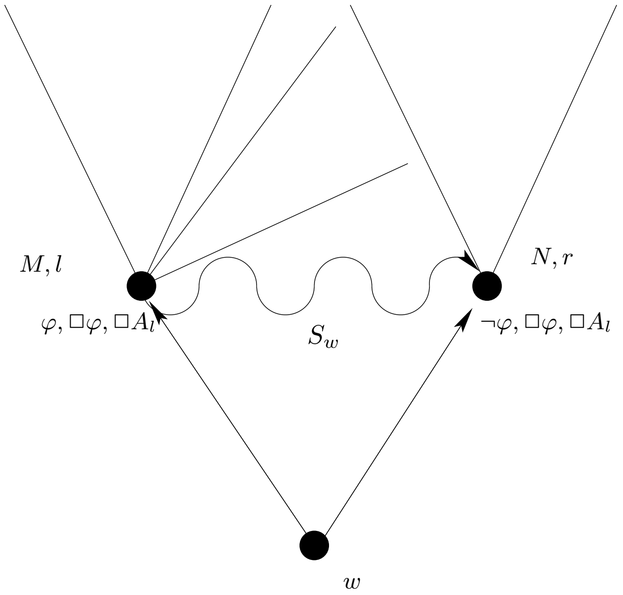
M, l
N, r
$\varphi, \Box\varphi, \Box A_l$
$\neg\varphi, \Box\varphi, \Box A_l$
S_w
w

Figure 6.1: T.s.g.'s

We can now "glue" a world w below l and r, set lS_wr and consider the smallest **IL**M-model extending this. We have depicted this construction in Figure 6.1. Let us also give a precise definition. If $M := \langle W_0, R_0, \Vdash_0\rangle$ and $N := \langle W_1, R_1, \Vdash_1\rangle$, then we define

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

We observe that, by Lemma 3.3.4 $L, r \Vdash \Box\varphi \wedge \Box A_l \wedge \neg\varphi$ and $L \models rRx \Rightarrow L, x \Vdash \varphi \wedge A_l$. Also, if $L \models lRx$, then $L, x \Vdash \varphi \wedge A_i$, whence $L, l \Vdash \Box\varphi \wedge \Box A_l$. As $M, l \Vdash \varphi_l$ and φ_l only contains literals and and diamond-formulas, we see that $L, l \Vdash \varphi_l$, whence $L, l \Vdash \varphi \wedge \Box\varphi$. As $L, r \Vdash \neg\varphi \wedge \Box\varphi$ we see that $L, w \Vdash \neg\Sigma(\varphi \wedge \Box\varphi)$.

As in the proof of Theorem 6.2.5, we can take some fresh p and q and define p to hold only at l and q to hold only at r. Now, clearly $w \nVdash p \rhd q \to p \wedge (\varphi \wedge \Box\varphi) \rhd q \wedge (\varphi \wedge \Box\varphi)$, whence, by Lemma 6.2.3 we conclude $\neg\Sigma(\varphi \wedge \Box\varphi)$. ⊣

To conclude this subsection, we remain in **GL** and shall settle the question for which φ we have that

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

We shall see how this question can be reduced to the characterization of t.s.g.'s.

Lemma 6.2.10.

For some (possibly empty) $\bigvee_i \Box C_i$ *we have* $\vdash \varphi \wedge \Box\neg\varphi \leftrightarrow \bigvee_i \Box C_i$
iff
$\vdash \Box\bot \to \varphi$ *or* $\vdash \neg\varphi$

Proof. For non-empty $\bigvee_i \Box C_i$ we have the following.

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

Here, the final step in the proof comes from Lemma 6.1.9.

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

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

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

Lemma 6.2.11.

$\Sigma(\varphi \wedge \Box\varphi) \wedge \Sigma(\varphi \wedge \Box\neg\varphi) \Rightarrow \Sigma(\varphi)$
iff
$\Sigma(\varphi \wedge \Box\varphi) \Rightarrow \Sigma(\varphi)$ *or* $\vdash \varphi \to \Diamond\top$

Proof. $\Uparrow$. Clearly, if $\Sigma(\varphi \wedge \Box\varphi) \Rightarrow \Sigma(\varphi)$, also $\Sigma(\varphi \wedge \Box\varphi) \wedge \Sigma(\varphi \wedge \Box\neg\varphi) \Rightarrow \Sigma(\varphi)$. Thus, suppose $\vdash \varphi \to \Diamond\top$, or put differently $\vdash \Box\bot \to \neg\varphi$. If now $\vdash \neg\varphi$, then clearly $\Sigma(\varphi)$, whence $\Sigma(\varphi \wedge \Box\varphi) \wedge \Sigma(\varphi \wedge \Box\neg\varphi) \Rightarrow \Sigma(\varphi)$, so, we may assume that $\nvdash \neg\varphi$. It is clear that now $\neg\Sigma(\varphi \wedge \Box\neg\varphi)$. For, suppose $\Sigma(\varphi \wedge \Box\neg\varphi)$, then by Lemma 6.2.10 we see $\vdash \Box\bot \to \varphi$, whence $\vdash \Diamond\top$. Quod non. Thus, $\vdash \Box\bot \to \neg\varphi \Rightarrow \neg\Sigma(\varphi\wedge\Box\neg\varphi)$ and thus certainly $\Sigma(\varphi\wedge\Box\varphi)\wedge\Sigma(\varphi\wedge\Box\neg\varphi) \Rightarrow \Sigma(\varphi)$.

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

As $\nvdash \Box\bot \to \neg\varphi$, also $\nvdash \neg\varphi \vee \Diamond\neg\varphi$. Thus, under the assumption that $\Sigma(\varphi\wedge\Box\varphi)$, we can find (a non-empty collection of) C_i with $\vdash \varphi\wedge\Box\varphi \leftrightarrow \bigvee_i \Box C_i$. In this case, clearly $\vdash \Box\bot \to \bigvee_i \Box C_i \to \varphi$, whence, by Lemma 6.2.10 we conclude $\Sigma(\varphi \wedge \Box\neg\varphi)$. $\dashv$

Chapter 7

More completeness results

In this chapter we prove two more modal completeness results. In Section 7.1 we prove the modal completeness of $\mathbf{IL}\mathsf{M}_0$. At some points we shall be rather sketchy in our proofs. Full proofs can be read in [GJ04]. By a minor modification of the completeness proof for $\mathbf{IL}\mathsf{M}_0$, we obtain in Section 7.2 a modal completeness proof for $\mathbf{IL}\mathsf{W}^*$.

7.1 The logic $\mathbf{IL}\mathsf{M}_0$

This section is devoted to showing the following theorem.[1]

Theorem 7.1.1. $\mathbf{IL}\mathsf{M}_0$ *is a complete logic.*

In the light of Remark 5.2.19 a proof of Theorem 7.1.1 boils down to giving the four ingredients mentioned there. Sections 7.1.3, 7.1.4, 7.1.5, 7.1.6 and 7.1.7 below contain those ingredients. Before these main sections, we have in Section 7.1.2 some preliminaries. We start in Section 7.1.1 with an overview of the difficulties we encounter during the application of the construction method to $\mathbf{IL}\mathsf{M}_0$.

7.1.1 Overview of difficulties

In the construction method we repeatedly eliminate problems and deficiencies by extensions that satisfy all the invariants. During these operations we need to keep track of two things.

1. If x has been added to solve a problem in w, say $\neg(A \rhd B) \in \nu(w)$. Then for all y such that xS_wy we have $\nu(w) \prec_B \nu(y)$.

2. If wRx then $\nu(w) \prec \nu(x)$

[1] A proof sketch of this theorem was first given in [Joo98].

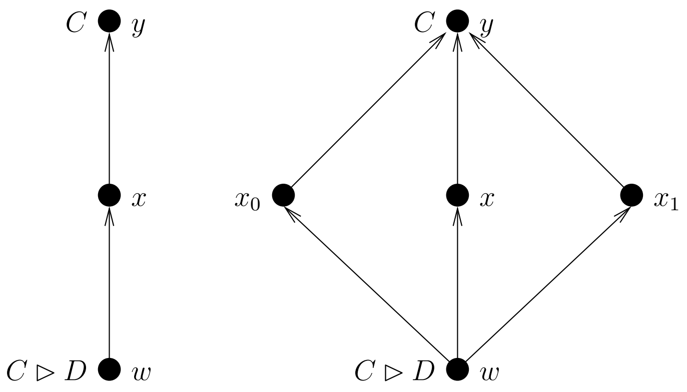

Figure 7.1: A deficiency in w w.r.t. y

Item 1. does not impose any direct difficulties. But some do emerge when we try to deal with the difficulties concerning Item 2. So let us see why it is difficult to ensure 2. Suppose we have $wRxRyS_wy'Rz$. The M_0–frame condition (Theorem 7.1.19) requires that we also have xRz. So, from 2. and the M_0–frame condition we obtain $wRxRyS_wy'Rz \to \nu(x) \prec \nu(z)$. A sufficient (and in certain sense necessary) condition is,

$$wRxRyS_wy' \to \nu(x) \subseteq_\Box \nu(y').$$

Let us illustrate some difficulties concerning this condition by some examples. Consider the left model in Figure 7.1. That is, we have a deficiency in w w.r.t. y. Namely, $C \rhd D \in \nu(w)$ and $C \in \nu(y)$. If we solve this deficiency by adding a world y', we thus require that for all x such that $wRxRy$ we have $\nu(x) \subseteq_\Box \nu(y')$. This difficulty is partially handled by Lemma 7.1.2 below. We omit a proof, but it can easily be given by replacing in the corresponding lemma for **IL**M, applications of the M-axiom by applications of the M_0-axiom.

Lemma 7.1.2. *Let Γ, Δ be* MCS*'s such that $C \rhd D \in \Gamma$, $\Gamma \prec_A \Delta$ and $\Diamond C \in \Delta$. Then there exists some Δ' with $\Gamma \prec_A \Delta'$, $\Box\neg D, D \in \Delta'$ and $\Delta \subseteq_\Box \Delta'$.*

Let us now consider the right most model in Figure 7.1. We have at least for two different worlds x, say x_0 and x_1, that $wRxRy$. Lemma 7.1.2 is applicable to $\nu(x_0)$ and $\nu(x_1)$ separately but not simultaneously. In other words we find y'_0 and y'_1 such that $\nu(x_0) \subseteq_\Box \nu(y'_0)$ and $\nu(x_1) \subseteq_\Box \nu(y'_1)$. But we actually want one single y' such that $\nu(x_0) \subseteq_\Box \nu(y')$ and $\nu(x_1) \subseteq_\Box \nu(y')$. We shall handle this difficulty by ensuring that it is enough to consider only one of the worlds in between w and y. To be precise, we shall ensure $\nu(x') \subseteq_\Box \nu(x)$ or $\nu(x) \subseteq_\Box \nu(x')$.

But now some difficulties concerning Item 1. occur. In the situations in Figure 7.1 we were asked to solve a deficiency in w w.r.t. y. As usual, if $w \prec_A y$ then we should be ably to choose a solution y' such that $w \prec_A y'$. But Lemma

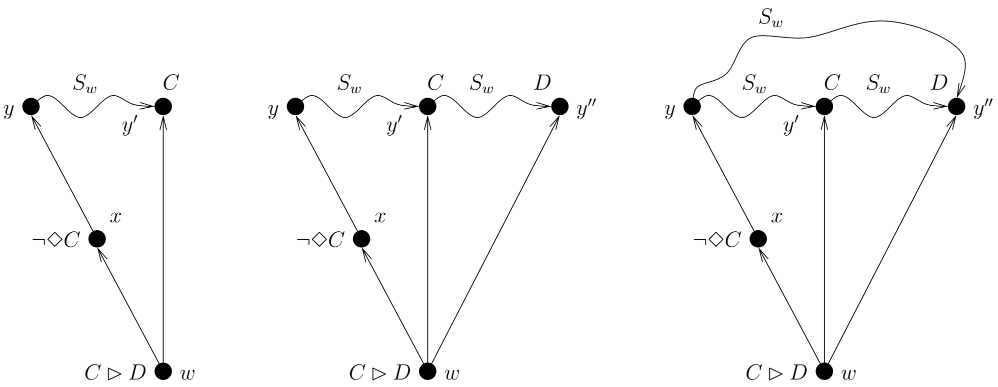

Figure 7.2: A deficiency in w w.r.t. y'

7.1.2 takes only criticallity of x w.r.t. w into account. This issue is solved by ensuring that $wRxRy \in \mathcal{C}_w^A$ implies $\nu(w) \prec_A \nu(x)$.

We are not there yet. Consider the leftmost model in Figure 7.2. That is, we have a deficiency in w w.r.t. y'. Namely, $C \rhd D \in \nu(w)$ and $C \in \nu(y')$. If we add a world y'' to solve this deficiency, as in the middle model, then by transitivity of S_w we have yS_wy'', as shown in the rightmost model. So, we require that $\nu(x) \subseteq_\Box \nu(y'')$. But we might very well have $\Diamond C \notin \nu(x)$. So the Lemma 7.1.2 is not applicable.

In Lemma 7.1.16 we formulate and prove a more complicated version of the Lemma 7.1.2 which basically says that if we have chosen $\nu(y')$ appropriately, then we can choose $\nu(y'')$ such that $\nu(x) \subseteq_\Box \nu(y'')$. And moreover, Lemma 7.1.16 ensures us that we can, indeed, choose $\nu(y')$ appropriate.

7.1.2 Preliminaries

Definition 7.1.3 ($T^{\mathbf{t}r}$, T^*, $T;T'$, T^1, $T^{\geq 2}$, $T \cup T'$). Let T and T' be binary relations on a set W. We fix the following fairly standard notations. $T^{\mathrm{t}r}$ is the transitive closure of T; T^* is the transitive reflexive closure of T; $xT;T'y \Leftrightarrow \exists t\ xTtT'y$; $xT^1y \Leftrightarrow xTy \wedge \neg\exists t\ xTtTy$; $xT^{\geq 2}y \Leftrightarrow xTy \wedge \neg(xT^1y)$ and $xT \cup T'y \Leftrightarrow xTy \vee xT'y$.

Definition 7.1.4 ($\mathcal{S}_w$). Let $F = \langle W, R, S, \nu\rangle$ be a quasi–frame. For each $w \in W$ we define the relation $\mathcal{S}_w$, of pure S_w transitions, as follows.

$$x\mathcal{S}_wy \Leftrightarrow xS_wy \wedge \neg(x = y) \wedge \neg(x(S_w \cup R)^*; R; (S_w \cup R)^*y)$$

Definition 7.1.5 (Adequate $\mathbf{IL}\mathsf{M}_0$–frame). Let $F = \langle W, R, S, \nu\rangle$ be an adequate frame. We say that F is an adequate $\mathbf{IL}\mathsf{M}_0$–frame iff the following additional properties hold.[2]

[2] One might think that 6. is superfluous. In finite frame this is indeed the case, but in the general case we need it as an requirement.

4. $wRxRyS_wy'Rz \to xRz$

5. $wRxRyS_wy' \to \nu(x) \subseteq_\Box \nu(y')$

6. $xS_wy \to x(\mathcal{S}_w \cup R)^*y$

7. $xRy \to x(R^1)^{\mathrm{tr}}y$

As usual, when we speak of $\mathbf{IL}\mathsf{M}_0$–frames we shall actually mean an adequate $\mathbf{IL}\mathsf{M}_0$–frame. Below we will construct $\mathbf{IL}\mathsf{M}_0$–frames out of frames belonging to a certain subclass of the class of quasi–frames. (Namely the quasi–$\mathbf{IL}\mathsf{M}_0$–frames, see Definition 7.1.10 below.) We would like to predict on forehand which extra R relations will be added during this construction. The following definition does just that.

Definition 7.1.6 ($K(F)$, K). Let $F = \langle W, R, S, \nu\rangle$ be a quasi–frame. We define $K = K(F)$ to be the smallest binary relation on W such that

1. $R \subseteq K$,

2. $K = K^{\mathrm{tr}}$,

3. $wKxK^1y(\mathcal{S}_w)^{\mathrm{tr}}y'K^1z \to xKz$.

Note that for $\mathbf{IL}\mathsf{M}_0$–frames we have $K = R$. The following lemma shows that K satisfies some stability conditions. The lemma will mainly be used to show that whenever we extend R within K, then K does not change.

Lemma 7.1.7. *Let $F_0 = \langle W, R_0, S, \nu\rangle$ and $F_1 = \langle W, R_1, S, \nu\rangle$ be quasi–frames. If $R_1 \subseteq K(F_0)$ and $R_0 \subseteq K(F_1)$. Then $K(F_0) = K(F_1)$.*

In a great deal of situations we have a particular interest in K^1. To determine some of its properties the following lemma comes in handy. It basically shows that we can compute K by first closing of under the M_0–condition and then take the transitive closure.

Lemma 7.1.8 (Calculation of K). *Let $F = \langle W, R, S, \nu\rangle$ be a quasi–frame. Let $K = K(F)$ and suppose K conversely well–founded. Let T be a binary relation on W such that*

1. $R \subseteq T^{tr} \subseteq K$,

2. $wT^{tr}xT^1y(\mathcal{S}_w)^{tr}y'T^1z \to xT^{tr}z$.

Then we have the following.

(a) $K = T^{tr}$

(b) $xK^1y \to xTy$

Proof. To see (a), it is enough to see that T^{tr} satisfies the three properties of the definition of K (Definition 7.1.6). Item (b) follows from (a). ⊣

Another entity that changes during the construction of an $\mathbf{IL}\mathsf{M}_0$–frame out of a quasi–frame is the critical cone In accordance with the above definition of $K(F)$, we also like to predict what eventually becomes the critical cone.

Definition 7.1.9 ($\mathcal{N}_w^C$). For any quasi–frame F we define $\mathcal{N}_w^C$ to be the smallest set such that

1. $\nu(w,x) = C \Rightarrow x \in \mathcal{N}_w^C$,
2. $x \in \mathcal{N}_w^C \wedge x(K \cup S_w)y \Rightarrow y \in \mathcal{N}_w^C$.

In accordance with the notion of a quasi–frame we introduce the notion of a quasi–$\mathbf{IL}\mathsf{M}_0$–frame. This gives sufficient conditions for a quasi–frame to be closeable, not only under the $\mathbf{IL}$–frameconditions, but under all the $\mathbf{IL}\mathsf{M}_0$–frameconditions.

Definition 7.1.10 (Quasi–$\mathbf{IL}\mathsf{M}_0$–frame). A quasi–$\mathbf{IL}\mathsf{M}_0$–frame is a quasi–frame that satisfies the following additional properties.

6. K is conversely well–founded.
7. $xKy \rightarrow \nu(x) \prec \nu(y)$
8. $x \in \mathcal{N}_w^A \rightarrow \nu(w) \prec_A \nu(x)$
9. $wKxKy(S_w \cup K)^* y' \rightarrow \nu(x) \subseteq_\Box \nu(y')$
10. $xS_w y \rightarrow x(\mathcal{S}_w \cup R)^* y$
11. $wKxK^1 y(\mathcal{S}_w)^{\mathrm{tr}} y' K^1 z \rightarrow x(K^1)^{\mathrm{tr}} z$
12. $xRy \rightarrow x(R^1)^{\mathrm{tr}} y$

Lemma 7.1.11. *If F is a quasi–$\mathbf{IL}\mathsf{M}_0$–frame, then $K = (K^1)^{tr}$.*

Proof. Using Lemma 7.1.8. ⊣

Lemma 7.1.12. *Suppose that F is a quasi–$\mathbf{IL}\mathsf{M}_0$–frame. Let $K = K(F)$. Let K', K'' and K''' the smallest binary relations on W satifying 1. and 2. of 7.1.6 and additionaly we have the following.*

3′. $wK'xK'^1 y(\mathcal{S}_w \cup K')^* y' K'^1 z \rightarrow xK'z$

3″. $wK''xK''y(\mathcal{S}_w)^{tr} y' K'' z \rightarrow xK''z$

3‴. $wK'''xK'''y(S_w \cup K''')^* y' K''' z \rightarrow xK'''z$

Then $K = K' = K'' = K'''$.

Proof. Using Lemma 7.1.11. ⊣

Before we move on, let us first sum up a few comments.

Corollary. *If $F = \langle W, R, S, \nu\rangle$ is an adequate* $\mathbf{IL}\mathsf{M}_0$*–frame. Then we have the following.*

1. $K(F) = R$
2. $F \models x \in \mathcal{N}^A_w \Leftrightarrow F \models x \in \mathcal{C}^A_w$
3. *F is a quasi–*$\mathbf{IL}\mathsf{M}_0$*–frame*

Lemma 7.1.13 ($\mathbf{IL}\mathsf{M}_0$–closure). *Any quasi–*$\mathbf{IL}\mathsf{M}_0$*–frame can be extended to an adequate* $\mathbf{IL}\mathsf{M}_0$*–frame.*

Proof. Given a quasi–$\mathbf{IL}\mathsf{M}_0$–frame F we construct a sequence

$$F = F_0 \subseteq F_1 \subseteq \cdots$$

very similar to the sequence constructed for the **IL** closure of a quasi–frame (Lemma 5.3.2). The only difference is that we add a fifth entry to the list of imperfections.

(v) $\gamma = \langle 4, w, a, b, b', c\rangle$ with $F_n \models wRaRbS_wb'Rc$ but $F_n \not\models aRc$

In this case we set, of course, $F_{n+1} := \langle W_n, R_n \cup \langle a, c\rangle, S_n, \nu_n\rangle$. First we will show by induction that each F_n is a quasi–$\mathbf{IL}\mathsf{M}_0$–frame. Then we show that the union $\hat{F} = \bigcup_{n\geq 0} F_n$, is quasi and satisfies all the $\mathbf{IL}\mathsf{M}_0$ frame conditions.

We assume that F_n is a quasi-$\mathbf{IL}\mathsf{M}_0$-frame and define $K^n := K(F_n)$, $R^n := R^{F_n}$ and $S^n := S^{F_n}$. Quasi-ness of F_{n+1} will follow from Claim 7.1.13a, and from Claim 7.1.13b we may conlude that F_{n+1} is indeed a quasi-$\mathbf{IL}\mathsf{M}_0$-frame.

Claim 7.1.13a. For all w, x, y and A we have the following.

(a) $R^{n+1} \subseteq K^n$

(b) $x(S^{n+1}_w \cup R^{n+1})^*y \Rightarrow x(S^n_w \cup K^n)^*y$

(c) $F_{n+1} \models x \in \mathcal{C}^A_w \Rightarrow F_n \models x \in \mathcal{N}^A_w$.

Proof. We distinguish cases according to which imperfection is dealt with in the step from F_n to F_{n+1}. The only interesting case is the new imperfection which is dealt with by Lemma 7.1.12, Item 3″. ⊣

Claim 7.1.13b. For all w, x and A we have the following.

1. $K^{n+1} \subseteq K^n$.
2. $x(S^{n+1}_w \cup K^{n+1})^*y \Rightarrow x(S^n_w \cup K^n)^*y$
3. $F_{n+1} \models x \in \mathcal{N}^A_w \Rightarrow F_n \models x \in \mathcal{N}^A_w$.

Proof. Item 1. follows by Claim 7.1.13a and Lemma 7.1.7. Item 2. follows from Item 1. and Claim 7.1.13a-(b). Item 3. is an immediate corollary of item 2. ⊣

Again, it is not hard to see that $\hat{F} = \bigcup_{n\geq 0} F_n$ is an adequate $\mathbf{IL}\mathsf{M}_0$-frame. ⊣

Lemma 7.1.14. *Let $F = \langle W, R, S, \nu\rangle$ be a quasi–$\mathbf{IL}\mathsf{M}_0$–frame and $K = K(F)$. Then*

$$xKy \to \exists z\,(\nu(x) \subseteq_\Box \nu(z) \wedge x(R \cup S)^* zRy).$$

Proof. We define $T := \{(x, y) \mid \exists z\,(\nu(x) \subseteq_\Box \nu(z) \wedge x(R \cup S)^* zRy)\}$. It is not hard to see that T is transitive and that $\{(x, y) \mid \exists t\,(\nu(x) \subseteq_\Box \nu(t) \wedge xT;(S \cup K)^* tTy)\} \subseteq T$. We now define $K' = K \cap T$. We have to show that $K' = K$. As $K' \subseteq K$ is trivial, we will show $K \subseteq K'$.

It is easy to see that K' satisfies properties 1., 2. and 3. of Definition 7.1.6; It follows on the two observations on T we just made. Since K is the smallest binary relation that satisfies these properties we conclude $K \subseteq K'$. ⊣

The next lemma shows that K is a rather stable relation. We show that if we extend a frame G to a frame F such that from worlds in $F - G$ we cannot reach worlds in G, then K on G does not change.

Lemma 7.1.15. *Let $F = \langle W, R, S, \nu\rangle$ be a quasi–$\mathbf{IL}\mathsf{M}_0$–frame. And let $G = \langle W^-, R^-, S^-, \nu^-\rangle$ be a subframe of F (which means $W^- \subseteq W$, $R^- \subseteq R$, $S^- \subseteq S$ and $\nu^- \subseteq \nu$). If*

(a) for each $f \in W - W^-$ and $g \in W^-$ not $f(R \cup S)g$ and

(b) $R{\restriction}_{W^-} \subseteq K(G)$.

Then $K(G) = K(F){\restriction}_{W^-}$.

Proof. Clearly $K(F){\restriction}_{W^-}$ satisfies the properties 1., 2. and 3. of the definition of $K(G)$ (Definition 7.1.6). Thus, since K_G is the smallest such relation, we get that $K(G) \subseteq K(F){\restriction}_{W^-}$.

Let $K' = K(F) - (K(F){\restriction}_{W^-} - K(G))$. Using Lemma 7.1.14 one can show that $K(F) \subseteq K'$. From this it immediately follows that $K(F){\restriction}_{W^-} \subseteq K(G)$. ⊣

We finish the basic preliminaries with a somewhat complicated variation of Lemma 5.2.18.

Lemma 7.1.16. *Let Γ and Δ be* MCS*'s. $\Gamma \prec_C \Delta$.*

$$P \rhd Q, S_1 \rhd T_1, \ldots, S_n \rhd T_n \in \Gamma \quad \textit{and} \quad \Diamond P \in \Delta.$$

There exist $k \leq n$. MCS*'s $\Delta_0, \Delta_1, \ldots, \Delta_k$ such that*

- *Each Δ_i lies C-critical above Γ,*
- *Each Δ_i lies $\subseteq_\Box$ above Δ (i.e. $\Delta \subseteq_\Box \Delta_i$),*
- *$Q \in \Delta_0$,*
- *For all $1 \leq j \leq n$, $S_j \in \Delta_h \Rightarrow$ for some $i \leq k$, $T_j \in \Delta_i$.*

Proof. First a definition. For each $I \subseteq \{1, \ldots, n\}$ put

$$\overline{S}_I :\Leftrightarrow \bigwedge\{\neg S_i \mid i \in I\}.$$

The lemma can now be formulated as follows. There exists $I \subseteq \{1, \ldots, n\}$ such that

$$\{Q, \overline{S}_I\} \cup \{\neg B, \Box\neg B \mid B \rhd C \in \Gamma\} \cup \{\Box A \mid \Box A \in \Delta\} \nvdash \bot$$

and, for all $i \notin I$,

$$\{T_i, \overline{S}_I\} \cup \{\neg B, \Box\neg B \mid B \rhd C \in \Gamma\} \cup \{\Box A \mid \Box A \in \Delta\} \nvdash \bot.$$

So let us assume, for a contradiction, that this is false. Then there exist finite sets $\mathcal{A} \subseteq \{A \mid \Box A \in \Delta\}$ and $\mathcal{B} \subseteq \{B \mid B \rhd C \in \Gamma\}$ such that, if we put

$$A :\Leftrightarrow \bigwedge \mathcal{A}, \text{ and } B :\Leftrightarrow \bigvee \mathcal{B},$$

then, for all $I \subseteq \{1, \ldots, n\}$,

$$Q, \overline{S}_I, \Box A, \neg B \wedge \Box\neg B \vdash \bot \tag{7.1}$$

or,

$$\text{for some } i \notin I, \quad T_i, \overline{S}_I, \Box A, \neg B \wedge \Box\neg B \vdash \bot. \tag{7.2}$$

We are going to define a permutation $i_1, \ldots, i_n$ of $\{1, \ldots, n\}$ such that if we put $I_k = \{i_j \mid j < k\}$ then

$$T_{i_k}, \overline{S}_{I_k}, \Box A, \neg B \wedge \Box\neg B \vdash \bot. \tag{7.3}$$

Additionally, we will verify that for each k

$$\text{(7.1) does not hold with } I_k \text{ for } I.$$

We will define i_k with induction on k. We define $I_1 = \emptyset$. And by Lemma 5.2.18, (7.1) does not hold with $I = \emptyset$. Moreover, because of this, (7.2) must be true with $I = \emptyset$. So, there exists some $i \in \{1, \ldots, n\}$ such that

$$T_i, \Box A, \neg B \wedge \Box\neg B \vdash \bot.$$

It is thus sufficient to take for i_1, for example, the least such i.

Now suppose i_k has been defined. We will first show that

$$Q, \overline{S}_{I_{k+1}}, \Box A, \neg B \wedge \Box\neg B \nvdash \bot. \tag{7.4}$$

Let us suppose that this is not so. Then

$$\vdash \Box(Q \to \Diamond\neg A \vee B \vee \Diamond B \vee S_{i_1} \vee \cdots \vee S_{i_k}). \tag{7.5}$$

So,

$$\begin{array}{lll}
\Gamma \vdash P \rhd Q & & \\
& \rhd \Diamond\neg A \vee B \vee \Diamond B \vee S_{i_1} \vee \cdots \vee S_{i_{k-1}} \vee S_{i_k} & \text{by (7.5)} \\
& \rhd \Diamond\neg A \vee B \vee \Diamond B \vee S_{i_1} \vee \cdots \vee S_{i_{k-1}} \vee T_{i_k} & \\
& \rhd \Diamond\neg A \vee B \vee \Diamond B \vee S_{i_1} \vee \cdots \vee S_{i_{k-1}} \vee (T_{i_k} \wedge \Box A \wedge \neg B \wedge \Box\neg B \wedge \overline{S}_{I_k}) & \\
& \rhd \Diamond\neg A \vee B \vee \Diamond B \vee S_{i_1} \vee \cdots \vee S_{i_{k-1}} & \text{by (7.3)} \\
& \vdots & \\
& \rhd \Diamond\neg A \vee B \vee \Diamond B \vee S_{i_1} & \\
& \rhd \Diamond\neg A \vee B \vee \Diamond B \vee T_{i_1} & \\
& \rhd \Diamond\neg A \vee B \vee \Diamond B \vee (T_{i_1} \wedge \Box A \wedge \neg B \wedge \Box\neg B) & \\
& \rhd \Diamond\neg A \vee B \vee \Diamond B. & \text{by (7.3), with } k = 1.
\end{array}$$

So by M_0,

$$\Diamond P \wedge \Box A \rhd (\Diamond\neg A \vee B \vee \Diamond B) \wedge \Box A \in \Gamma.$$

But $\Diamond P \wedge \Box A \in \Delta$. So, by Lemma 5.2.18 there exists some MCS Δ with $\Gamma \prec_C \Delta$ that contains $B \vee \Diamond B$. This is a contradiction, so we have shown (7.4).

But now, since (7.4) is indeed true, and thus (7.1) with I_{k+1} for I is false, (7.2) must hold. Thus there must exist some $i \notin I_{k+1}$ such that

$$T_i, \overline{S}_{I_{k+1}}, \Box A, \neg B \wedge \Box\neg B \vdash \bot.$$

So we can take for i_{k+1}, for example, the smallest such i.

It is clear that for $I = \{1, 2, \ldots, n\}$, (7.2) cannot be true. Thus, for $I = \{1, 2, \ldots, n\}$, (7.1) must be true. This implies

$$\vdash \Box(Q \to \Diamond\neg A \vee B \vee \Diamond B \vee S_{i_1} \vee \cdots \vee S_{i_n}).$$

Now exactly as above we can show $\Gamma \vdash P \rhd \Diamond\neg A \vee B \vee \Diamond B$. And again as above, this leads to a contradiction. ⊣

In order to formulate the invariants needed in the main lemma applied to $\mathbf{IL}\mathsf{M}_0$, we need one more definition and a corollary.

Definition 7.1.17 ($\subset_1$, $\subset$). Let $F = \langle W, R, S, \nu\rangle$ be a quasi–frame. Let $K = K(F)$. We define $\subset_1$ and $\subset$ as follows.

1. $x \subset_1 y \Leftrightarrow \exists w y' wKxK^1y'(\mathcal{S}_w)^{\mathrm{tr}}y$
2. $x \subset y \Leftrightarrow x(\subset_1 \cup K)^*y$

Corollary 7.1.18. *Let $F = \langle W, R, S, \nu\rangle$ be a quasi–frame. And let $K = K(F)$.*

1. $x \subset y \wedge yKz \to xKz$
2. *If F is a quasi–$\mathbf{IL}\mathsf{M}_0$–frame, then $x \subset y \Rightarrow \nu(x) \subseteq_{\Box} \nu(y)$.*

7.1.3 Frame condition

The following theorem is well known.

Theorem 7.1.19. *For an* **IL***-frame* $F = \langle W, R, S, \nu\rangle$ *we have*

$$\forall wxyy'z\ (wRxRyS_wy'Rz \to xRz) \Leftrightarrow F \models \mathsf{M_0}.$$

7.1.4 Invariants

Let $\mathcal{D}$ be some finite set of formulas, closed under subformulas and single negation.

During the construction we will keep track of the following main–invariants.

$\mathcal{I}_\Box$ for all y, $\{\nu(x) \mid xK^1y\}$ is linearly ordered by $\subseteq_\Box$

$\mathcal{I}_\mathrm{d}$ $wK^1x \wedge wK^{\geq 2}x'(S_w \cup K)^*x \to$ 'there does not exists a deficiency in w w.r.t. x'

$\mathcal{I}_S$ $wKxKy(S_w \cup K)^*y' \to$
'the $\subseteq_\Box$-max of $\{\nu(t) \mid wKtK^1y'\}$, if it exists, is $\subseteq_\Box$-larger than $\nu(x)$'

$\mathcal{I}_\mathcal{N}$ $wKxKy \wedge y \in \mathcal{N}^A_w \to x \in \mathcal{N}^A_w$

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

$\mathcal{I}_{\mathsf{M_0}}$ All conditions for an adequate $\mathbf{IL}\mathsf{M_0}$–frame hold

In order to ensure that the main–invariants are preserved during the construction we need to consider the following sub–invariants.[3]

$\mathcal{J}_\mathrm{u}$ $wK^{\geq 2}x(\mathcal{S}_w)^{\mathrm{tr}}y \wedge wK^{\geq 2}x'(\mathcal{S}_w)^{\mathrm{tr}}y \to x = x'$

$\mathcal{J}_{K^1}$ $wKxK^1y(\mathcal{S}_w)^{\mathrm{tr}}y'K^1z \to xK^1z$

$\mathcal{J}_\subset$ $y \subset x \wedge x \subset y \to y = x$

$\mathcal{J}_{\mathcal{N}_1}$ $x(\mathcal{S}_v)^{tr}y \wedge wKy \wedge x \in \mathcal{N}^A_w \to y \in \mathcal{N}^A_w$

$\mathcal{J}_{\mathcal{N}_2}$ $x(\mathcal{S}_w)^{tr}y \wedge y \in \mathcal{N}^A_w \to x \in \mathcal{N}^A_w$

$\mathcal{J}_{\nu_1}$ '$\nu(w,y)$ is defined' $\wedge\, vKy \to v \subset w$

$\mathcal{J}_{\nu_2}$ '$\nu(w,y)$ is defined' $\to wK^1y$

$\mathcal{J}_{\nu_4}$ If $x(\mathcal{S}_w)^{\mathrm{tr}}y$, then $\nu(w,y)$ is defined

$\mathcal{J}_{\nu_3}$ If $\nu(v,y)$ and $\nu(w,y)$ are defined then $w = v$

[3] We call them sub–invariants since they merely serve the purpose of showing that the main-invariants are, indeed, invariant.

What can we say about these invariants? $\mathcal{I}_\Box$, $\mathcal{I}_S$, $\mathcal{I}_\mathcal{N}$ and $\mathcal{I}_\mathrm{d}$ were discussed in Section 7.1.1. $\mathcal{I}_{\mathsf{M}_0}$ is there to ensure that our final frame is an $\mathbf{IL}\mathsf{M}_0$–frame. About the sub–invariants there is not much to say. They are merely technicalities that ensure that the main–invariants are invariant.

Let us first show that if we have a quasi–$\mathbf{IL}\mathsf{M}_0$–frame that satisfies all the invariants, possibly $\mathcal{I}_{\mathsf{M}_0}$ excluded, then we can assume, nevertheless, that $\mathcal{I}_{\mathsf{M}_0}$ holds as well.

Corollary 7.1.20. *Any quasi–$\mathbf{IL}\mathsf{M}_0$–frame that satisfies all of the above invariants, except possibly $\mathcal{I}_{\mathsf{M}_0}$, can be extended to an $\mathbf{IL}\mathsf{M}_0$–frame that satisfies all the invariants.*

Proof. Only $\mathcal{I}_\mathcal{D}$ and $\mathcal{I}_\mathrm{d}$ need some attention. All the other invariants are given in terms of relations that do not change during the construction of the $\mathbf{IL}\mathsf{M}_0$-closure (Lemma 7.1.13). ⊣

Lemma 7.1.21. *Let $F = \langle W, R, S, \nu \rangle$ be a quasi–$\mathbf{IL}\mathsf{M}_0$–frame. Then $F \models x \in \mathcal{N}_w^A$ iff one of the following cases applies.*

1. *$\nu(w, x) = A$*
2. *There exists $t \in \mathcal{N}_w^A$ such that tKx*
3. *There exists $t \in \mathcal{N}_w^A$ such that $t\mathcal{S}_w x$*

Corollary 7.1.22. *Let F be a quasi–$\mathbf{IL}\mathsf{M}_0$–frame that satisfies $\mathcal{J}_{\nu_4}$. Let $w, x \in F$ and let A be a formula. Then $x \in \mathcal{N}_w^A$ implies $\nu(w, x) = A$ or there exists some $t \in \mathcal{N}_w^A$ such that tKx.*

Lemma 7.1.23. *Let F be a quasi–frame which satisfies $\mathcal{J}_{\mathcal{N}_2}$, $\mathcal{J}_{\nu_1}$, $\mathcal{J}_{\nu_3}$ and $\mathcal{J}_{\nu_4}$. Then $x\mathcal{S}_v y$, $y \in \mathcal{N}_w^A \Rightarrow x \in \mathcal{N}_w^A$.*

Proof. Suppose $x\mathcal{S}_v y$ and $y \in \mathcal{N}_w^A$. Then, by Corollary 7.1.22, $\nu(w, y) = A$ or, for some $t \in \mathcal{N}_w^A$, tKy. In the first case we obtain $w = v$ by $\mathcal{J}_{\nu_3}$ and $\mathcal{J}_{\nu_4}$. And thus by $\mathcal{J}_{\mathcal{N}_2}$, $x \in \mathcal{N}_w^A$. In the second case we have, by $\mathcal{J}_{\nu_4}$ and $\mathcal{J}_{\nu_1}$ that $t \subset v$. Which implies, by Lemma 7.1.18–1., tKx. ⊣

7.1.5 Solving problems

Let $F = \langle W, R, S, \nu \rangle$ be a quasi–$\mathbf{IL}\mathsf{M}_0$–frame that satisfies all the invariants. Let $(\mathbf{a}, \neg(A \rhd B))$ be a $\mathcal{D}$-problem in F. We fix some $\mathbf{b} \notin W$. Using Lemma 5.2.17 we find a MCS $\Delta_\mathbf{b}$, such that $\nu(\mathbf{a}) \prec_B \Delta_\mathbf{b}$ and $A, \Box\neg A \in \Delta_\mathbf{b}$. We put

$$\begin{aligned}\hat{F} &= \langle \hat{W}, \hat{R}, \hat{S}, \hat{\nu} \rangle \\ &= \langle W \cup \{\mathbf{b}\}, R \cup \{\langle \mathbf{a}, \mathbf{b} \rangle\}, S, \nu \cup \{\langle \mathbf{b}, \Delta_\mathbf{b} \rangle, \langle \langle \mathbf{a}, \mathbf{b} \rangle, B \rangle\} \rangle,\end{aligned}$$

and define $\hat{K} = K(\hat{F})$. The frames F and $\hat{F}$ satisfy the conditions of Lemma 7.1.15. Thus we have

$$\forall xy{\in}F \; xKy \Leftrightarrow x\hat{K}y. \tag{7.6}$$

Since $\hat{S}{=}S$, this implies that all simple enough properties expressed in $\hat{K}$ and $\hat{S}$ using only parameters from F are true if they are true with $\hat{K}$ replaced by K.

Claim. *$\hat{F}$ is a quasi–$\mathbf{IL}\mathsf{M}_0$–frame.*

Proof. A simple check of Properties (1.–5.) of Definition 5.3.1 (quasi–frames) and Properties (6.–10.) of Definition 7.1.10 (quasi–$\mathbf{IL}\mathsf{M}_0$–frames) and the remaining ones in Definition 5.3.1 (quasi–frames). Let us comment on two of them.

$x\hat{K}y \to \hat{\nu}(x) \prec \hat{\nu}(y)$ follows from Lemma 7.1.14 and (7.6).

Let us show $\hat{F} \models x \in \mathcal{N}^C_w \Rightarrow \hat{\nu}(w) \prec_C \hat{\nu}(x)$. We have $\forall xw{\in}F\ F \models x \in \mathcal{N}^C_w \Leftrightarrow \hat{F} \models x \in \mathcal{N}^C_w$. So we only have to consider the case $\hat{F} \models \mathbf{b} \in \mathcal{N}^C_w$. If $w = \mathbf{a}$ then we are done by choice of $\hat{\nu}(\mathbf{b})$. Otherwise, by Lemma 7.1.23, we have for some $x \in F$, $F \models x \in \mathcal{N}^C_w$ and $x\hat{K}\mathbf{b}$. By the first property we proved, we get $\hat{\nu}(x) \prec \hat{\nu}(\mathbf{b})$. So, since $\hat{\nu}(w) \prec_C \hat{\nu}(x)$ we have $\hat{\nu}(w) \prec_C \hat{\nu}(\mathbf{b})$. ⊣

Before we show that $\hat{F}$ satisfies all the invariants we prove some lemmata.

Lemma 7.1.24. *If for some $x \neq \mathbf{a}$, $x\hat{K}^1\mathbf{b}$. Then there exist unique u and w (independent of x) such that $wK^{\geq 2}u(\mathcal{S}_w)^{tr}\mathbf{a}$.*

Proof. If such w and u do not exists then $T = K \cup \{\mathbf{a}, \mathbf{b}\}$ satisfies the conditions of Lemma 7.1.8. In which case $xK^1\mathbf{b}$ gives $xT\mathbf{b}$ which implies $x = \mathbf{a}$. The uniqueness of w follows from $\mathcal{J}_{\nu_3}$ and $\mathcal{J}_{\nu_4}$. The uniqueness of u follows from $\mathcal{J}_{\mathrm{u}}$ and the uniqueness of w. ⊣

In what follows we will denote these w and u, if they exist, by $\mathbf{w}$ and $\mathbf{u}$.

Lemma 7.1.25. *For all x. If $x\hat{K}^1\mathbf{b}$ then $x \subset \mathbf{a}$.*

Proof. Let $K' = K \cup \{(x, \mathbf{b}) \mid x\hat{K}\mathbf{b} \wedge x \subset \mathbf{a}\}$. It is not hard to show that K' satisfies the conditions of T in Lemma 7.1.8. ⊣

Lemma 7.1.26. *Suppose the conditions of Lemma 7.1.24 are satisfied and let $\mathbf{u}$ be the u asserted to exist. Then for all $x \neq \mathbf{a}$, if $x\hat{K}^1\mathbf{b}$, then $xK^1\mathbf{u}$.*

Proof. By Lemma 7.1.25 we have $x \subset \mathbf{a}$. Let

$$x = x_0(\subset_1 \cup K)x_1(\subset_1 \cup K)\cdots(\subset_1 \cup K)x_n = \mathbf{a}.$$

First we show $x = x_0 \subset_1 x_1 \subset_1 \cdots \subset_1 x_n = \mathbf{a}$. Suppose, for a contradiction, that for some $i < n$, x_iKx_{i+1}. Then, by Lemma 7.1.18, $xKx_{i+1}K\mathbf{b}$. So, $xK^{\geq 2}\mathbf{b}$. A contradiction. The lemma now follows by showing, with induction on i and using $F \models \mathcal{J}_{K^1}$, that for all $i \geq 0$, $x_{n-(i+1)}K^1\mathbf{u}$.

⊣

Lemma. *$\hat{F}$ satisfies all the sub-invariants.*

Proof. We only comment on $\mathcal{J}_{K^1}$ and $\mathcal{J}_{\nu_1}$. Let $K = K(\hat{F})$.

$\mathcal{J}_{\nu_1}$ follows from Lemma 7.1.25, so let us treat $\mathcal{J}_{K^1}$. Suppose $w\hat{K}x\hat{K}^1y(\hat{\mathcal{S}}_w)^{\mathrm{tr}}y'\hat{K}^1z$. We can assume that at least one of w, x, y, y', z is not in F and the only candidate for this is z. So we have $z = \mathbf{b}$. We can assume that $x \neq y'$ (otherwise we are done at once), so the conditions of Lemma 7.1.24 are fulfilled and thus $\mathbf{w}$ and $\mathbf{u}$ as stated there exist.

Suppose now, for a contradiction, that for some t, $x\hat{K}t\hat{K}^1\mathbf{b}$. Then by Lemma 7.1.26, $t = \mathbf{a}$ or $t\hat{K}^1\mathbf{u}$. Suppose we are in the case $t = \mathbf{a}$. Since $\nu(\mathbf{w}, \mathbf{a})$ is defined and $x\hat{K}\mathbf{a}$ we obtain by $\mathcal{J}_{\nu_1}$, that $x \subset \mathbf{w}$. Since $\mathbf{w}\hat{K}^{\geq 2}\mathbf{u}$ we obtain by Lemma 7.1.18 that $x\hat{K}^{\geq 2}\mathbf{u}$. In the case $t\hat{K}^1\mathbf{u}$ we have $x\hat{K}^{\geq 2}\mathbf{u}$ trivially. So in any case we have

$$x\hat{K}^{\geq 2}\mathbf{u}.$$

However, by Lemma 7.1.26 and since $y'\hat{K}^1z$ we have $y'\hat{K}^1\mathbf{u}$ or $y' = \mathbf{a}$. In the first case, since $F \models \mathcal{J}_{K^1}$, we have $x\hat{K}^1\mathbf{u}$. In the second case we obtain, by the uniqueness of $\mathbf{u}$, that $y = \mathbf{u}$ and thus $x\hat{K}^1\mathbf{u}$. So in any case we have

$$x\hat{K}^1\mathbf{u}.$$

A contradiction.

⊣

Lemma. *Possibly with the exception of* $\mathcal{I}_{\mathsf{M}_0}$, $\hat{F}$ *satisfies all the main-invariants.*

Proof. Let $K = K(\hat{F})$. We only comment on $\mathcal{I}_\Box$ and $\mathcal{I}_\mathcal{N}$.

First we treat $\mathcal{I}_\Box$. So we have to show that for all y, $\{\hat{\nu}(x) \mid x\hat{K}^1y\}$ is linearly ordered by $\subseteq_\Box$. We only need to consider the case $y = \mathbf{b}$. If $\{\mathbf{a}\} = \{x \mid x\hat{K}^1\mathbf{b}\}$ then the claim is obvious. So we can assume that the condition of Lemma 7.1.24 is fulfilled and we fix $\mathbf{u}$ as stated. The claim now follows by $F \models \mathcal{I}_\Box$ (with $y = \mathbf{u}$) and noting that, by Lemma 7.1.14, $x\hat{K}^1\mathbf{b} \Rightarrow x \subseteq_\Box \mathbf{a}$.

Now we look at $\mathcal{I}_\mathcal{N}$: $w\hat{K}x\hat{K}y \wedge \hat{F} \models y \in \mathcal{N}^A_w \to \hat{F} \models x \in \mathcal{N}^A_w$. Suppose $w\hat{K}x\hat{K}y$ and $\hat{F} \models y \in \mathcal{N}^A_w$. We only have to consider the case $y = \mathbf{b}$. Then, by Lemma 7.1.21, $\hat{\nu}(w, \mathbf{b}) = A$ or for some $t \in \mathcal{N}^A_w$ we have $t\hat{\mathcal{S}}_w\mathbf{b}$ or $t\hat{K}^1\mathbf{b}$. The first case is impossible by $\mathcal{J}_{\nu_2}$. The second is also clearly not so. Thus we have

$$t\hat{K}^1\mathbf{b}. \tag{7.7}$$

We suppose that the conditions of Lemma 7.1.24 are fulfilled (the other case is easy). If $t\hat{K}^1\mathbf{u}$ and $x\hat{K}^*\mathbf{u}$ then we are done simmilarly as the case above. So assume $t\hat{K}^1\mathbf{a}$ or $x\hat{K}^*\mathbf{a}$. Since wRt and wRx in any case we have $w\hat{K}\mathbf{a}$. Now by Lemma 7.1.23 and $\mathcal{J}_{\mathcal{N}_1}$ we have $\mathbf{u} \in \mathcal{N}^A_w \Leftrightarrow \mathbf{a} \in \mathcal{N}^A_w$. Also, by (7.7), $\mathbf{u} \in \mathcal{N}^A_w \vee \mathbf{a} \in \mathcal{N}^A_w$. So since $x\hat{K}\mathbf{u}$ or $x = \mathbf{a}$ or $x\hat{K}\mathbf{a}$ we obtain $x \in \mathcal{N}^A_w$ by $F \models \mathcal{I}_\mathcal{N}$.

⊣

To finish this subsection we note that by Lemma 7.1.13 and Corollary 7.1.20 we can extend $\hat{F}$ to an adequate $\mathbf{IL}\mathsf{M}_0$–frame that satisfies all invariants.

7.1.6 Solving deficiencies

Let $F = \langle W, R, S, \nu \rangle$ be an $\mathbf{IL}\mathsf{M}_0$–frame satisfing all the invariants. Let $(\mathbf{a}, \mathbf{b}, C \rhd D)$ be a $\mathcal{D}$-deficiency in F.

Suppose $\mathbf{a}R^{\geq 2}\mathbf{b}$ (the case $\mathbf{a}R^1\mathbf{b}$ is easy). Let x be the $\subseteq_\Box$-maximum of $\{x \mid \mathbf{a}KxK^1\mathbf{b}\}$. This maximum exists by $\mathcal{I}_\Box$. Pick some A such that $\mathbf{b} \in \mathcal{N}^A_{\mathbf{a}}$. (If such an A exists, then by adequacy of F, it is unique. If no such A exists, take $A = \bot$.) By $\mathcal{I}_\mathcal{N}$ and adequacy we have $\nu(\mathbf{a}) \prec_A \nu(x)$. So we have $C \rhd D \in \nu(\mathbf{a}) \prec_A \nu(x) \ni \Diamond C$. We apply Lemma 7.1.16 to obtain, for some set Y, disjoint from W, a set $\{\Delta_y \mid y \in Y\}$ of MCS's with all the properties as stated in that lemma. We define

$$\begin{aligned}\hat{F} = \langle & W \cup Y, R \cup \{\langle \mathbf{a}, y\rangle \mid y \in Y\}, \\ & S \cup \{\langle \mathbf{a}, \mathbf{b}, y\rangle \mid y \in Y\} \cup \{\langle \mathbf{a}, y, y'\rangle \mid y, y' \in Y, y \neq y'\}, \\ & \nu \cup \{\langle y, \Delta_y\rangle, \langle\langle \mathbf{a}, y\rangle, A\rangle \mid y \in Y\}\rangle.\end{aligned}$$

Claim. *$\hat{F}$ is a quasi–$\mathbf{IL}\mathsf{M}_0$–frame.*

Proof. An easy check of Properties (1.–5.) of Definition 5.3.1 (quasi–frames) and Properties (6.–10.) of Definition 7.1.10 (quasi–$\mathbf{IL}\mathsf{M}_0$–frames). Let us comment on two cases.

First we see that $x\hat{K}y \rightarrow \hat{\nu}(x) \prec \hat{\nu}(y)$. We can assume $y \in Y$. By Lemma 7.1.14 we obtain some z with $\hat{\nu}(x) \subseteq_\Box \hat{\nu}(z)$ and $x(\hat{R} \cup \hat{S})^* z\hat{R}y$. This z can only be $\mathbf{a}$. By choice of $\hat{\nu}(y)$ we have $\hat{\nu}(\mathbf{a}) \prec \hat{\nu}(y)$. And thus $\hat{\nu}(x) \prec \hat{\nu}(y)$.

We now see that $w\hat{K}x\hat{K}y(\hat{S}_w \cup \hat{K})^* y' \rightarrow \hat{\nu}(x) \subseteq_\Box \hat{\nu}(y')$. We can assume at least one of w, x, y, y' is in Y. The only candidates for this are y and y'. If both are in Y then $w = \mathbf{a}$ and an x as stated does not exists. So only $y' \in Y$ and thus in particular $y \neq y'$. Now there are two cases to consider.

The first case is that for some t, $w\hat{K}x\hat{K}y(\hat{S}_w \cup \hat{K})^* t\hat{K}y'$. But, $\hat{\nu}(y')$ is $\subseteq_\Box$-larger than $\hat{\nu}(t)$ by $x\hat{K}y \rightarrow \hat{\nu}(x) \prec \hat{\nu}(y)$. Also we have $wKxKy(S_w \cup K)^* t$. So, $\hat{\nu}(x) = \nu(x) \subseteq_\Box \nu(t) = \hat{\nu}(t)$.

The second case is $w\hat{K}x\hat{K}y(\hat{S}_w \cup \hat{K})^* \mathbf{b}\hat{S}_w y'$. In this case we have $w = \mathbf{a}$. y' is chosen to be $\subseteq_\Box$–larger than the $\subseteq_\Box$-maximum of $\{\nu(r) \mid \mathbf{a}KrK^1\mathbf{b}\}$. We have $wKxKy(S_w \cup K)^* \mathbf{b}$ So, by $F \models \mathcal{I}_S$, this $\subseteq_\Box$–maximum is $\subseteq_\Box$–larger than $\nu(x)$. $\dashv$

Lemma 7.1.27. *For any $x \in \hat{F}$ and $y \in Y$ we have $x\hat{K}^1 y \rightarrow x \subset \mathbf{a}$.*

Proof. We put $K' = K \cup \{(x, y) \mid y \in Y,\ x\hat{K}y,\ x \subset \mathbf{a}\}$. By showing that K' satisfies the conditions of T in Lemma 7.1.8. we obtain $x\hat{K}^1 y \rightarrow xK'y$. So if $x\hat{K}^1 y$ then $xK'y$. But if $y \in Y$ then xKy does not hold. Thus we have $x \subset \mathbf{a}$. $\dashv$

Lemma 7.1.28. *Suppose $y \in Y$ and $\mathbf{a}\hat{K}^1 z$. Then for all x, $x\hat{K}^1 y \rightarrow x\hat{K}^1 z$.*

Proof. Suppose xK^1y. By Lemma 7.1.27 we have $x \subset \mathbf{a}$. There exist $x_0, x_1, x_2, \ldots, x_n$ such that $x = x_0(\subset_1 \cup K)x_1(\subset_1 \cup K) \cdots (\subset_1 \cup K)x_n = \mathbf{a}$. First we show that $x = x_0 \subset_1 x_1 \subset_1 \cdots \subset_1 \mathbf{a}$. Suppose, for a contradiction that for some $i < n$, we have x_iKx_{i+1}. Then $xKx_{i+1}Ky$ and thus $xK^{\geq 2}y$. A contradiction. The lemma now follows by showing, with induction on i, using $\mathcal{J}_{K^1}$, that for all $i \leq n$, $x_{n-i}K^1z$. ⊣

Lemma. *$\hat{F}$ satisfies all the sub-invariants.*

Proof. The proofs are rather straightforward. We give two examples.

First we show $\mathcal{J}_{\mathrm{u}}$: $w\hat{K}^{\geq 2}x(\hat{\mathcal{S}}_w)^{\mathrm{tr}}y \wedge w\hat{K}^{\geq 2}x'(\hat{\mathcal{S}}_w)^{tr}y \to x = x'$. Suppose $w\hat{K}^{\geq 2}x(\hat{\mathcal{S}}_w)^{\mathrm{tr}}y$ and $w\hat{K}^{\geq 2}x'(\hat{\mathcal{S}}_w)^{\mathrm{tr}}y$. We can assume that $y \in Y$. (Otherwise all of w, x, x', y are in F and we are done by $F \models \mathcal{J}_{\mathrm{u}}$.) We clearly have $w \in F$. If $x \in Y$ then $w = \mathbf{a}$ and thus $w\hat{K}^1x$. So, $x \notin Y$. Next we show that both $x, x' \neq \mathbf{b}$.

Assume, for a contradiction, that at least one of them equals $\mathbf{b}$. W.l.o.g. we assume it is x. But then $wK^{\geq 2}\mathbf{b}$ and $wK^{\geq 2}x'(\mathcal{S}_w)^{tr}\mathbf{b}$. By $F \models \mathcal{J}_{\nu_4}$ we now obtain that $\nu(w, \mathbf{b})$ is defined. And thus by $F \models \mathcal{J}_{\nu_2}$, $wK^1\mathbf{b}$. A contradiction.

So, both $x, x' \neq \mathbf{b}$. But now $wK^{\geq 2}x(\mathcal{S}_w)^{tr}\mathbf{b}$ and $wK^{\geq 2}x'(\mathcal{S}_w)^{tr}\mathbf{b}$. So, by $F \models \mathcal{J}_{\mathrm{u}}$, we obtain $x = x'$.

Now let us see that $\mathcal{J}_{K^1}$ holds, that is $w\hat{K}x\hat{K}^1y(\hat{\mathcal{S}}_w)^{\mathrm{tr}}y'\hat{K}^1z \to x\hat{K}^1z$. Suppose $w\hat{K}x\hat{K}^1y(\hat{\mathcal{S}}_w)^{\mathrm{tr}}y'\hat{K}^1z$. We can assume that $z \in Y$. (Otherwise all of w, x, y, y', z are in F and we are done by $F \models \mathcal{J}_{K^1}$.) Fix some $a_1 \in F$ for which $\mathbf{a}K^1a_1$. By Lemma 7.1.28 we have $y'K^1a_1$ and thus, since $F \models \mathcal{J}_{K^1}$, xK^1a_1. By definition of $\hat{K}$ we have $x\hat{K}z$. Now, if for some t, we have $x\hat{K}t\hat{K}^1z$, then similarly as above,tK^1a_1. So, this implies $xK^{\geq 2}a_1$. A contradiction, conclusion: xK^1z. ⊣

Lemma. *Except for $\mathcal{I}_{\mathsf{M}_0}$, $\hat{F}$ satisfies all main-invariants.*

Proof. We only comment on $\mathcal{I}_{\Box}$ and $\mathcal{I}_{\mathcal{N}}$.

First we show $\mathcal{I}_{\Box}$: For all y, $\{\hat{\nu}(x) \mid x\hat{K}^1y\}$ is linearly ordered by $\subseteq_{\Box}$. Let $y \in \hat{F}$ and consider the set $\{x \mid xK^1y\}$. Since $\hat{K}\restriction_F = K$ and for all $y \in Y$ there does not exists z with $y\hat{K}^1z$ we only have to consider the case $y \in Y$. Fix some a_1 such that $\mathbf{a}K^1a_1K^*\mathbf{b}$. By Lemma 7.1.27 for any such y we have

$$\{x \mid xK^1y\} \subseteq \{x \mid xK^1a_1\}.$$

And by $F \models \mathcal{I}_{\Box}$ with a_1 for y, we know that $\{\nu(x) \mid xK^1a_1\}$ is linearly ordered by $\subseteq_{\Box}$.

Now let us see $\mathcal{I}_{\mathcal{N}}$: $w\hat{K}x\hat{K}y \wedge \hat{F} \models y \in \mathcal{N}_w^A \to \hat{F} \models x \in \mathcal{N}_w^A$. Suppose $w\hat{K}x\hat{K}y$ $\hat{F} \models y \in \mathcal{N}_w^A$. We can assume $y \in Y$. By Lemma 7.1.27, $x \subset \mathbf{a}$. So, $wKxK\mathbf{b}$. By Lemma 7.1.23, $F \models \mathbf{b} \in \mathcal{N}_w^A$ and thus $\hat{F} \models x \in \mathcal{N}_w^A$. ⊣

To finish this section we noting that by Lemma 7.1.13 and Corollary 7.1.20 we can extend $\hat{F}$ to an adequate $\mathbf{IL}\mathsf{M}_0$–frame that satisfies all invariants.

7.1.7 Rounding up

It is clear that the union of a bounded chain of $\mathbf{IL}\mathsf{M}_0$–frames is itself an $\mathbf{IL}\mathsf{M}_0$–frame.

7.2 The logic $\mathbf{IL}\mathsf{W}^*$

In this section we are going to prove the following theorem.

Theorem 7.2.1. $\mathbf{IL}\mathsf{W}^*$ *is a complete logic.*

The completeness proof of $\mathbf{IL}\mathsf{W}^*$ lifts almost completely along with the completeness proof for $\mathbf{IL}\mathsf{M}_0$. We only need some minor adaptations.

7.2.1 Preliminaries

The frame condition of W is well known.

Theorem 7.2.2. *For any* $\mathbf{IL}$*-frame* F *we have that* $F \models \mathsf{W} \Leftrightarrow \forall w\ (S_w; R)$ *is conversely well-founded.*

In [dJV99] a completeness proof for $\mathbf{IL}\mathsf{W}$ was given. We can define a new principle M_0^* that is equivalent to W^*, as follows.

$$\mathsf{M}_0^* : \quad A \rhd B \to \Diamond A \wedge \Box C \rhd B \wedge \Box C \wedge \Box\neg A$$

Lemma 7.2.3. $\mathbf{IL}\mathsf{M}_0\mathsf{W} = \mathbf{IL}\mathsf{W}^* = \mathbf{IL}\mathsf{M}_0^*$

Proof. The proof we give consists of four natural parts.

First we see $\mathbf{IL}\mathsf{W}^* \vdash \mathsf{M}_0$. We reason in $\mathbf{IL}\mathsf{W}^*$ and assume $A \rhd B$. Thus, also $A \rhd (B \vee \Diamond A)$. Applying the W^* axiom to the latter yields $(B \vee \Diamond A) \wedge \Box C \rhd (B \vee \Diamond A) \wedge \Box C \wedge \Box\neg A$. From this we may conclude

$$\begin{array}{rcl} \Diamond A \wedge \Box C & \rhd & (B \vee \Diamond A) \wedge \Box C \\ & \rhd & (B \vee \Diamond A) \wedge \Box C \wedge \Box\neg A \\ & \rhd & B \wedge \Box C \end{array}$$

Secondly, we see that $\mathbf{IL}\mathsf{W}^* \vdash \mathsf{W}$. Again, we reason in $\mathbf{IL}\mathsf{W}^*$. We assume $A \rhd B$ and take the C in the W^* axiom to be $\top$. Then we immediately see that $A \rhd B \rhd B \wedge \Box\top \rhd B \wedge \Box\top \wedge \Box\neg A \rhd B \wedge \Box\neg A$.

We now easily see that $\mathbf{IL}\mathsf{M}_0\mathsf{W} \vdash \mathsf{M}_0^*$. For, reason in $\mathbf{IL}\mathsf{M}_0\mathsf{W}$ as follows. By W^*, $A \rhd B \rhd B \wedge \Box\neg A$. Now an application of M_0 on $A \rhd B \wedge \Box\neg A$ yields $\Diamond A \wedge \Box C \rhd B \wedge \Box C \wedge \Box\neg A$.

Finally we see that $\mathbf{IL}\mathsf{M}_0^* \vdash \mathsf{W}^*$. So, we reason in $\mathbf{IL}\mathsf{M}_0^*$ and assume $A \rhd B$. Thus, we have also $\Diamond A \wedge \Box C \rhd B \wedge \Box C \wedge \Box\neg A$. We now conclude $B \wedge \Box C \rhd B \wedge \Box C \wedge \Box\neg A$ easily as follows. $B \wedge \Box C \rhd (B \wedge \Box C \wedge \Box\neg A) \vee (\Box C \wedge \Diamond A) \rhd B \wedge \Box C \wedge \Box\neg A$. ⊣

Corollary 7.2.4. *For any* $\mathbf{IL}$*-frame we have that* $F \models \mathsf{W}^*$ *iff both (for each* w*,* $(S_w; R)$ *is conversely well-founded) and* $(\forall w, x, y, y', z\ (wRxRyS_wy'Rz \to xRz))$*.*

The frame condition of W^* tells us how to correctly define the notions of adequate $\mathbf{IL}\mathsf{W}^*$-frames and quasi-$\mathbf{IL}\mathsf{W}^*$-frames.

Definition 7.2.5 ($\subsetneq_{\Box}^{\mathcal{D}}$)**.** Let $\mathcal{D}$ be a finite set of formulas. Let $\subsetneq_{\Box}^{\mathcal{D}}$ be a binary relation on MCS's defined as follows. $\Delta \subsetneq_{\Box}^{\mathcal{D}} \Delta'$ iff

1. $\Delta \subseteq_{\Box} \Delta'$,

2. For some $\Box A \in \mathcal{D}$ we have $\Box A \in \Delta' - \Delta$.

Lemma 7.2.6. *Let F be a quasi-frame and $\mathcal{D}$ be a finite set of formulas. If $wRxRyS_wy' \rightarrow \nu(x) \subsetneq^{\mathcal{D}}_{\Box} \nu(y')$ then $(R; S_w)$ is conversely well-founded.*

Proof. By the finiteness of $\mathcal{D}$. $\dashv$

Lemma 7.2.7. *Let F be a quasi-*$\mathbf{IL}\mathsf{M}_0$*-frame. If $wRxRyS_wy' \rightarrow \nu(x) \subsetneq^{\mathcal{D}}_{\Box} \nu(y')$ then $wRxRy(S_w \cup R)^*y' \rightarrow \nu(x) \subsetneq^{\mathcal{D}}_{\Box} \nu(y')$*

Proof. Suppose $wRxRy(S_w \cup R)^*y'$. $\nu(x) \subsetneq^{\mathcal{D}}_{\Box} \nu(y')$ follows with induction on the minimal number of R-steps in the path from y to y'. $\dashv$

Definition 7.2.8 (Adequate ILW***-frame).** Let $\mathcal{D}$ be a set of formulas. We say that an adequate $\mathbf{IL}\mathsf{M}_0$-frame is an adequate **IL**W*-frame (w.r.t. $\mathcal{D}$) iff the following additional property holds.

8. $wRxRy(\mathcal{S}_w)^{\mathrm{tr}}y' \rightarrow x \subsetneq^{\mathcal{D}}_{\Box} y'$

Definition 7.2.9 (Quasi-ILW***-frame).** Let $\mathcal{D}$ be a set of formulas. We say that a quasi-$\mathbf{IL}\mathsf{M}_0$-frame is a quasi-**IL**W*-frame (w.r.t. $\mathcal{D}$) iff the following additional property holds.

13. $wKxKy(\mathcal{S}_w)^{\mathrm{tr}}y' \rightarrow x \subsetneq^{\mathcal{D}}_{\Box} y'$

In what follows we might simply talk of adequate **IL**W*-frames and quasi-**IL**W* In these cases $\mathcal{D}$ is clear from context.

Lemma 7.2.10. *Any quasi-***IL**W**-frame can be extended to an adequate* **IL**W**-frame. (Both w.r.t. the same set of formulas $\mathcal{D}$.)*

Proof. Let F be a quasi-**IL**W*-frame. Then in particular F is a quasi-$\mathbf{IL}\mathsf{M}_0$-frame. So consider the proof of Lemma 7.1.13. There we constructed a sequence of quasi-$\mathbf{IL}\mathsf{M}_0$-frames $F = F_0 \subseteq F_1 \subseteq \bigcup_{i<\omega} F_i = \hat{F}$. What we have to do, is to show that if $F_0(= F)$ is a quasi-**IL**W*-frame, then each F_i is as well. Additionally we have to show that $\hat{F}$ is an adequate **IL**W*-frame.

But this is rather trivial. As noted in the proof of Lemma 7.1.13, The relation K and the relations $(\mathcal{S}_w)^{tr}$ are constant throughout the whole process. So clearly each F_i is a quasi-**IL**W*-frame.

Also the extra property of quasi-**IL**W*-frames is preserved under unions of bounded chains. So, $\hat{F}$ is an adequate **IL**W*-frame. $\dashv$

Lemma 7.2.11. *Let Γ and Δ be* MCS*'s with $\Gamma \prec_C \Delta$,*

$$P \rhd Q, S_1 \rhd T_1, \ldots, S_n \rhd T_n \in \Gamma \quad \textit{and} \quad \Diamond P \in \Delta.$$

There exist $k \leq n$. MCS*'s $\Delta_0, \Delta_1, \ldots, \Delta_k$ such that*

- *Each Δ_i lies C-critical above Γ,*

- *Each* Δ_i *lies* $\subseteq_\Box$ *above* Δ*,*
- $Q \in \Delta_0$*,*
- *For each* $i \geq 0$*,* $\Box\neg P \in \Delta_i$*,*
- *For all* $1 \leq j \leq n$*,* $S_j \in \Delta_h \Rightarrow$ *for some* $i \leq k$*,* $T_j \in \Delta_i$*.*

Proof. The proof is a straightforward adaptation of the proof of Lemma 7.1.16. In that proof, a trick was to postpone an application of $\mathsf{M_0}$ as long as possible. We do the same here but let an application of $\mathsf{M_0}$ on $P \rhd \Diamond P \vee \psi$ be preceded by an application of W to obtain $P \rhd \psi$. ⊣

7.2.2 Completeness

Again, we specify the four ingredients from Remark 5.2.19. The **Frame condition** is contained in Corollary 7.2.4.

The **Invariants** are all those of **IL**$\mathsf{M_0}$ and additionally

$$\mathcal{I}_{w^*} \quad wKxKy(\mathcal{S}_w)^{\mathrm{tr}}y' \rightarrow x \subsetneq_\Box^{\mathcal{D}} y'$$

Here, $\mathcal{D}$ is some finite set of formulas closed under subformulas and single negation.

Problems. We have to show that we can solve problems in an adequate **IL**W^*-frame in such a way that we end up with a quasi-**IL**W^*-frame. If we have such a frame then in particular it is an **IL**$\mathsf{M_0}$-frame. So, as we have seen we can extend this frame to a quasi-**IL**$\mathsf{M_0}$-frame. It is easy to see that whenever we started with an adequate **IL**W^*-frame we end up with a quasi **IL**W^*-frame. (This is basically Lemma 7.2.10.)

Deficiencies. We have to show that we can solve any deficiency in an adequate **IL**W^*-frame such that we end up with an quasi-**IL**W^*-frame. It is easily seen that the process as described in the case of **IL**$\mathsf{M_0}$ works if we use Lemma 7.2.11 instead of Lemma 7.1.16.

Rounding up. We have to show that the union of a bounded chain of quasi-**IL**W^*-frames that satisfy all the invariants is an **IL**W^*-frame. The only novelty is that we have to show that in this union for each w we have that $(R; S_w)$ is conversely well-founded. But this is ensured by $\mathcal{I}_{w^*}$ and Lemma 7.2.6.

Chapter 8

Incompleteness and full labels

In this chapter we shall prove the modal incompleteness of $\mathbf{IL}\mathsf{P_0W^*}$. Furthermore, we shall see that $\mathbf{IL}\mathsf{R}^*$ is a real extension of $\mathbf{IL}\mathsf{P_0W^*}$. In Section 8.2 we introduce the new notion of *full labels*. They seem to simplify many things concerning modal completeness proofs. We develop some general theory of these full labels. As an application we give a short and perspicuous completeness proof of $\mathbf{IL}\mathsf{W}$ in Section 8.3.

8.1 Incompleteness of $\mathbf{IL}\mathsf{P_0W^*}$

We shall now see the modal incompleteness of the logic $\mathbf{IL}\mathsf{P_0W^*}$. We do this by showing that the principle R follows semantically from $\mathbf{IL}\mathsf{P_0W^*}$ but is not provable in $\mathbf{IL}\mathsf{P_0W^*}$.

Let us first calculate the frame condition of R. It turns out to be the same frame condition as for $\mathsf{P_0}$ (see [Joo98]).

Lemma 8.1.1. $F \models \mathsf{R} \Leftrightarrow [xRyRzS_xuRv \to zS_yv]$

Proof. "$\Leftarrow$" Suppose that at some world $x \Vdash A \rhd B$. We are to show $x \Vdash \neg(A \rhd \neg C) \rhd B \wedge \Box C$. Thus, if $xRy \Vdash \neg(A \rhd \neg C)$ we need to go via an S_x to a u with $u \Vdash B \wedge \Box C$.

As $y \Vdash \neg(A \rhd \neg C)$, we can find z with $yRz \Vdash A$. Now, by $x \Vdash A \rhd B$, we can find u with $yS_xu \Vdash B$. We shall now see that $u \Vdash B \wedge \Box C$. For, if uRv, then by our assumption, zS_yv, and by $y \Vdash \neg(A \rhd \neg C)$, we must have $v \Vdash C$. Thus, $u \Vdash B \wedge \Box C$ and clearly yS_xu.

"$\Rightarrow$" We suppose that R holds. Now we consider arbitrary a, b, c, d and e with $aRbRcS_adRe$. For propositional variables p, q and r we define a valuation

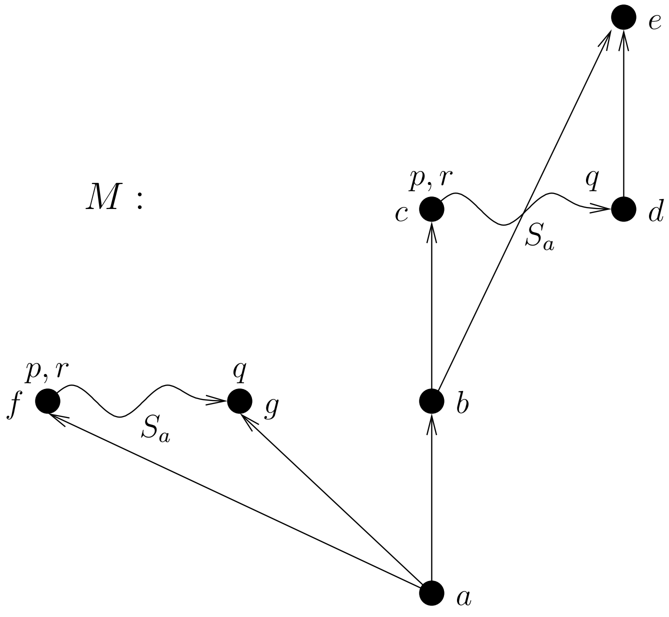

Figure 8.1: $\mathbf{IL}\mathsf{P}_0\mathsf{W}^*$ is incomplete

$\Vdash$ as follows.

$$\begin{array}{lll} x \Vdash p & :\Leftrightarrow & x = c \\ x \Vdash q & :\Leftrightarrow & x = d \\ x \Vdash r & :\Leftrightarrow & cS_bx \end{array}$$

Clearly, $a \Vdash p \rhd q$ and $b \Vdash \neg(p \rhd \neg r)$. By R we conclude $a \Vdash \neg(p \rhd \neg r) \rhd q \wedge \Box r$. Thus, $d \Vdash q \wedge \Box r$ which implies cS_be. ⊣

Theorem 8.1.2. $\mathbf{IL}\mathsf{P}_0\mathsf{W}^* \nvdash \mathsf{R}$

Proof. We consider the model M from Figure 8.1 and shall see that $M \models \mathbf{IL}\mathsf{P}_0\mathsf{W}^*$ but $M, a \nVdash \mathsf{R}$. By Lemma 3.3.10 we conclude that $\mathbf{IL}\mathsf{P}_0\mathsf{W}^* \nvdash \mathsf{R}$.

As M satisfies the frame condition for W^*, it is clear that $M \models \mathsf{W}^*$. We shall now see that $M \models A \rhd \Diamond B \rightarrow \Box(A \rhd B)$ for any formulas A and B.

A formula $\Box(A \rhd B)$ can only be false at some world with at least two successors. Thus, in M, we only need to consider the point a. So, supppose $a \Vdash A \rhd \Diamond B$. For which x with aRx can we have $x \Vdash A$?

As we have to be able to go via an S_x-transition to a world where $\Diamond B$ holds, the only candidates for x are b, c and d. But clearly, c and f make true the same modal formulas. From f it is impossible to go to a world where $\Diamond B$ holds.

Thus, if $a \Vdash A \rhd \Diamond B$, the A can only hold at b or at d. But this automatically implies that $a \Vdash \Box\Box\neg A$, whence $a \Vdash \Box(A \rhd B)$ and $M \models \mathsf{P}_0$.

It is not hard to see that $a \nVdash \mathsf{R}$. Clearly, $a \Vdash p \rhd q$ and $b \Vdash \neg(p \rhd \neg r)$. However, $d \nVdash q \wedge \Box r$ and thus $a \nVdash \neg(p \rhd \neg r) \rhd q \wedge \Box r$. ⊣

The following lemma tells us that $\mathbf{IL}\mathsf{R}$ is a proper extension of $\mathbf{IL}\mathsf{M}_0\mathsf{P}_0$.

Lemma 8.1.3. $\mathbf{IL}\mathsf{R} \vdash \mathsf{M}_0, \mathsf{P}_0$

Proof. As $\mathbf{IL} \vdash \Diamond A \wedge \Box C \rightarrow \neg(A \rhd \neg C)$ we get that $A \rhd B \rightarrow \Diamond A \wedge \Box C \rhd \neg(A \rhd \neg C)$ and M_0 follows from R.

The principle P_0 follows directly from R by taking $C = \neg B$. ⊣

It is not hard to see that R and W are fully independent over $\mathbf{IL}$. We can consider the principle R^* that can be seen, in a sense, as the union of W and R.

$$\mathsf{R}^* : \quad A \rhd B \rightarrow \neg(A \rhd \neg C) \rhd B \wedge \Box C \wedge \Box\neg A$$

Lemma 8.1.4. $\mathbf{IL}\mathsf{RW} = \mathbf{IL}\mathsf{R}^*$

Proof. $\supseteq$: $A \rhd B \rightarrow A \rhd B \wedge \Box\neg A \rightarrow \neg(A \rhd \neg C) \rhd B \wedge \Box C \wedge \Box\neg A$.

$\subseteq$: $A \rhd B \rightarrow \neg(A \rhd \neg C) \rhd B \wedge \Box C \wedge \Box\neg A \rhd B \wedge \Box C$; and if $A \rhd B$, then $A \rhd B \rhd ((B \wedge \Box\neg A) \vee \Diamond A) \rhd B \wedge \Box\neg A$, as $A \rhd B \rightarrow \neg(A \rhd \bot) \rhd B \wedge \Box\top \wedge \Box\neg A$. ⊣

8.2 Full labels

In this section we will expose a generalization of critical successor and show how it can be used to solve, in a uniform way, certain problematic aspects of modal completeness proofs.

The main idea behind our full labels is as simple as it is powerful. When we employ critical successors, we use a label to signal this. The label is there to remind us to keep our promise: once we go B-critical, we stay B-critical. That is, we never meet any B or $\Diamond B$.

Similar promises are needed to incorporate specific frame conditions, like "whenever we have some $\Box$-formulas there, we also want them there and there". The notion of criticality can, however, only deal with finitely many promises at the time. And here is precisely our generalization. We shall introduce a means to register infinitely many promises and a technique that helps us keeping our word.

First, we find it convenient to make positive promises. Rather than saying "we never meet B or $\Diamond B$", we prefer to say "we shall guarantee the presence of $\neg B$ and $\Box\neg B$".

Once we have made such a promise B, $\Box B$, say in Γ, what could force us to break our word? If $A \rhd \neg B \in \Gamma$, we could by Lemma 5.2.18 be in trouble. And thus, we should simply demand that, under our promise, we never meet A or $\Diamond A$. Or positively formulated, we will always have $\neg A$ and $\Box\neg A$. So far, we have only reformulated the definition of Δ being a $\neg B$-critical successor of Γ:

$$A \rhd \neg B \in \Gamma \Rightarrow \neg A, \Box\neg A \in \Delta. \qquad (8.1)$$

The generalization to infinitely many promises is readily made. Let Σ be a set of promises to which we would like to commit ourselves, say in Γ. We could be in trouble whenever for some collection of $S_i \in \Sigma$ and some formula A, we have $A \rhd \bigvee_i \neg S_i \in \Gamma$. In this case, in complete analogy with (8.1) we shall have to commit ourselves also to guaranteeing $\neg A$ and $\Box\neg A$.

Definition 8.2.1 (Assuring successor). Let S be a set of formulas. We define $\Gamma{\prec_S}\Delta$, and say that Δ is an S-*assuring successor* of Γ, if for any finite $S' \subseteq S$ we have[1] $A \rhd \bigvee_{S_j \in S'} \neg S_j \in \Gamma \Rightarrow \neg A, \Box\neg A \in \Delta$.

Lemma 8.2.2. *Let* Γ, Δ *and* Δ' *be* MCS*'s. For the relation* $\prec_S$ *we have the following observations.*

1. $\Gamma{\prec_\emptyset}\Delta \Leftrightarrow \Gamma \prec \Delta$

2. Δ *is a* B*-critical successor of* Γ $\Leftrightarrow \Gamma{\prec_{\{\neg B\}}}\Delta$

3. $S \subseteq T$ & $\Gamma{\prec_T}\Delta \Rightarrow \Gamma{\prec_S}\Delta$

4. $\Gamma{\prec_S}\Delta \prec \Delta' \Rightarrow \Gamma{\prec_S}\Delta'$

5. $\Gamma{\prec_S}\Delta \Rightarrow S, \Box S \subseteq \Delta, \Diamond S \subseteq \Gamma$ *and for all* A, $\Diamond A \notin S$

Theorem 8.2.3. *Let* Γ *be a MCS and* S *a set of formulas. If for any choice of* $S_i \in S$ *we have that* $\neg(B \rhd \bigvee \neg S_i) \in \Gamma$*, then*[2] *there exists a MCS* Δ *such that* $\Gamma{\prec_S}\Delta \ni B, \Box\neg B$.

Proof. Suppose for a contradiction there is no such Δ. Then there is a formula A such that for some $S_i \in S$, $(A \rhd \bigvee \neg S_i) \in \Gamma$ and $\Box\neg B, B, \Box\neg A, \neg A \vdash \bot$. Then $\vdash \Box\neg B \wedge B \rhd A \vee \Diamond A$ and we get $\vdash B \rhd A$. As $(A \rhd \bigvee \neg S_i) \in \Gamma$, also $(B \rhd \bigvee \neg S_i) \in \Gamma$. A contradiction. ⊣

Lemma 8.2.4. *Let* Γ *be a MCS such that* $\neg(B \rhd C) \in \Gamma$*. Then there is a MCS* Δ *such that* $\Gamma{\prec_{\{\neg C\}}}\Delta$ *and* $B, \Box\neg B \in \Delta$*.*

Proof. Taking $S = \{\neg C\}$ in Theorem 8.2.3. ⊣

Lemma 8.2.5. *Let* Γ *and* Δ *be MCS's such that* $A \rhd B \in \Gamma{\prec_S}\Delta \ni A$*. Then there is a MCS* Δ' *such that* $\Gamma{\prec_S}\Delta' \ni B, \Box\neg B$*.*

Proof. First we see that for any choice of S_i, $\neg(B \rhd \bigvee \neg S_i) \in \Gamma$. Suppose not. Then for some S_i, $(B \rhd \bigvee \neg S_i) \in \Gamma$ because Γ is a MCS. But then $(A \rhd \bigvee \neg S_i) \in \Gamma$ and by $\Gamma{\prec_S}\Delta$ we have $\neg A \in \Delta$. A contradiction. So $\neg(B \rhd \bigvee \neg S_i) \in \Gamma$ for any choice of S_i and we can apply Theorem 8.2.3. ⊣

Lemmata 8.2.4, 8.2.5 are the obvious generalizations of the corresponding lemmata involving criticality instead of assuringness (Lemma 5.2.17 and Lemma 5.2.18). To clarify the benefits of assuringness over criticality let us roughly

[1] Often, it is convenient to also demand that for some $\Box C \in \Delta$ we have $\Box C \notin \Gamma$.

[2] It is easy to see that we actually have iff

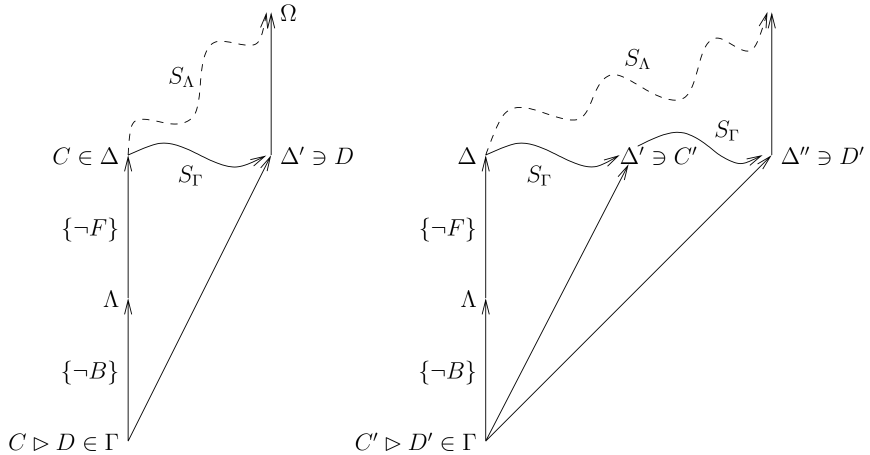

Figure 8.2: R frame condition

identify the three main points when building a counter model $\langle W, R, S, V\rangle$ for some unprovable formula (in some extension of **IL**). We take W a multi-set of MCS's and build the model in a step-by-step fashion.

1. For each $\Gamma \in W$ with $\neg(A \rhd B) \in \Gamma$ we should add some B-critical successor (equivalently $\{\neg B\}$-assuring successor) Δ to W for which $A \in \Delta$.
2. For each $\Gamma, \Delta \in W$ with $C \rhd D \in \Gamma R \Delta \ni C$ we should add a Δ' to W for which $\Gamma \prec \Delta' \ni D$. Moreover if Δ is a B-critical successor of Γ then then we should be able to choose Δ' a B-critical successor of Γ as well.
3. We should take care of the frame conditions.

When working in **IL**, Lemma 8.2.4 handles Item 1. and Lemma 8.2.5 handles Item 2. Making sure that the frame conditions are satisfied does not impose any problems (see [dJJ98]).

With extensions of **IL** the situation regarding the frame conditions becomes more complicated ([dJV90], [GJ04]). Let us clarify this by looking at **IL**R. We first recall the frame condition of R from Lemma 8.1.1.

$$wRxRyS_wy'Rz \Rightarrow yS_xz$$

This is depicted in the leftmost picture in Figure 8.2. Let us use the notation as in Item 2: Δ' was added to the model since $C \rhd D \in \Gamma R \Delta \ni C$. Since Δ lies F-critical (equivalently $\{\neg F\}$-assuring) above Λ, we should not only make sure that Δ' lies B-critical above Γ, but also that for any successor Ω of Δ' lies F-critical above Λ.

One way to guarantee this is to actually require that $\Box\neg H \in \Delta'$ whenever $H \rhd F \in \Lambda$. As one easily checks, it is quite easy to prove such a lemma in

ILR, but we have oversimplified[3] the situation. Consider the rightmost picture in Figure 8.2. That is, after having added Δ' to the model we are required to add some Δ'' with $D' \in \Delta'$ to the model since $C' \rhd D' \in \Gamma$ and $C' \in \Delta'$. By the transitivity of S_Γ we require that $\Box\neg H \in \Delta''$ whenever $H \rhd F \in \Lambda$. In this situation it is not so clear what to do.

Although for **IL**$\mathsf{M_0}$ ([GJ04]) and **IL**W ([dJV99]) there where add hoc solutions to similar problems, criticality seemed too weak a notion for a more uniform solution. As the lemmata below will show, assuringness does give us a uniform method for handling these kind of situations.

In what follows, we put for any set of formulas T,

$$\Delta_T^{\Box} = \{\Box\neg A \mid T' \subseteq T \text{ finite }, A \rhd \bigvee_{T_i \in T'} \neg T_i \in \Delta\},$$

$$\Delta_T^{\boxdot} = \{\Box\neg A, \neg A \mid T' \subseteq T \text{ finite }, A \rhd \bigvee_{T_i \in T'} \neg T_i \in \Delta\}.$$

Lemma 8.2.6. *For any logic (i.e. extension of* **IL***) we have* $\Gamma{\prec_S}\Delta \Rightarrow \Gamma{\prec_{S\cup\Gamma_S^{\boxdot}}}\Delta$.

Proof. Suppose $\Gamma{\prec_S}\Delta$ and $C{\rhd}\bigvee\neg S_i{\vee}\bigvee A_j{\vee}\Diamond A_j \in \Gamma$. Then $C{\rhd}\bigvee\neg S_i{\vee}\bigvee A_j \in \Gamma$ and thus $C \rhd \bigvee\neg S_i \vee \bigvee\neg S_k^j \in \Gamma$ which implies $\neg C, \Box\neg C \in \Delta$. ⊣

Lemma 8.2.7. *For logics containing* M *we have* $\Gamma{\prec_S}\Delta \Rightarrow \Gamma{\prec_{S\cup\Delta_\emptyset^{\Box}}}\Delta$.

Proof. Note that $\Delta_\emptyset^{\Box} = \{\Box C \mid \Box C \in \Delta\}$. We consider A such that for some $S_i \in S$ and $\Box C_j \in \Delta_\emptyset^{\Box}$, $(A \rhd \bigvee\neg S_i \vee \bigvee\neg\Box C_j) \in \Gamma$. By M, $(A \wedge \bigwedge\Box C_j \rhd \bigvee\neg S_i) \in \Gamma$, whence $\boxdot\neg(A \wedge \bigwedge\Box C_j) \in \Delta$. As $\bigwedge\Box C_j \in \Delta$, we conclude $\neg A, \Box\neg A \in \Delta$. ⊣

Lemma 8.2.8. *For logics containing* P *we have* $\Gamma{\prec_S}\Lambda{\prec_T}\Delta \Rightarrow \Gamma{\prec_{S\cup\Lambda_T^{\boxdot}}}\Delta$.

Proof. Suppose $C \rhd \bigvee\neg S_i \vee \bigvee A_j \vee \Diamond A_j \in \Gamma$, where $\Box\neg A_j, \neg A_j \in \Lambda_T^{\boxdot}$. Then $C \rhd \bigvee\neg S_i \vee \bigvee A_j \in \Gamma$ and thus by P we obtain $C \rhd \bigvee\neg S_i \vee \bigvee A_j \in \Lambda$. Since $\Gamma{\prec_S}\Lambda$ we have $\Box\bigwedge S_i \in \Lambda$ so we obtain $C \rhd \bigvee A_j \in \Lambda$. But for each A_j we have $A_j \rhd \bigvee\neg T_{jk} \in \Lambda$ and thus $C \rhd \bigvee T_{jk} \in \Lambda$. Since $\Lambda \prec_T \Delta$ we conclude $\neg C, \Box\neg C \in \Delta$. ⊣

Lemma 8.2.9. *For logics containing* $\mathsf{M_0}$ *we have* $\Gamma{\prec_S}\Delta \prec \Delta' \Rightarrow \Gamma{\prec_{S\cup\Delta_\emptyset^{\Box}}}\Delta'$.

Proof. Suppose $C \rhd \bigvee S_i \vee \bigvee\Diamond A_j \in \Gamma$, where $\Box\neg A_j \in \Delta_\emptyset^{\Box}$. By $\mathsf{M_0}$ we obtain $\Diamond C{\wedge}\bigwedge\Box\neg A_j{\rhd}\bigvee S_i \in \Gamma$. So, since $\Gamma{\prec_S}\Delta$ and $\bigwedge\Box\neg A_j \in \Delta$ we obtain $\Box\neg C \in \Delta$ and thus $\Box\neg C, \neg C \in \Delta'$. ⊣

Lemma 8.2.10. *For logics containing* R *we have* $\Gamma{\prec_S}\Delta{\prec_T}\Delta' \Rightarrow \Gamma{\prec_{S\cup\Delta_T^{\Box}}}\Delta'$.

Proof. We consider A such that for some $S_i \in S$ and some $\Box\neg A_j \in \Delta_T^{\Box}$, we have $(A \rhd \bigvee\neg S_i \vee \bigvee\Diamond A_j) \in \Gamma$. By R we obtain $(\neg(A \rhd \bigvee A_j) \rhd \bigvee\neg S_i) \in \Gamma$, thus by $\Gamma{\prec_S}\Delta$ we get $(A \rhd \bigvee A_j) \in \Delta$. As $(A_j \rhd \bigvee\neg T_{kj}) \in \Delta$, also $(A \rhd \bigvee\neg T_{kj}) \in \Delta$. By $\Delta{\prec_T}\Delta'$ we conclude $\boxdot\neg A \in \Delta'$. ⊣

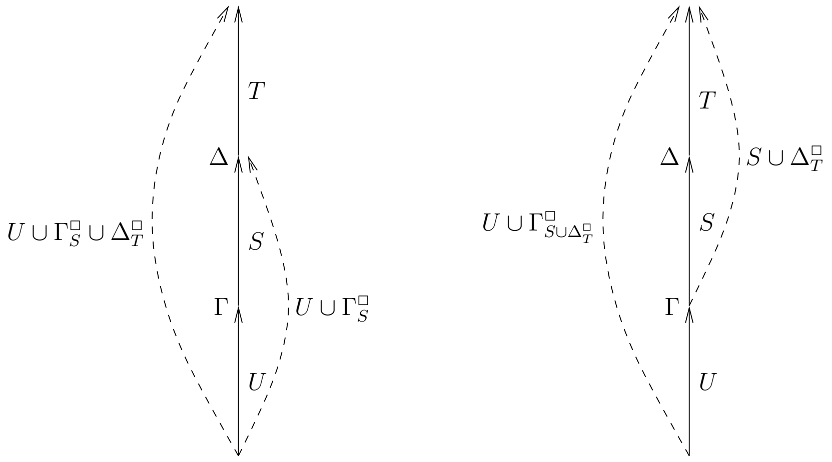
T
Δ
S
Γ
U
$U \cup \Gamma^{\Box}_{S} \cup \Delta^{\Box}_{T}$
$U \cup \Gamma^{\Box}_{S}$
T
Δ
$S \cup \Delta^{\Box}_{T}$
$U \cup \Gamma^{\Box}_{S\cup\Delta^{\Box}_{T}}$
S
Γ
U

Figure 8.3: Two ways of computing the transitive closure in **IL**R.

What lemmata 8.2.8, 8.2.9 and 8.2.10 actually tell us is how to label R-relations when we take R transitive while working in the lemma's respective logic. However, there is an easily identifiable problem here. Suppose we are working in **IL**R. Consider the two pictures in Figure 8.3. If we compute the label between the lower world and the upper world it does make a difference whether we first compute the label between the lower world and Δ (left picture) or the label between Γ and the upper world (right picture). We will show in Lemma 8.2.11 below that in the situation as given in Figure 8.3 we have

$$U \cup \Gamma^{\Box}_{S\cup\Delta^{\Box}_{T}} \subseteq U \cup \Gamma^{\Box}_{S} \cup \Delta^{\Box}_{T}.$$

And we should thus opt for the strategy as depicted in the leftmost picture when computing the transitive closure of R.

Lemma 8.2.11. *For logics containing*[4] R *we have* $\Gamma \prec_S \Delta \Rightarrow \Gamma^{\Box}_{S\cup\Delta^{\Box}_{T}} \subseteq \Delta^{\Box}_{T}$.

Proof. Consider $\Box\neg A \in \Gamma^{\Box}_{S\cup\Delta^{\Box}_{T}}$, that is, for some $S_i \in S$ and $\Box\neg B_j \in \Delta^{\Box}_{T}$, $A \rhd \bigvee \neg S_i \vee \bigvee \neg\Box\neg B_j \in \Gamma$. By R, $\neg(A \rhd \bigvee B_j) \rhd \bigvee \neg S_i \in \Gamma$, whence by $\Gamma \prec_S \Delta$, we get $A \rhd \bigvee B_j \in \Delta$. But for each B_j there is $T_{jk} \in T$ with $B_j \rhd \bigvee \neg T_{jk} \in \Delta$, whence $A \rhd \bigvee \neg T_{jk} \in \Delta$ and $\Box\neg A \in \Delta^{\Box}_{T}$. ⊣

Lemmata as Lemma 8.2.7, 8.2.9 and 8.2.10 are examples of what we call *labeling lemma*. We propose the following slogan.

> **Slogan:** Every complete logic with a first order frame condition has its own labeling lemma.

[3] We do not give a completeness proof for **IL**R here. We only indicate a few problems one will encounter and indicate the usefulness of assuringness by overcoming these.

[4] For the other logics we get similar lemmata.

Let us state two lemmata for **IL**W, a logic without a first order frame property. As predicted by our slogan, these do not fit in very nicely with the previous ones.

Lemma 8.2.12. *Suppose* $\neg(A \rhd B) \in \Gamma$. *There exists some* Δ *with* $\Gamma \prec_{\{\Box\neg A, \neg B\}} \Delta$ *and* $A \in \Delta$.

Proof. Suppose for a contradiction that there is no such Δ. Then there is a formula E with $(E \rhd \Diamond A \vee B) \in \Gamma$ such that $A, \neg E, \Box\neg E \vdash \bot$ and so $\vdash A \rhd E$. Then $(A \rhd \Diamond A \vee B) \in \Gamma$ and by the principle W we have $A \rhd B \in \Gamma$. A contradiction. $\dashv$

Lemma 8.2.13. *For logics containing* W *we have that if* $B \rhd C \in \Gamma \prec_S \Delta \ni B$ *then there exists* Δ *with* $\Gamma \prec_{S \cup \{\Box\neg B\}} \Delta \ni C, \Box\neg C$.

Proof. Suppose for a contradiction that no such Δ exists. Then for some formula A with $(A \rhd \bigvee \neg S_i \vee \Diamond B) \in \Gamma$, we get $C, \Box\neg C, \neg A, \Box\neg A \vdash \bot$, whence $\vdash C \rhd A$. Thus $B \rhd C \rhd A \rhd \bigvee \neg S_i \vee \Diamond B \in \Gamma$. By W, $B \rhd \bigvee \neg S_i \in \Gamma$ which contradicts $\Gamma \prec_S \Delta \ni B$. $\dashv$

We conclude this section with one more general lemma about assuring successors.

Lemma 8.2.14. *The following holds in extensions of* **IL**.

1. $\Gamma \prec_\Sigma \Delta$ *and* $\Sigma \vdash S$ *implies* $\Gamma \prec_{\Sigma \cup \{S, \Box S\}} \Delta$.

2. $\Gamma \prec_\Sigma \Delta$ *and* $\neg A \rhd \neg S \in \Gamma$ *(with* $S \in \Sigma$*) implies* $\Gamma \prec_{\Sigma \cup \{A\}} \Delta$.

3. $\Gamma \prec_\Sigma \Delta$ *and* $\Box A \in \Gamma$ *implies* $\Gamma \prec_{\Sigma \cup \{A\}} \Delta$.

4. $\Gamma \prec_\Sigma \Delta \prec_\Theta \Delta'$ *implies* $\Delta \prec_{\Sigma \cup \Theta} \Delta'$.

Proof. First we show 1. If $\vdash \bigwedge S_k \to S$ and $A \rhd \bigvee \neg S_i \vee \neg S \vee \neg\Box S \in \Gamma$, then $A \rhd \bigvee \neg S_i \vee \neg S \in \Gamma$ and $S \rhd \bigvee \neg S_k \in \Gamma$. So, $A \rhd \bigvee \neg S_i \vee \bigvee S_k \in \Gamma$ and thus $\neg A, \Box\neg A \in \Delta$.

We now show 2. If $C \rhd \bigvee \neg S_i \vee \neg A \in \Gamma$ and $\neg A \rhd \neg S \in \Gamma$, then $C \rhd \bigvee \neg S_i \vee \neg S \in \Gamma$ and thus $\neg C, \Box\neg C \in \Delta$. Item 3. follows at once from Item 2. by noting that $\Box A$ is equivalent to $\neg A \rhd \bot$.

Finally we show Item 4. If $\Gamma \prec_\Sigma \Delta$ then $\Gamma \prec_{\Sigma \cup \{\Box S \mid S \in \Sigma\}} \Delta$ by Item 1. So, $\{\Box S \mid S \in \Sigma\} \subseteq \Delta$, and thus by 3. we obtain hat if $\Delta \prec_\Theta \Delta'$, then $\Delta \prec_{\Theta \cup \{\Box S \mid S \in \Sigma\}} \Delta'$. $\dashv$

We note that in particular, by Item 1. we can assume that our labels are closed under logical consequence. So, if $\Gamma \prec_\Sigma \Delta$ and $A \rhd \bigvee \neg S_i \in \Gamma$ then we can just as well write $A \rhd \neg S \in \Gamma$ (where $S = \bigwedge S_i$).

8.3 The logic ILW

As a demonstration of the use of assuringness we will give in this section a relatively simple proof of the known fact ([dJV99]) that **IL**W is a complete logic.

Proving the decidability of an interpretability logic is in all known cases done by showing that the logic has the finite model property. The finite model property is easier to achieve if the building blocks of the model are finite sets instead of infinite maximal consistent sets.

A turn that is usually made to obtain finite building blocks, is to work with truncated parts of maximal consistent sets. This part should be large enough to allow for the basic reasoning. This gives rise to the notion of so-called adequate sets where different logics yield different notions of adequacy. In order to obtain the finite model property along with modal completeness of **IL**W, in this section we will use the following notion of adequacy.

Definition 8.3.1 (Adequate set). We say that a set of formulas Φ is *adequate* iff

1. $\bot \rhd \bot \in \Phi$,
2. Φ is closed under single negation and subformulas,
3. If A is an antecedent or consequent of some $\rhd$ formula in Φ and so is B then $A \rhd B \in \Phi$.

It is clear that any formula is contained in some finite and minimal adequate set. For a formula F we will denote this set by $\Phi(F)$. Since our maximal consistent sets are more restricted we should also modify the notion of an assuring successor a bit

Definition 8.3.2 ($\langle S, \Phi\rangle$-assuring successor). Let Φ be a finite adequate set, $S \subseteq \Phi$ and $\Gamma, \Delta \subseteq \Phi$ be maximal consistent sets. We say that Δ is an $\langle S, \Phi\rangle$-assuring successor of Γ ($\Gamma \prec_S^\Phi \Delta$) iff for each $\Box\neg A \in \Phi$ we have

$$\Gamma \vdash A \rhd \bigvee_{S_i \in S} \neg S_i \Rightarrow \neg A, \Box\neg A \in \Delta.$$

Moreover for some $\Box C \in \Delta$ we have $\Box C \notin \Gamma$.

Note that by the requirement $\Box\neg A \in \Phi$ the usual reading of $\prec$ in extensions of **GL** coincides with $\prec_\emptyset^\Phi$. So, we will write $\prec$ for $\prec_\emptyset^\Phi$. The following two lemmata are proved exactly as their infinite counterparts.

Lemma 8.3.3. *Let $\Gamma \subseteq \Phi$ be maximal consistent. If $\neg(A \rhd B) \in \Gamma$ then there exists some maximal consistent set $\Delta \subseteq \Phi$ such that $A \in \Delta$ and $\Gamma \prec_{\{\neg B, \Box\neg A\}}^\Phi \Delta$.*

Lemma 8.3.4. *Let $\Gamma, \Delta \subseteq \Phi$ be maximal consistent and $S \subseteq \Phi$. If $A \rhd B \in \Gamma$, $\Gamma \prec_S^\Phi \Delta$ and $A \in \Delta$, then there exists some maximal consistent $\Delta' \subseteq \Phi$ with $B \in \Delta'$ and $\Gamma \prec_{S\cup\{\Box\neg A\}}^\Phi \Delta'$.*

In what follows we let Φ be some fixed finite adequate set and reason with $\mathbf{IL}\mathsf{W}$ (e.g. $\vdash$ is $\mathbf{IL}\mathsf{W}$-provable, and consistent is $\mathbf{IL}\mathsf{W}$-consistent). The rest of this section is devoted to the proof of the following theorem.

Theorem 8.3.5 (Completeness of ILW **[dJV99]).** **IL**W *is complete with respect to finite Veltman frames $\langle W, R, S\rangle$ in which, for each $w \in W$, $(S_w; R)$ is conversely well-founded.*

Suppose $\nvdash G$. Let $\Phi = \Phi(\neg G)$ and let $\Gamma \subseteq \Phi$ be a maximal consistent set that contains $\neg G$. We will construct a Veltman model $\langle W, R, S, \Vdash\rangle$ in which for each $w \in W$ we have that $(S_w; R)$ is conversely well-founded. Each $w \in W$ will be a tuple, the second component, denoted by $(w)_1$, of which will be a maximal consistent subset of Φ. For some $w \in W$ we will have $(w)_1 = \Gamma$ and we will finish the proof by proving a truth lemma: $w \Vdash A$ iff $A \in (w)_1$.

Let the *height* of a maximal consistent $\Delta \subseteq \Phi$ be defined as the number of $\Box$-formulas in Δ minus the number of $\Box$-formulas in Γ. For sequences σ_0 and σ_1 we write $\sigma_0 \subseteq \sigma_1$ iff σ_0 is an initial, but not necessarily proper subsequence of σ_1. For two sequences σ_0 and σ_1, $\sigma_0 * \sigma_1$ denotes the concatenation of the two sequences. If S is a set of formulas then $\langle S\rangle$ is the sequence of length one whose only element is S. Let us now define $\langle W, R, S, \Vdash\rangle$.

1. W is the set of tuples $\langle \sigma, \Delta\rangle$ where $\Delta \subseteq \Phi$ is maximal consistent such that either $\Gamma = \Delta$ or $\Gamma \prec \Delta$ and σ is a finite sequence of subsets of Φ, the length of which does not exceed the height of Δ. For $w = \langle \sigma, \Delta\rangle$, we write $(w)_0$ for σ and $(w)_1$ for Δ.

2. wRv iff for some S we have $(v)_0 \supseteq (w)_0 * \langle S\rangle$ and $(w)_1 \prec^{\Phi}_{S} (v)_1$.

3. xS_wy iff wRx, y and, xRy or $x = y$ or both 3.a and 3.b hold.

   (a) If $(x)_0 = (w)_0 * \langle S\rangle * \tau_x$, $(y)_0 = (w)_0 * \langle T\rangle * \tau_y$ then $S \subseteq T$

   (b) For some $C \rhd D \in (w)_1$ we have $\Box\neg C \in T$ and, $C \in (x)_1$ or $\Diamond C \in (x)_1$

4. $w \Vdash p$ iff $p \in (w)_1$.

Lemma 8.3.6. *R is transitive and conversely well-founded.*

Proof. Transitivity follows from the fact that $(x)_1 \prec^{\Phi}_{S} (y)_1 \prec (z)_1$ implies $(x)_1 \prec^{\Phi}_{S} (z)_1$. Conversely well-foundedness now follows from the fact that our model is finite and R is irreflexive. ⊣

Lemma 8.3.7. *$wRxRy$ implies xS_wy, wRx implies xS_wx, S_w is transitive*

Proof. The first two assertions hold by definition. So suppose xS_wyS_wz. Let us fix $(x)_0 \supseteq (w)_0 * \langle S\rangle$, $(y)_0 \supseteq (w)_0 * \langle T\rangle$ and $(z)_0 \supseteq (w)_0 * \langle U\rangle$. We distinguish two cases.

Case 1: xRy or $x = y$. If $x = y$ then we are done so we assume xRy. If yRz or $y = z$ then we are also easily done. So, we assume that for some $C \rhd D \in (w)_1$

we have $\Box\neg C \in U$ and, $C \in (y)_1$ or $\Diamond C \in (y)_1$. Since $(x)_1 \prec (y)_1$, we have that $\Diamond C \in (x)_1$ and thus we conclude $xS_w z$.

Case 2: $\neg(xRy)$ and $x \neq y$. In this case there exists some $C \rhd D \in (w)_1$ with $\Box\neg C \in T$ and $C \in (x)_1$ or $\Diamond C \in (x)_1$. Whatever the reason for $yS_w z$ is, we always have $T \subseteq U$ and thus $\Box\neg C \in U$. So, we conclude $xS_w z$. ⊣

Lemma 8.3.8. *$(S_w; R)$ is conversely well-founded.*

Proof. Suppose we have an infinite sequence

$$x_0 S_w y_0 R x_1 S_w y_1 R \cdots .$$

For each $i \geq 0$, fix X_i and Y_i such that $(x_i)_0 \supseteq (w)_0 * \langle X_i \rangle$ and $(y_i)_0 \supseteq (w)_0 * \langle Y_i \rangle$. We may assume that, for each i, $x_i \neq y_i$ and $\neg(x_i R y_i)$. Thus, let $C_i \rhd D_i$ be the formula as given by condition 3.b. We thus have $C_i \rhd D_i \in (w)_1$, $\Box\neg C_i \in Y_i$ and, $C_i \in (x_i)_1$ or $\Diamond C_i \in (x_i)_1$. For any $j > i$, this implies $\Box\neg C_i \in X_j$ which gives $\Box\neg C_i \in (x_j)_1$ and thus $\neg C_i, \Box\neg C_i \in (y_j)_1$. The latter gives $C_i \neq C_j$, which is a contradiction since Φ is finite. ⊣

Lemma 8.3.9 (Truth lemma). *For all $F \in \Phi$ and $w \in W$ we have $F \in (w)_1$ iff $w \Vdash F$.*

Proof. By induction on F. The cases of the propositional variables and the connectives are easily provable using properties of MCS's and the $\Vdash$ relation. So suppose $F = A \rhd B$.

($\Rightarrow$) Suppose we have $A \rhd B \in (w)_1$. Then for all v such that wRv and $v \Vdash A$ we have to find a u such that $vS_w u \Vdash B$ which, by the induction hypothesis, is equivalent to $B \in (u)_1$. Consider such a v. We have for some S that $(v)_0 = (w)_0 * \langle S \rangle * \tau$ and $(w)_1 \prec^{\Phi}_{S} (v)_1$. By Lemma 8.3.4 there is a MCS Δ such that $(w)_1 \prec^{\Phi}_{S\cup\{\Box\neg A\}} \Delta \ni B$. We take $u = \langle (w)_0 * \langle S \cup \{\Box\neg A\}\rangle, \Delta\rangle$. Now 3.b holds, whence $vS_w u$.

($\Leftarrow$) Suppose that $A \rhd B \notin (w)_1$. Then $\neg(A \rhd B) \in (w)_1$ whence by Lemma 8.3.3 there is a MCS Δ such that $(w)_1 \prec^{\Phi}_{\{\Box\neg A, \neg B\}} \Delta \ni A$. Consider $v' = \langle (w)_0 * \langle \{\Box\neg A, \neg B\}\rangle * \tau, \Delta\rangle$. Clearly, there is no u' such that $v' S_w u' \Vdash B$. ⊣

Part III

On Primitive Recursive Arithmetic

Chapter 9

Comparing PRA and $\mathrm{I}\Sigma_1$

In this chapter we introduce the main subject of Part III of this dissertation; Primitive Recursive Arithmetic (PRA). We shall consider various formulations of PRA. Then we shall focus on how PRA relates to $\mathrm{I}\Sigma_1$, a theme that comes back in Chapters 10 and 11 too.

We shall see that $\mathrm{I}\Sigma_1$ is Π_2-conservative over PRA, but that proofs in PRA can become non-elementary larger than proofs in $\mathrm{I}\Sigma_1$. All proofs in this chapter are proof-theoretical and easily formalizable.

9.1 PRA, what and why?

Primitive Recursive Arithmetic, as a formal system, was first introduced by Skolem in 1923 [Sko67]. Throughout literature there exist many different variants of PRA. In a sense though, they are all the same, as they are easily seen to be equi-interpretable in a faithful way. In this part of the thesis we shall consider theories modulo faithful interpretability. There are two reasons why we are interested in PRA.

Firstly, of course, PRA has an intrinsic foundational importance. It has often been associated with finitism and Hilbert's programme ([Sko67], [HB68], [Tai81]) and can, in a sense, be seen as a system common to both classical and intuitionistic mathematics.

Secondly, a study of PRA, with an interpretability perspective can give us insight on the interpretability logic of all numberized theories. PRA is neither finitely axiomatized nor essentially reflexive. And clearly, $\mathbf{IL}(\mathrm{All}) \subseteq \mathbf{IL}(\mathrm{PRA})$.

Since Π_1-sentences or open formulas played a prominent role in Hilbert's programme, the first versions of PRA were formulated in a quasi-equational setting without quantifiers but with a symbol for every primitive recursive function. (See for example Goodstein [Goo57], or Schwartz [Sch87a], [Sch87b].)

Other formulations are in the full language of predicate logic and also contain a function symbol for every primitive recursive function. The amount of induction can either be for Δ_0-formulas or for open formulas. Both choices yield the same set of theorems. This definition of PRA has, for example, been used in [Smo77].[1]

Reading convention: All statements about PRA and other theories in this part of the thesis, will refer to the definition given in the (sub)section in which the statement appears. If no such specific definition is given, we shall refer to the standard definitions as given in [HP93].

In this subsection, we shall refer with PRA to the theory that is formulated in a language that contains for every primitive recursive function a function symbol plus its defining axioms and that allows for induction over open formulas.

Definition 9.1.1 ($\mathrm{I}\Sigma_n^R$). $\mathrm{I}\Sigma_n^R$ is the predicate logical theory in the pure language of arithmetic $\{+, \cdot, 0, 1, <\}$ that contains Robinson's arithmetic Q plus the Σ_n-induction rule. The Σ_n-induction rule allows one to conclude $\forall x\ \varphi(x, \vec{y})$ from $\varphi(0, \vec{y}) \wedge \forall x\ (\varphi(x, \vec{y}) \rightarrow \varphi(x+1, \vec{y}))$ whenever $\varphi \in \Sigma_n$.

It is well known that PRA is faithfully interpretable in $\mathrm{I}\Sigma_1^R$ in the expected way, that is, every function symbol is replaced by its definition in terms of sequences. For a comparison the other way around, we have the following lemma.

Lemma 9.1.2. $\mathrm{I}\Sigma_1^R\ \subseteq\ \mathrm{PRA}$.

Proof. The proof goes by induction on the length of a proof in $\mathrm{I}\Sigma_1^R$. If $\mathrm{I}\Sigma_1^R \vdash \varphi$ without any applications of the Σ_1 induction rule, it is clear that $\mathrm{PRA} \vdash \varphi$.

So, suppose that the last step in the $\mathrm{I}\Sigma_1^R$-proof of φ were an application of the Σ_1-induction rule. Thus φ is of the form $\forall x\, \exists y\ \varphi_0(x, y, \vec{z})$ and we obtain shorter $\mathrm{I}\Sigma_1^R$-proofs of the Σ_1-statements $\exists y\ \varphi_0(0, y, \vec{z})$ and $\exists y'\ (\varphi_0(x, y, \vec{z}) \rightarrow \varphi_0(x+1, y', \vec{z}))$. The induction hypothesis tells us that these statements are also provable in PRA. Herbrand's theorem for PRA provides us with primitive recursive functions $g(\vec{z})$ and $h(x, y, \vec{z})$ such that

$$\mathrm{PRA} \vdash \varphi_0(0, g(\vec{z}), \vec{z}) \qquad (1)$$

and

$$\mathrm{PRA} \vdash \varphi_0(x, y, \vec{z}) \rightarrow \varphi_0(x+1, h(x, y, \vec{z}), \vec{z}) \qquad (2)$$

Let $f(x, \vec{z})$ be the primitive recursive function defined by

$$\begin{cases} f(0, \vec{z}) = g(\vec{z}), \\ f(x+1, \vec{z}) = h(x, f(x, \vec{z}), \vec{z}). \end{cases}$$

By (1) and (2) it follows from Δ_0-induction in PRA that $\mathrm{PRA} \vdash \forall x\ \varphi_0(x, f(x, \vec{z}), \vec{z})$ whence $\mathrm{PRA} \vdash \forall x \exists y\ \varphi_0(x, y, \vec{z})$. ⊣

[1] Confusingly enough Smoryński later defines in [Smo85] a version of PRA which is equivalent to IΣ$_1$.

In [Bek97], a characterization of $\mathrm{I}\Sigma_n^R$ is given in terms of reflection principles. Reflection principles turn out to be very useful in axiomatizing arithmetical theories.

For a theory T and a class of formulas Γ we define the uniform reflection principle for Γ over T to be a set of formulas in the following way: $\mathrm{RFN}_\Gamma(T) := \{\forall x\ (\Box_T \gamma(\dot{x}) \rightarrow \gamma(x)) \mid \gamma \in \Gamma\}$. This set of formulas is often equivalent to a single formula also denoted by $\mathrm{RFN}_\Gamma(T)$. For ordinals $\alpha \leq \omega$ we define $(T)_0^\Gamma := T$, $(T)_{\alpha+1}^\Gamma := (T)_\alpha^\Gamma + \mathrm{RFN}_\Gamma((T)_\alpha^\Gamma)$ and $(T)_\omega^\Gamma := \cup_{\beta<\omega}(T)_\beta^\Gamma$. This can be extended to transfinite ordinals, provided an elementary system of ordinal notation is given. If Γ is just the class of Π_n-formulas we write $(T)_\alpha^n$ instead of $(T)_\alpha^{\Pi_n}$.

The following facts illustrate the usefulness of reflection principles in describing theories. Identity in the statements below, refer to the set of theorems of a theory.

Theorem 9.1.3 (Leivant ([Lei83])). $\mathrm{I}\Sigma_n = \mathrm{RFN}_{\Pi_{n+2}}(\mathrm{EA})$ $(n \geq 1)$

Theorem 9.1.4 (Beklemishev ([Bek97])). $\mathrm{I}\Sigma_n^R = (\mathrm{EA})_\omega^{n+1}$ $(n \geq 1)$

Fact 9.1.5 (Kreisel, Levy ([KL68])). *$T + \mathrm{RFN}_{\Pi_n}(T)$ is not contained in any consistent Σ_n-extension of T.*

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

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

By compactness we have for some particular Σ_n-sentence σ that $T + \sigma \vdash \forall x\ (\Box_T(\mathsf{Tr}_{\Pi_n}(\dot{x})) \rightarrow \mathsf{Tr}_{\Pi_n}(x))$. Consequently, $T + \sigma \vdash \forall x\ (\Box_T(\mathsf{Tr}_{\Pi_n}(\ulcorner\neg\sigma\urcorner)) \rightarrow \mathsf{Tr}_{\Pi_n}(\ulcorner\neg\sigma\urcorner))$ and thus $T \vdash \sigma \rightarrow (\Box_T(\ulcorner\neg\sigma\urcorner) \rightarrow \neg\sigma)$. But we also have $T \vdash \neg\sigma \rightarrow (\Box_T(\ulcorner\neg\sigma\urcorner) \rightarrow \neg\sigma)$ hence $T \vdash \Box_T(\ulcorner\neg\sigma\urcorner) \rightarrow \neg\sigma$. Löb's rule gives us $T \vdash \neg\sigma$ in which case $T+S$ is inconsistent. ⊣

Corollary 9.1.6. *$\mathrm{I}\Sigma_n$ $(n \geq 1)$ is not contained in any consistent Σ_{n+2}-extension of* EA.

Proof. By Theorem 9.1.3 and Fact 9.1.5. ⊣

From Theorem 9.1.4 we see that we can also take $(\mathrm{EA})_\omega^2$ as a definition for PRA. Whenever we shall do so, we shall also work with the canonical axiomatization of $(\mathrm{EA})_\omega^2$, that is, the following.

Since we have partial truth definitions and we are talking global reflection we have that $\{\forall x\ (\Box_T \pi(\dot{x}) \rightarrow \pi(x)) \mid \pi \in \Pi_2\}$ can be expressed by a single sentence $\mathrm{RFN}_{\Pi_2}(T)$. Let ϵ denote the sentence axiomatizing EA. We define a sequence of axioms π_i as follows.

$$\pi_0 := \epsilon,$$

$$\pi_{m+1} := \pi_m \wedge \forall^{\Pi_2}\varphi\ (\Box_{\pi_m}\varphi \rightarrow \mathsf{True}_{\Pi_2}(\varphi)).$$

Clearly, π_m axiomatizes $(\mathrm{EA})^2_m$. An advantage of this axiomatization, is that the fact that any Σ_2-extension of PRA is reflexive, becomes immediate.

Lemma 9.1.7. *(In* EA.*) Any* Σ_2*-extension of* $(\mathrm{EA})^2_\omega$ *is reflexive.*

Proof. Let σ be a Σ_2-sentence. Reason in $(\mathrm{EA})^2_\omega + \sigma$ and assume for a contradiction that $\Box_{(\mathrm{EA})^2_m+\sigma}\bot$ for some m. We then have $\Box_{(\mathrm{EA})^2_m}\neg\sigma$ and by π_{m+1} we obtain $\neg\sigma$. $\dashv$

9.2 Parsons' theorem

Parsons' theorem says that $I\Sigma_1$ is Π_2-conservative over PRA. It was proved independently by C. Parsons ([Par70], [Par72]), G. Mints ([Min72]) and G. Takeuti ([Tak75]). As PRA is often associated with finitism, Parsons' theorem can be considered of great importance as a partial realization of Hilbert's programme.

The first proofs of Parsons' theorem were all of proof-theoretical nature. Parsons' first proof, [Par70], is based upon Gödel's Dialectica interpretation. His second proof, [Par72], merely relies on a cut elimination. Mints' proof, [Min72], employs the no-counterexample interpretation of a special sequent calculus. The proof by Takeuti, [Tak75], employs an ordinal analysis in the style of Gentzen.

Over the years, many more proofs of Parsons' theorem have been published. In many accounts Herbrand's theorem plays a central role in providing primitive recursive Skolem functions for Π_2-statements provable in $I\Sigma_1$. (Cf. Sieg's method of Herbrand analysis [Sie91], Avigad's proof by his notion of Herbrand saturated models [Avi02], Buss's proof by means of his witness functions [Bus98], and Ferreira's proof using Herbrand's theorem for Σ_3 and Σ_1-formulas [Fer02].) A first model-theoretic proof is due to Paris and Kirby. They employ semi-regular cuts in their proof (cf. [Sim99]:373-381).

In this thesis, we will add two more proofs to the long list. The first proof is given in Section 9.2.1. It is a proof-theoretic proof and can be seen as a modern version of Parsons' second proof. The main ingredient is the cut elimination theorem for Tait's sequent calculus.

The second proof is given in Section 10.1. It is a model-theoretic proof. A central ingredient is an analysis of the difference between PRA and $I\Sigma_1$ in terms of iteration of total functions.

9.2.1 A proof-theoretic proof of Parsons' theorem

The first proof we give of Parsons' theorem is proof-theoretic. Our presentation is due to L. Beklemishev. It will become evident that the whole argument is easily formalizable as soon as the superexponential function[2] is provably total. This is because our proof only uses the standard cut elimination theorem.

In this section we will work with a fragment of first order predicate logic that only contains $\wedge, \vee, \forall, \exists$ and $\neg$, where $\neg$ may only occur on the level of atomic

[2] We shall write supexp, both for the function itself, as for the sentence asserting its totality.

formulae. We can define → and unrestricted negation as usual. We shall thus freely use these connectives too.

A proof system for this fragment of logic in the form of a Tait calculus is provided in [Sch77]. We will use this calculus in our proof. The calculus works with sequents which are finite sets and should be read disjunctively in the sense that $\Gamma = \{\varphi_1, \ldots, \varphi_n\}$ stands for $\varphi_1 \vee \ldots \vee \varphi_n$. We will omit the set-brackets $\{\}$. The axioms of the Tait calculus are:

$$\Gamma, \varphi, \neg\varphi \quad \text{for atomic } \varphi.$$

The rules are:

$$\frac{\Gamma, \varphi \quad \Gamma, \psi}{\Gamma, \varphi \wedge \psi}, \qquad \frac{\Gamma, \varphi}{\Gamma, \varphi \vee \psi}, \qquad \frac{\Gamma, \psi}{\Gamma, \varphi \vee \psi},$$

$$\frac{\Gamma, \varphi(a)}{\Gamma, \forall x\, \varphi(x)}, \qquad \frac{\Gamma, \varphi(t)}{\Gamma, \exists x\, \varphi(x)},$$

plus the cut rule

$$\frac{\Gamma, \varphi \quad \Gamma, \neg\varphi}{\Gamma}.$$

In the rule for the universal quantifier introduction it is necessary that the a does not occur free anywhere else in Γ. And in the rule for the introduction of the existential quantifier one requires t to be substitutable for x in φ. In our proof we use the nice properties that this calculus is known to posses. Most notably the cut elimination theorem and some inversion properties.

Let us now fix our versions of PRA and $\mathrm{I}\Sigma_1$ for this section.

Definition 9.2.1 ($\mathrm{I}\Sigma_1$)**.** The theory $\mathrm{I}\Sigma_1$ is an extension of predicate logic with some easy Π_1-fragment of arithmetic (for example the Π_1-part of Robinson's arithmetic Q), together with all axioms of the form

$$\forall x\ (\neg\mathsf{Progr}(\varphi, x) \vee \varphi(x, \vec{y})).$$

Here φ is some Σ_1-formula and $\mathsf{Progr}(\varphi, x)$ is the Π_2-formula that is equivalent to

$$\varphi(0, \vec{y}) \wedge \forall x\ (\varphi(x, \vec{y}) \to \varphi(x+1, \vec{y})).$$

Definition 9.2.2 (PRA)**.** The theory $\mathrm{I}\Sigma_1^R$, also called Primitive Recursive Arithmetic, is the extension of predicate logic that arises by adding a simple Π_1-fragment of arithmetic[3] together with the Σ_1-induction rule to it. Here, the Σ_1-induction rule is

$$\frac{\Gamma, \varphi(0, \vec{y}) \quad \Gamma, \forall x\ (\neg\varphi(x, \vec{y}) \vee \varphi(x+1, \vec{y}))}{\Gamma, \varphi(t, \vec{y})}.$$

where Γ is a Π_2-sequent, φ a Σ_1-formula and t is free for x in φ.

[3] The same fragment as in Definition 9.2.1.

Theorem 9.2.3. $\mathrm{I}\Sigma_1$ *is* Π_2*-conservative over* $\mathrm{I}\Sigma_1^R$.

Proof. So, our aim is to prove that if $\mathrm{I}\Sigma_1 \vdash \pi$ then $\mathrm{I}\Sigma_1^R \vdash \pi$ whenever π is a Π_2-sentence. If $\mathrm{I}\Sigma_1 \vdash \pi$, then by induction on the length of such a proof we see that some sequent Σ, π is provable in the pure predicate calculus. Here Σ is a finite set of negations of axioms of $\mathrm{I}\Sigma_1$. By the cut elimination theorem for the Tait calculus we know that there exists a cut-free derivation of the sequent. Thus we also have the sub-formula property (modulo substitution of terms) for our cut-free proof of Σ, π.

The proof is concluded by showing by induction on the length of cut-free derivations that if a sequent of the form Σ, Π is derivable then $\mathrm{I}\Sigma_1^R \vdash \Pi$. Here Σ is a finite set of negations of induction axioms of Σ_1-formulas and Π is a finite non-empty set of Π_2-formulas.

The basis case is trivial. So, for the inductive step, suppose we have a cut-free proof of Σ, Π. What can be the last step in the proof of this sequent? Either the last rule yielded something in the Π-part of the sequent or in the Σ-part of it. In the first case nothing interesting happens and we almost automatically obtain the desired result by the induction hypothesis.

So, suppose something had happened in the Σ-part. All formulas in this part are of the form $\exists\, a\ [\varphi(0) \wedge \forall\, x\ (\varphi(x) \to \varphi(x+1)) \wedge \neg\varphi(a)]$, with $\varphi \in \Sigma_1$.

The last deduction step thus must have been the introduction of the existential quantifier and we can by a one step shorter proof derive for some term t the following sequent.

$$\Sigma', \varphi(0) \wedge \forall\, x\ (\varphi(x) \to \varphi(x+1)) \wedge \neg\varphi(t), \Pi$$

By the inversion property of the Tait calculus (for a proof and precise formulation of the statement consult e.g. [Sch77] page 873) we obtain proofs of the same length of the following sequents

$$\Sigma', \varphi(0), \Pi\ , \quad \Sigma', \forall\, x\ (\varphi(x) \to \varphi(x+1)), \Pi \quad \text{and} \quad \Sigma', \neg\varphi(t), \Pi.$$

As all of $\varphi(0), \forall\, x\ (\varphi(x) \to \varphi(x+1))$ and $\neg\varphi(t)$ are Π_2-formulas, we can apply the induction hypothesis to conclude that we have the following.

$$\begin{array}{lcll} \mathrm{I}\Sigma_1^R & \vdash & \varphi(0), \Pi & (1) \\ \mathrm{I}\Sigma_1^R & \vdash & \forall\, x\ (\varphi(x) \to \varphi(x+1)), \Pi & (2) \\ \mathrm{I}\Sigma_1^R & \vdash & \neg\varphi(t), \Pi & (3) \end{array}$$

Recall that Π consists of Π_2-statements. So, we can apply the Σ_1-induction rule to (1) and (2) and obtain $\varphi(t), \Pi$. This together with (3) yields by one application of the cut rule (in $\mathrm{I}\Sigma_1^R$) the desired result, that is, $\mathrm{I}\Sigma_1^R \vdash \Pi$. ⊣

Corollary 9.2.4. $\mathrm{I}\Sigma_n$ *is* Π_{n+1}*-conservative over* $\mathrm{I}\Sigma_n^R$.

Proof. $\mathrm{I}\Sigma_n^R$ is defined as the canonical generalization of Definition 9.2.2. Changing the indices in the proof of Theorem 9.2.3 immediately yields the result. ⊣

In [Bek03a] this result is stated as Corollary 4.8. It is a corollary of his Reduction property, Theorem 2, which is also formalizable in the presence of the superexponential function. The proof of Parsons' theorem we have presented here is very close to the proof of the reduction property.

9.3 Cuts, consistency and length of proofs

A direct consequence of the formalizability of Parsons' theorem is that PRA and $\mathrm{I}\Sigma_1$ are equi-consistent. To be more precise, for every theory T proving the totality of the superexponentiation we have that

$$T \vdash \mathsf{Con}(\mathrm{PRA}) \leftrightarrow \mathsf{Con}(\mathrm{I}\Sigma_1).$$

Consequently $\mathrm{I}\Sigma_1 \nvdash \mathsf{Con}(\mathrm{PRA})$. In this section we take PRA to be $\mathrm{I}\Sigma_1^R$ and shall see that we can find a definable $\mathrm{I}\Sigma_1$-cut J such that $\mathrm{I}\Sigma_1 \vdash \mathsf{Con}^J(\mathrm{PRA})$. More generally, we shall show that for all n, we can find an $\mathrm{I}\Sigma_n$-cut J_n, such that[4] $\mathrm{I}\Sigma_n + \sigma \vdash \mathsf{Con}^{J_n}(\mathrm{I}\Sigma_n^R + \sigma)$ for any $\sigma \in \Sigma_{n+1}$.

As in [Pud86] and [Ign90] we note that Theorem 9.3.1 implies that certain proofs in PRA must be non-elementary larger than their counterparts in $\mathrm{I}\Sigma_1$. This, in a sense, says that the use of the cut elimination, whence the super exponential blow-up, in the proof of Theorem 9.2.3 was essential.

To the best of our knowledge Ignjatovic ([Ign90]) showed for the first time that $\mathrm{I}\Sigma_1$ proves the consistency of PRA on some definable cut. His reasoning was based on a paper by Pudlák ([Pud86]). Pudlák showed in this paper by model-theoretic means that GB proves the consistency of ZF on a cut. The cut that Ignjatovic exposes is actually an RCA_0-cut. (See for example [Sim99] for a definition of RCA_0.)

The elements of Ignjatovic' cut correspond to complexities of formulas for which a sort of truth-predicate is available. By an interpretability argument it is shown that a corresponding cut can be defined in $\mathrm{I}\Sigma_1$. It seems straight-forward to generalize his result to obtain Corollary 9.3.2

Theorem 9.3.1. *(In* EA*.) For each $n{\in}\omega$ with $n{\geq}\ 1$, there exists some $\mathrm{I}\Sigma_n$-cut J_n such that for all Σ_{n+1}-sentences σ, $\mathrm{I}\Sigma_n + \sigma \vdash \mathsf{Con}^{J_n}(\mathrm{I}\Sigma_n^R + \sigma)$.*

Proof. We recall that $\mathrm{I}\Sigma_n^R \equiv (\mathrm{EA})_\omega^{n+1}$. Let ϵ be the arithmetical sentence axiomatizing EA. In analogy with Section 9.1 we fix the following axiomatization $\{i_m^n\}_{m\in\omega}$ of $\mathrm{I}\Sigma_n^R$:

$$i_0^n := \epsilon,$$

$$i_{m+1}^n := i_m^n \wedge \forall^{\Pi_{n+1}}\pi\ (\Box_{i_m^n}\pi \rightarrow \mathsf{True}_{\Pi_{n+1}}(\pi)).$$

The map that sends m to the code of i_m^n is clearly primitive recursive. We will assume that the context makes clear if we are talking about the formula or its

[4] With $\mathrm{I}\Sigma_n^R + \sigma$ we mean the theory axiomatized by σ and all theorems of $\mathrm{I}\Sigma_n^R$.

code when writing i^n_m. Similarly for other formulas. An IΣ_n-cut J_n is defined in the following way.

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

We will now see that J'_n defines an initial segment in IΣ_n. Clearly I$\Sigma_n \vdash J'_n(0)$. It remains to show that I$\Sigma_n \vdash J'_n(m) \to J'_n(m{+}1)$.

So, we reason in IΣ_n and assume $J'_n(m)$. We need to show that $\mathsf{True}_{\Pi_{n+1}}(i^n_{m+1})$, that is,

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

Our assumption gives us $\mathsf{True}_{\Pi_{n+1}}(i^n_m)$ thus we need to show $\mathsf{True}_{\Pi_{n+1}}(\forall^{\Pi_{n+1}}\pi\ (\Box_{i^n_m}\pi \to \mathsf{True}_{\Pi_{n+1}}(\pi)))$ or, equivalently, $\forall^{\Pi_{n+1}}\pi\ (\Box_{i^n_m}\pi \to \mathsf{True}_{\Pi_{n+1}}(\pi))$. The latter is equivalent to

$$\forall^{\Pi_{n+1}}\pi\ \Box_{\mathrm{EA}}(\mathsf{True}_{\Pi_{n+1}}(i^n_m) \to \mathsf{True}_{\Pi_{n+1}}(\pi)) \to \mathsf{True}_{\Pi_{n+1}}(\pi). \tag{9.1}$$

But as $\mathsf{True}_{\Pi_{n+1}}(i^n_m) \to \mathsf{True}_{\Pi_{n+1}}(\pi) \in \Pi_{n+2}$, and as I$\Sigma_n \equiv \mathrm{RFN}_{\Pi_{n+2}}(\mathrm{EA})$, we get that

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

We again use our assumption $\mathsf{True}_{\Pi_{n+1}}(i^n_m)$ to obtain (9.1). Thus indeed, $J'_n(x)$ defines in initial segment. By well known techniques, J'_n can be shortened to a definable cut.

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

Corollary 9.3.2. *There exists an* IΣ_1*-cut* J *such that for any* Σ_2 *sentence* σ *we have* I$\Sigma_1 + \sigma \vdash \mathsf{Con}^J(\mathrm{PRA} + \sigma)$.

Proof. Immediate from Theorem 9.3.1 as PRA $=$ IΣ^R_1. ⊣

Chapter 10

Models for PRA and $\mathrm{I}\Sigma_1$

In this chapter we study how models of PRA compare to models of $\mathrm{I}\Sigma_1$. As a result of this study, we give a model-theoretic proof of Parsons' theorem. In Section 10.2 we shall see a second proof of the fact that $\mathrm{I}\Sigma_1$ proves the consistency of PRA on a definable cut. The proof does not make use of our previous study on models. However, we shall work with a formulation of PRA that is reminiscent to the one used in Section 10.1.

10.1 A model theoretic proof of Parsons' theorem

In this section we shall give a model theoretic proof of Parsons' theorem. Our proof has the following outline.

In Subsection 10.1.1 we give a slightly renewed proof of a theorem by Gaifman and Dimitracopoulos. This theorem says that under certain conditions a definitional extension of a theory has nice properties, like proving enough induction.

In Subsection 10.1.2 we use this theorem to give a characterization of $\mathrm{I}\Sigma_1$ in terms of PRA and closure under iteration of a certain class of functions. In Theorem 10.1.12 we will see what it takes for a model $\mathcal{M}$ of PRA to also be a model of $\mathrm{I}\Sigma_1$: A class of functions of this model should be majorizable by another class of functions.

This theorem is at the heart of our model theoretic proof of Parsons' theorem in Subsection 10.1.3. We will show that any countable model $\mathcal{N}$ of PRA falsifying $\pi \in \Pi_2$ can be extended to a countable model $\mathcal{N}'$ of $\mathrm{I}\Sigma_1 + \neg\pi$ whence $\mathrm{I}\Sigma_1 \nvdash \pi$. In extending the model we will, having Theorem 10.1.12 in the back of our mind, repeatedly majorize functions to finally obtain a model of $\mathrm{I}\Sigma_1 + \neg\pi$.

Our proof is based on a proof sketch in an unpublished note of Visser ([Vis90b]). The very same note inspired Zambella in his [Zam96] for a proof of a conservation result of Buss' S^1_2 over PV.

First, we fix some formulation of PRA and IΣ_1 that suits the purposes of this section.

Definition 10.1.1. The *language* of PRA is the language of PA plus a family of new function symbols $\{\mathsf{Sup}_n \mid n{\in}\omega\}$. The *non-logical axioms of* PRA come in three sorts.

- Defining axioms for $+$, $\cdot$, and $<$,[1]
- Defining axioms for the new symbols
  - $\forall x\ \mathsf{Sup}_0(x) = 2x$,
  - $\{\mathsf{Sup}_{n+1}(0) = 1\}$,
  - $\{\forall x\ \mathsf{Sup}_{n+1}(x+1) = \mathsf{Sup}_n(\mathsf{Sup}_{n+1}(x)) \mid n \in \omega\}$,
- Induction axioms for $\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$-formulas in the following form: $\forall x\ (\varphi(0) \wedge \forall y{<}x\ (\varphi(y) \rightarrow \varphi(y+1)) \rightarrow \varphi(x))$.

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

The functions Sup_i describe on the standard model a well-known hierarchy; Sup_0 is the doubling function, Sup_1 is the exponentiation function, Sup_2 is superexponentiation, Sup_3 is superduperexponentiation and so on. It is also known that the Sup_i form an envelope for PRA, that is, every provably total recursive function of PRA gets eventually majorized by some Sup_i. (Essentially this is Parikh's theorem [Par71].) Consequently all terms of PRA are majorizable by a strictly monotone one.

PRA proves all the evident properties of the Sup_i functions like $\mathsf{Sup}_n(1) = 2$, $1 \leq \mathsf{Sup}_{n+1}(y)$, $x \leq y \rightarrow \mathsf{Sup}_n(x) \leq \mathsf{Sup}_n(y)$, $n{\leq}m \rightarrow \mathsf{Sup}_n(x){\leq}\mathsf{Sup}_m(y)$ and so on. Of course PRA proves in a trivial way the totality of all the Sup_i as these symbols form part of our language. We have chosen an equivalent variant of the usual induction axiom so that we end up with a Π_1-axiomatization of PRA. It is easy to see that our definition of PRA is equivalent, or more precisely equi-interpretable, to any other of our definitions of PRA.

Definition 10.1.2. The theory IΣ_1 is the theory that is obtained by adding to PRA induction axioms $\varphi(0) \wedge \forall x\ (\varphi(x) \rightarrow \varphi(x+1)) \rightarrow \forall x\ \varphi(x)$ for all $\Sigma_1(\{\mathsf{Sup}_i\}_{i\in\omega})$-formulas $\varphi(x)$ that may contain additional parameters.

Reading conventions Throughout this section we will adhere to the following notational convention. Arithmetical formulas defining the graph of a function are denoted by lowercase Greek letters. The corresponding lower case Roman letter is reserved to be the symbol that refers to the function described by its graph. By the corresponding upper case Roman letter we will denote the very short formula that defines the graph using the lower case Roman letter and

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

the identity symbol only. Context, like indices and so forth, are inherited in the expected way.

For example, if $\chi_n(x, y)$ is an arithmetical formula describing a function, in a richer language this function will be referred to by the symbol g_n. The corresponding G_n will refer to the simple formula $g_n(x) = y$ in the enriched language.

10.1.1 Introducing a new function symbol

In our discussion we shall like to work with a theory that arises as an extension of PRA by a definition. We will add a new function symbol f to the language of PRA together with the axiom φ that defines f. Moreover we would like to employ induction that involves this new function symbol, possibly also in the binding terms of the bounded quantifiers. We will see that if the function f allows for a simple definition and has some nice properties we have indeed access to the extended form of induction.

Essentially the justification boils down to a theorem of Gaifman and Dimitracopoulos [GD82] a proof of which can also be found in [HP93] (Theorem 1.48 and Proposition 1.3). We will closely follow here a proof of Beklemishev from [Bek97] which we slightly improved and modified.

We first give the necessary definitions before we come to formulate the main result, Theorem 10.1.8

Definition 10.1.3 ($\Delta_0(\{g_i\}_{i\in I})$-formulas, $\mathrm{I}\Delta_0(\{g_i\}_{i\in I})$).
Let $\{g_i\}_{i\in I}$ be a set of function symbols. The $\Delta_0(\{g_i\}_{i\in I})$-formulas are the bounded formulas in the language of PA enriched with the function symbols $\{g_i\}_{i\in I}$. The new function symbols are also allowed to occur in the binding terms of the bounded quantifiers. By $\mathrm{I}\Delta_0(\{g_i\}_{i\in I})$ we mean the theory that comprises

- some open axioms describing some minimal arithmetic[2],
- induction axioms for all $\Delta_0(\{g_i\}_{i\in I})$-formulas and
- (possibly) defining axioms of the symbols $\{g_i\}_{i\in I}$.

The defining axioms of the symbols $\{g_i\}_{i\in I}$ are denoted by $\mathcal{D}(\{g_i\}_{i\in I})$.

From now on, we may thus write $\mathrm{I}\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$ instead of PRA.

Definition 10.1.4 ($\mathsf{Tot}(\varphi)$, $\mathsf{Mon}(\varphi)$).
Let $\varphi(x, y)$ be a $\Delta_0(\{g_i\}_{i\in I})$ formula. By $\mathsf{Tot}(\varphi)$ we shall denote the formula $\forall x\, \exists!\, y\ \varphi(x, y)$[3] stating that φ can be regarded as a total function. By $\mathsf{Mon}(\varphi)$ we shall denote the formula $\forall x, x', y, y'\ (x \leq x' \wedge \varphi(x, y) \wedge \varphi(x', y') \to y \leq y') \wedge \mathsf{Tot}(\varphi)$ stating the monotonicity of the total φ.

[2] For example the open part of Robinson's arithmetic.

[3] That is, $\forall x\, \exists y\ \varphi(x, y) \wedge \forall x\, \forall y\, \forall y'\ (\varphi(x, y) \wedge \varphi(x, y') \to y = y')$.

Definition 10.1.5 ($\Delta_0(\{g_i\}_{i\in I}, F)$-formula, $\mathrm{I}\Delta_0(\{g_i\}_{i\in I}, F)$).
Let φ be such that $\mathrm{I}\Delta_0(\{g_i\}_{i\in I}) \vdash \mathsf{Tot}(\varphi)$. Recall that the uppercase letter F paraphrases the formula $f(x) = y$. A $\Delta_0(\{g_i\}_{i\in I}, F)$-formula is a $\Delta_0(\{g_i\}_{i\in I})$-formula possibly containing occurrences of F. By $\mathrm{I}\Delta_0(\{g_i\}_{i\in I}, F)$ we denote the theory $\mathrm{I}\Delta_0(\{g_i\}_{i\in I})$ where we now also have induction for $\Delta_0(\{g_i\}_{i\in I}, F)$ formulas. The defining axiom of f, in our case φ, is also in $\mathrm{I}\Delta_0(\{g_i\}_{i\in I}, F)$.

Note that f cannot occur in a bounding term in an induction axiom of $\mathrm{I}\Delta_0(\{g_i\}_{i\in I}, F)$. Also note that F is nothing but a formula containing f stating $f(x) = y$ and consists of just six symbols (if f is unary). Of course later we will substitute for F an arithmetical definition of the graph of f, that is, $\varphi(x, y)$.

The main interest of the extension of $\mathrm{I}\Delta_0(\{g_i\}_{i\in I})$ by a definition of f is in Theorem 10.1.8 and in its Corollary 10.1.9. The latter says that we can freely use $f(x)$ as an abbreviation of $\varphi(x, y)$ and have access to $\Delta_0(\{g_i\}_{i\in I}, f)$-induction whenever f has a $\Delta_0(\{g_i\}_{i\in I})$ graph and is provably total and monotone in $\mathrm{I}\Delta_0(\{g_i\}_{i\in I})$.

First we prove some technical but rather useful lemmata. They are slight improvements of Beklemishev's Lemma 5.12 and 5.13 from [Bek97]. From now on we will work under the assumptions of Theorem 10.1.8 so that $\mathrm{I}\Delta_0(\{g_i\}_{i\in I})$ is such that any term t in its language is provably majorizable by some other term $\tilde{t}$ that is strictly increasing in all of its arguments. Throughout the forthcoming proofs we will for any term t denote by $\tilde{t}$ such a term that is provably strictly monotone (in all of its arguments) and majorizing t.

Lemma 10.1.6. *For every term $s(\vec{a})$ of $\mathrm{I}\Delta_0(\{g_i\}_{i\in I}, f)$ and every $R \in \{\leq, \geq, =, <, >\}$ there are terms t^R_s and $\tilde{s}(a)$ strictly increasing in all of their arguments and a $\Delta_0(\{g_i\}_{i\in I}, F)$-formula $\psi^R_s(\vec{a}, b, y)$ such that $\mathrm{I}\Delta_0(\{g_i\}_{i\in I}, F) + \mathsf{Mon}(\varphi) \vdash \forall\, y{\geq}t^R_s(\vec{a})\ (s(\vec{a})Rb \leftrightarrow \psi^R_s(\vec{a}, b, y))$ and $\mathrm{I}\Delta_0(\{g_i\}_{i\in I}, F) + \mathsf{Mon}(\varphi) \vdash \forall\, \vec{x}\ (s(\vec{x}) \leq \tilde{s}(\vec{x}))$.*

Proof. The proof proceeds by induction on $s(\vec{a})$. In the basis case nothing has to be done as x_iRb, $0Rb$ and $1Rb$ are all atomic $\Delta_0(\{g_i\}_{i\in I}, F)$-formulas. Moreover all of the x_i, 0 and 1 are (provably) strictly monotone in all of their arguments. For the induction case consider $s(\vec{a}) = h(s_1(\vec{a}))$, where h is either one of the g_i or $h = f$. For simplicity we assume here that h is a unary function.

The induction hypothesis provides us with a $\Delta_0(\{g_i\}_{i\in I}, F)$-formula $\psi^=_{s_1}(\vec{a}, b, y)$ and terms $t^=_{s_1}(\vec{a})$ and $\tilde{s}_1(\vec{a})$ such that

$$\mathrm{I}\Delta_0(\{g_i\}_{i\in I}, F) + \mathsf{Mon}(\varphi) \vdash \forall\, y{\geq}t^=_{s_1}(\vec{a})\ (s_1(\vec{a}) = b \leftrightarrow \psi^=_{s_1}(\vec{a}, b, y)),$$

and

$$\mathrm{I}\Delta_0(\{g_i\}_{i\in I}, F) + \mathsf{Mon}(\varphi) \vdash \forall\, \vec{x}\ (s_1(\vec{x}) \leq \tilde{s}_1(\vec{x})).$$

We now want to say that $h(s_1(\vec{a}))Rb$ in a $\Delta_0(\{g_i\}_{i\in I}, F)$ way. This can be done by $\exists\, y', y''{\leq}y\ (\psi^=_{s_1}(\vec{a}, y', y) \wedge h(y') = y'' \wedge y''Rb)$ whenever $y \geq t^=_{s_1}(\vec{a}) + \tilde{s}(\vec{a})$. Here we define $\tilde{s}(\vec{a})$ to be just $f(\tilde{s}_1(\vec{a}))$ in case $h = f$ and $\tilde{g}_i(\tilde{s}_1(\vec{a}))$ in case $h = g_i$. Clearly $\mathrm{I}\Delta_0(\{g_i\}_{i\in I}, F) + \mathsf{Mon}(\varphi) \vdash \forall\, \vec{x}\ (s(\vec{x}) \leq \tilde{s}(\vec{x}))$. Indeed one easily sees that

$$\mathrm{I}\Delta_0(\{g_i\}_{i\in I}, F) + \mathsf{Mon}(\varphi) \vdash \forall\, y{\geq}t^{=}_{s_1}(\vec{a}) + \tilde{s}(\vec{a})\ \ [h(s_1(\vec{a}))Rb \leftrightarrow \exists\, y', y''{\leq}y\ (\psi^{=}_{s_1}(\vec{a}, y', y) \wedge h(y') = y'' \wedge y''Rb)].$$

It is also easy to see that $t^{=}_{s_1}(\vec{a}) + \tilde{s}(\vec{a})$ is indeed monotone. In case $h = f$ we need $\mathsf{Mon}(\varphi)$ here.

A similar reduction applies to the case when the function g has more than one argument. ⊣

It is possible to simplify the above reduction a bit by distinguishing between $h = f$ and $h \neq f$ and also $R ==$ and $R \neq=$, or by proving the lemma just for $R ==$ and showing that all the other cases can be reduced to this. We are not very much interested in optimality at this point though.

Lemma 10.1.7. *For every* $\Delta_0(\{g_i\}_{i\in I}, f)$*-formula* $\theta(\vec{a})$ *there is a* $\Delta_0(\{g_i\}_{i\in I}, F)$*-formula* $\theta_0(\vec{a}, y)$ *and a provably monotonic term* $t_\theta(\vec{a})$ *such that* $\mathrm{I}\Delta_0(\{g_i\}_{i\in I}, F) + \mathsf{Mon}(\varphi) \vdash \forall\, y{\geq}t_\theta(\vec{a})\ \ (\theta(\vec{a}) \leftrightarrow \theta_0(\vec{a}, y))$.

Proof. The lemma is proved by induction on θ.

- Basis. In this case $\theta(\vec{a})$ is $s_1(\vec{a})R\, s_2(\vec{a})$. Applying Lemma 10.1.6 we see that[4] $s_1(\vec{a})R\, s_2(\vec{a}) \leftrightarrow \exists\, b{\leq}y\ (\psi^{=}_{s_2}(\vec{a}, b, y) \wedge \psi^{R}_{s_1}(\vec{a}, b, y))$ whenever $y \geq t_{s_1}(\vec{a}) + t_{s_2}(\vec{a})$.

- The only interesting induction case is where a bounded quantifier is involved. We consider the case when $\theta(\vec{a})$ is $\exists\, x{\leq}s(\vec{a})\ \xi(\vec{a}, x)$. The induction hypothesis yields a provably monotonic term $t_\xi(\vec{a}, x)$ and a $\Delta_0(\{g_i\}_{i\in I}, F)$-formula $\xi_0(\vec{a}, x, y)$ such that provably

  $$\forall\, y{\geq}t_\xi(\vec{a}, x)\ (\xi(\vec{a}, x) \leftrightarrow \xi_0(\vec{a}, x, y))$$

  . Combining this with Lemma 10.1.6 we get that provably

  $$\exists\, x{\leq}s(\vec{a})\ \xi(\vec{a}, x) \leftrightarrow \exists\, x'{\leq}y\ (\psi^{=}_{s}(\vec{a}, x', y) \wedge \exists\, x{\leq}x'\ \xi_0(\vec{a}, x, y))^{5}$$

  whenever $y \geq \tilde{s}(\vec{a}) + t^{=}_{s}(\vec{a}) + t_\xi(\vec{a}, \tilde{s}(\vec{a}))$.

⊣

Theorem 10.1.8. *Let* $\mathrm{I}\Delta_0(\{g_i\}_{i\in I})$ *be such that any term* t *in its language is provably majorizable by some other term* $\tilde{t}$ *that is strictly increasing in all of its arguments. We have that* $\mathrm{I}\Delta_0(\{g_i\}_{i\in I}, F) + \mathsf{Mon}(\varphi) \vdash \mathrm{I}\Delta_0(\{g_i\}_{i\in I}, f)$.

[4] If we only want to use Lemma 10.1.6 with R being $=$ we can observe that $s_1(\vec{a})R\, s_2(\vec{a}) \leftrightarrow \exists\, b, c{\leq}y\ (\psi^{=}_{s_1}(\vec{a}, b, y) \wedge \psi^{=}_{s_2}(\vec{a}, c, y) \wedge bRc)$ whenever $y \geq t_{s_1}(\vec{a}) + t_{s_2}(\vec{a})$.

[5] Alternatively, one could take $\exists\, x{\leq}y\ (\psi^{\geq}_{s}(\vec{a}, x, y) \wedge \xi_0(\vec{a}, x, y))$ for $y \geq t^{\geq}_{s}(\vec{a}) + t_\xi(\vec{a}, \tilde{s}(\vec{a}))$.

Proof. We will prove the least number principle for $\Delta_0(\{g_i\}_{i\in I}, f)$-formulas in $I\Delta_0(\{g_i\}_{i\in I}, F) + \mathsf{Mon}(\varphi)$ as this is equivalent to induction for $\Delta_0(\{g_i\}_{i\in I}, f)$-formulas. So, let $\theta(x, \vec{a})$ be a $\Delta_0(\{g_i\}_{i\in I}, f)$-formula and reason in $I\Delta_0(\{g_i\}_{i\in I}, F)$ $+\mathsf{Mon}(\varphi)$. By Lemma 10.1.7 we have a strict monotonic term $t_\theta(x, \vec{a})$ and a $\Delta_0(\{g_i\}_{i\in I}, F)$-formula $\theta_0(x, \vec{a}, y)$ such that $\theta(x, \vec{a}) \leftrightarrow \theta_0(x, \vec{a}, y)$ whenever $y \geq t_\theta(x, \vec{a})$.

Now assume $\exists\, x\; \theta(x, \vec{a})$. We will show that $\exists\, x\; (\theta(x, \vec{a}) \wedge \forall\, x'{<}x\; \neg\theta(x', \vec{a}))$. Let x be such that $\theta(x, \vec{a})$. We now fix some $y \geq t_\theta(x, \vec{a})$. Thus we have $\theta_0(x, \vec{a}, y)$. Applying the least number principle to the $\Delta_0(\{g_i\}_{i\in I}, F)$-formula $\theta_0(x, \vec{a}, y)$ we get a minimal x_0 such that $\theta_0(x_0, \vec{a}, y)$. As $x_0 < x$ and t_θ is monotone we have $y \geq t_\theta(x, \vec{a}) \geq t_\theta(x_0, \vec{a})$ and thus $\theta(x_0, \vec{a})$. If now $x' < x_0$ such that $\theta(x', \vec{a})$ then also $\theta_0(x', \vec{a}, y)$ which would conflict the minimality of x_0 for θ_0. Thus x_0 is the minimal element such that $\theta(x_0, \vec{a})$. ⊣

As in [Bek97] (Remark 5.14) we note here that Theorem 10.1.8 shows that $\Delta_0(\{g_i\}_{i\in I}, f)$-induction is actually provable from $\Delta_0(\{g_i\}_{i\in I}, F)$-induction where the bounding terms are just plain variables. Also we note that Lemma 10.1.6 and Lemma 10.1.7 do not use the full strength of $I\Delta_0(\{g_i\}_{i\in I}, F)$.

Corollary 10.1.9. *Let $I\Delta_0(\{g_i\}_{i\in I})$ be such that any term t in its language is provably majorizable by some other term $\tilde{t}$ that is strictly increasing in all of its arguments. Let f be $\Delta_0(\{g_i\}_{i\in I})$-definable by φ. Then, $I\Delta_0(\{g_i\}_{i\in I}) + \mathsf{Mon}(\varphi) \vdash I\Delta_0(\{g_i\}_{i\in I}, f)$.*

Proof. Immediate from Theorem 10.1.8 by replacing every occurrence of F by φ. ⊣

10.1.2 PRA, $I\Sigma_1$ and iterations of total functions

This subsection contains two main results. In Theorem 10.1.11 we shall characterize the difference between $I\Sigma_1$ and PRA in terms of provable closure of iteration of a certain class of functions.

In Theorem 10.1.12 we use this characterization to give a sufficient condition for a model of PRA to be also a model of $I\Sigma_1$.

Let us first specify what we mean by function iteration. If f denotes a function we will denote by f^{it} the (unique) function satisfying the following primitive recursive schema: $f^{\mathsf{it}}(0){=}1$, $f^{\mathsf{it}}(x+1){=}f(f^{\mathsf{it}}(x))$.

Definition 10.1.10. Let $\varphi(x, y)$ be some formula. By $\varphi^{\mathsf{it}}(x, y)$ we denote
$\exists\, \sigma\; \tilde{\varphi}^{\mathsf{it}}(\sigma, x, y)$ where $\tilde{\varphi}^{\mathsf{it}}(\sigma, x, y)$ is the formula
$\mathsf{Finseq}(\sigma) \wedge \mathsf{lh}(\sigma) = x + 1 \wedge \sigma_0 = 1 \wedge \sigma_x = y \wedge \forall\, i{<}x\; \varphi(\sigma_i, \sigma_{i+1})$.

Note that if PRA proves the functionality of a $\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$-formula φ, it also proves the functionality of $\tilde{\varphi}^{\mathsf{it}}$, for example by proving by induction on σ that $\forall\, \sigma\, \forall\, x, y, y', \sigma'{\leq}\sigma\; (\tilde{\varphi}^{\mathsf{it}}(\sigma, x, y) \wedge \tilde{\varphi}^{\mathsf{it}}(\sigma', x, y') \rightarrow \sigma = \sigma' \wedge y = y')$.

As we will need upperbounds on sequences of numbers a short remark on coding is due here. By $[a_0, \ldots, a_n]$ we will denote the code of the sequence $a_0, \ldots, a_n$ of natural numbers via some fixed coding technique. By

$[a_0, \dots, a_n] \sqcap [b_0, \dots, b_m]$ we will denote the code of the sequence $a_0, \dots, a_n, b_0, \dots, b_m$ that arises from concatenating $b_0, \dots, b_m$ to $a_0, \dots, a_n$ (to the right).

The projection functions are referred to by sub indexing. So, σ_i will be a_i if $\sigma = [a_0, \dots, a_n]$ and $i \leq n$ and zero if $i > n$, and $n+1$ is called the length of σ. We say that σ is an initial subsequence of σ' if $\sigma = [a_0, \dots a_n]$ and $\sigma' = [a_0, \dots a_n, \dots a_m]$ and $m \geq n$. We denote this by $\sigma \sqsubseteq \sigma'$.

Further, we shall employ well known expressions like $\mathsf{lh}(\sigma)$, giving the length of a sequence σ. If we write down statements involving sequences we will tacitly assume that the statements actually make sense. For example, $\forall\, i{<}\mathsf{lh}(x)\ \psi$ will thus actually denote $\mathsf{Finseq}(x) \wedge \forall\, i{<}\mathsf{lh}(x)\ \psi$.

We shall not fix any specific coding protocol as any protocol with elementary projections, concatenation etcetera is good for us.

The following theorem tells us what is the difference between PRA and $\mathrm{I}\Sigma_1$ in terms of totality statements of $\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$-definable functions.

Theorem 10.1.11. $\mathrm{I}\Sigma_1 \equiv \mathrm{PRA} + \{\mathsf{Tot}(\varphi) \to \mathsf{Tot}(\varphi^{\mathsf{it}}) \mid \varphi \in \Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})\}$.

Proof. For one inclusion we only need to show that $\mathrm{I}\Sigma_1 \vdash \mathsf{Tot}(\varphi) \to \mathsf{Tot}(\varphi^{\mathsf{it}})$ but this follows easily from a Σ_1-induction on x in $\exists\, \sigma\, \exists\, y\ \tilde{\varphi}^{\mathsf{it}}(\sigma, x, y)$ under the assumption that $\forall\, x\, \exists\, y\ \varphi(x, y)$. We shall thus concentrate on the harder direction $\mathrm{PRA} + \{\mathsf{Tot}(\varphi) \to \mathsf{Tot}(\varphi^{\mathsf{it}}) \mid \varphi \in \Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})\} \vdash \mathrm{I}\Sigma_1$.

To this end we reason in $\mathrm{PRA} + \{\mathsf{Tot}(\varphi) \to \mathsf{Tot}(\varphi^{\mathsf{it}}) \mid \varphi \in \Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})\}$ and assume $\exists\, y\ \psi(0, y) \wedge \forall\, x\ (\exists y\ \psi(x, y) \to \exists\, y\ \psi(x{+}1, y))$ for some $\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$-formula $\psi(x, y)$. Our aim is to obtain $\forall\, x\, \exists\, y\ \psi(x, y)$.

Let $\mathsf{Least}_{\psi,x}(y)$ denote the formula $\psi(x, y) \wedge \forall\, y'{<}y\ \neg\psi(x, y')$. We are going to define in a $\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$-way a formula $\varphi(x, y)$ so that $f^{\mathsf{it}}(x{+}1) = [y_0, \cdots, y_x]$ with $\forall\, i{\leq}x\ \mathsf{Least}_{\psi,i}(y_i)$.

$$\varphi(x,y) := \begin{cases} (i) & (x = 0 \wedge y = 0) \hfill \vee \\ (ii) & (x = 1 \wedge \exists\, y'{<}y\ (y = [y'] \wedge \mathsf{Least}_{\psi,0}(y'))) \hfill \vee \\ (iii) & (x > 1 \wedge \forall\, i{<}\mathsf{lh}(x)\ \mathsf{Least}_{\psi,i}(x_i) \wedge \\ & \quad \exists\, y'{<}y\ (y = x \sqcap [y'] \wedge \mathsf{Least}_{\psi,\mathsf{lh}(x)}(y'))) \hfill \vee \\ (iv) & (x > 1 \wedge \neg(\forall\, i{<}\mathsf{lh}(x)\ \mathsf{Least}_{\psi,i}(x_i)) \wedge y = 0) \end{cases}$$

Thus, the function f defined by φ has the following properties. It is always zero unless $x{=}1$ or x is of the form $[y_0, \cdots, y_n]$ where each y_i is the smallest witness for $\exists y\ \psi(i, y)$.

We note, that by our assumptions $\exists y\, \psi(0, y)$ and $\forall x\, (\exists y\, \psi(x, y) \to \exists y\, \psi(x{+}1, y))$, the function f is total. As the definition of φ is clearly $\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$ we may conclude $\mathsf{Tot}(f^{\mathsf{it}})$.

We shall show that f^{it} is $\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$-definable, and that provably $\mathsf{Mon}(f^{\mathsf{it}})$. If we know this, then our result follows immediately. Because, by an easy $\mathrm{I}\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega}, f^{\mathsf{it}})$-induction we conclude $\forall x\, \psi(x, (f^{\mathsf{it}}(x{+}1))_x)$, whence $\forall x\, \exists y\, \psi(x, y)$. By Corollary 10.1.9 we conclude $\mathrm{PRA}{+}\{\mathsf{Tot}(\varphi) \to \mathsf{Tot}(\varphi^{\mathsf{it}}) \mid \varphi \in \Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})\} \vdash \forall x\, \exists y\ \psi(x, y)$ and we are done.

We will first see inside our theory that $\mathsf{Mon}(f^{\mathsf{it}})$. The monotonicity of f^{it} is intuitively clear but we have to show that we can catch this intuition using only $\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$-induction.

For example, we can first prove by induction on x that all of the $f^{\mathsf{it}}(x{+}1)$ are 'good sequences' where by a good sequence we mean one of the form $[y_0, \ldots, y_x]$ with the y_i minimal witnesses to $\exists y\ \psi(i, y)$. To make this a $\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$-induction we should reformulate the statement as for example $\forall z\,\forall\,\sigma, x, y{\leq}z\ (\tilde{\varphi}^{\mathsf{it}}(\sigma, x+1, y) \rightarrow \mathsf{Goodseq(y)})$.

Now assume $\tilde{\varphi}^{\mathsf{it}}(\sigma', x', y')$. We will show by induction on x that

$$\forall\,x{\leq}x'\,\exists\,\sigma{\leq}\sigma'\,\exists\,y{\leq}y'\ \tilde{\varphi}^{\mathsf{it}}(\sigma, x'-x, y) \qquad (+)$$

from which monotonicity follows. If $x = 0$ we take $\sigma' = \sigma$ and $y = y'$. For the inductive step, let $\sigma \leq \sigma'$ and $y \leq y'$ be such that $\tilde{\varphi}^{\mathsf{it}}(\sigma, x'-x, y)$. We assume that $x + 1 \leq x'$ hence $\mathsf{lh}(\sigma) > 1$, for if not, the solution is trivial.

By σ_{-1} we denote the sequence that is obtained from σ by deleting the last element. Clearly $\tilde{\varphi}^{\mathsf{it}}(\sigma_{-1}, x'-x-1, (\sigma_{-1})_{x'-x-1})$ and $\varphi((\sigma_{-1})_{x'-x-1}, y)$. Thus $(\sigma_{-1})_{x'-x-1}$ is a good sequence which implies that clause (iii) in the definition of φ is used to determine y. Consequently $(\sigma_{-1})_{x'-x-1} \sqsubseteq y$ and thus $(\sigma_{-1})_{x'-x-1} \leq y \leq y'$. Moreover we note that $\sigma_{-1} \sqsubseteq \sigma$ and thus $\sigma_{-1} \leq \sigma \leq \sigma'$.

We now want to show the $\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$-ness of $\varphi^{\mathsf{it}}(x, y)$ by providing an upperbound on the σ in $\tilde{\varphi}^{\mathsf{it}}(\sigma, x, y)$. Under any reasonable choice of our coding machinery, we can find an $n \in \omega$ such that

$$\begin{array}{ll} (a) & [\overbrace{y, \cdots, y}^{x \text{ times}}] \leq \mathsf{Sup}_n(x+y), \\ (b) & \mathsf{Sup}_n(x+y)\sqcap[y] \leq \mathsf{Sup}_n(x+y+1). \end{array}$$

For such an n it is not hard to see that

$$\exists\sigma\ \tilde{\varphi}^{\mathsf{it}}(\sigma, x, y) \leftrightarrow \exists\,\sigma'{\leq}\mathsf{Sup}_n(x+y)\ \tilde{\varphi}^{\mathsf{it}}(\sigma, x, y).$$

This, we see by proving by induction on σ that

$$\forall\sigma\,\forall\,x, y{\leq}\sigma\ (\tilde{\varphi}^{\mathsf{it}}(\sigma, x, y) \rightarrow \exists\,\sigma'{\leq}\mathsf{Sup}_n(x+y)\ \tilde{\varphi}^{\mathsf{it}}(\sigma, x, y)).$$

We note that this is sufficient as $\tilde{\varphi}^{\mathsf{it}}(\sigma, x, y) \rightarrow x, y \leq \sigma$. The only interesting possibility in the induction step is when we get for some new $x+1, y$ that $\tilde{\varphi}^{\mathsf{it}}(\sigma{+}1, x{+}1, y)$. For $\sigma'' := (\sigma{+}1)_{-1}$ we have that $\sigma'' < \sigma{+}1$ and $\tilde{\varphi}^{\mathsf{it}}(\sigma'', x, y_{-1})$. By the induction hypothesis we may assume that $\sigma'' \leq \mathsf{Sup}_n(x + y_{-1})$. By the definition of $\tilde{\varphi}^{\mathsf{it}}$, we now see that $\tilde{\varphi}^{\mathsf{it}}(\sigma''\sqcap[y], x+1, y)$. But,

$$\begin{aligned} \sigma''\sqcap[y] &\leq \mathsf{Sup}_n(x + y_{-1})\sqcap[y] \\ &\leq \mathsf{Sup}_n(x+y)\sqcap[y] \\ &\leq \mathsf{Sup}_n(x+y+1). \end{aligned}$$

$\dashv$

We note that we filled the gap between PRA and IΣ_1 by transforming an admissible rule of PRA to axiom form. Indeed $\mathsf{Tot}(\varphi) \mathrel{|\!\sim} \mathsf{Tot}(\varphi^{\mathsf{it}})$ is an admissible rule of PRA. For if PRA $\vdash \mathsf{Tot}(\varphi)$, then f is a primitive recursive function as is well known. But f^{it} is constructed from f by a simple recursion. Thus f^{it} is primitive recursive and hence provably total in PRA. The same phenomenon occurs in passing from IΣ_1^R to IΣ_1 where the (trivially) admissible Σ_1-induction rule is added in axiom form to PRA to obtain IΣ_1.

The fact that we allow for variables in Theorem 10.1.11 is essential. For if not, the logical complexity of PRA + $\{\mathsf{Tot}(\varphi) \to \mathsf{Tot}(\varphi^{\mathsf{it}}) \mid \varphi{\in}\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})\}$ would be[6] Δ_3 and so would be the logical complexity of IΣ_1. But it is well known that IΣ_1 can not be proved by any consistent collection of Σ_3-sentences.

A parameter-free version of PRA + $\{\mathsf{Tot}(\varphi) \to \mathsf{Tot}(\varphi^{\mathsf{it}}) \mid \varphi{\in}\Delta_0^-(\{\mathsf{Sup}_i\}_{i\in\omega})\}$ will be equivalent to parameter-free Σ_1-induction, I$\Sigma_1{}^-$.

We now come to prove a theorem that tells us when a model of PRA is also a model of IΣ_1. This lemma is formulated in terms of majorizability behavior of some total functions. A total function of a model M is a relation $\varphi(x,y)$ (possibly with parameters from M) for which $M \models \mathsf{Tot}(\varphi)$. Often we will write $f \leq g$ as short for $\forall x\ (\exists y\ \varphi(x,y) \to \exists y'\ (\chi(x,y') \wedge y \leq y'))$ and say that f is majorized by g. Thus if $f \leq g$ we automatically have $\mathsf{Tot}(\varphi) \to \mathsf{Tot}(\chi)$.

Theorem 10.1.12. *Let $\mathcal{M}$ be a model of* PRA. *If every $\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$-definable total function (with parameters) of $\mathcal{M}$ is majorized by $m{+}\mathsf{Sup}_n$ for some $m \in \mathcal{M}$ and some $n \in \omega$, then $\mathcal{M}$ is also a model of* IΣ_1.

Proof. Let $\mathcal{M}$ be satisfying our conditions. To see that $\mathcal{M} \models$ IΣ_1 we need in the light of Theorem 10.1.11 to show that $\mathcal{M} \models \mathsf{Tot}(\varphi) \to \mathsf{Tot}(\varphi^{\mathsf{it}})$ for any $\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$ function φ with parameters in $\mathcal{M}$. So, we consider some function f such that $\mathcal{M} \models \mathsf{Tot}(\varphi)$. We choose $m \in \mathcal{M} \setminus \{0\}$ and $n \in \omega$ large enough so that

(*a*.) $\mathcal{M} \models f \leq m + \mathsf{Sup}_n$,

(*b*.) $\mathcal{M} \models \forall x\ (m + \mathsf{Sup}_{n+1}(mx + m + 1) \leq \mathsf{Sup}_{n+1}(mx + m + m))$.

The second condition is automatically satisfied if m is a non-standard element.

An easy $\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$-induction shows that $(m{+}\mathsf{Sup}_n)^{\mathsf{it}}(x) \leq \mathsf{Sup}_{n+1}(mx{+}m)$. (Remember that we have excluded $m = 0$.) The case $x = 0$ is trivial as

[6] Actually we should be more careful here as we work in a richer language. However this makes no essential difference as all the Sup_n are Δ_1-definable over PRA.

$1 \leq \mathsf{Sup}_{n+1}(m)$. For the inductive step we see that[7]

$$\begin{array}{ll}
(m+\mathsf{Sup}_n)^{\mathsf{it}}(x+1) & = \\
(m+\mathsf{Sup}_n)((m+\mathsf{Sup}_n)^{\mathsf{it}}(x)) & \leq_{\text{i.h.}} \\
m+\mathsf{Sup}_n(\mathsf{Sup}_{n+1}(mx+m)) & \leq_{\text{def.}} \\
m+\mathsf{Sup}_{n+1}(mx+m+1) & \leq_{(b.)} \\
\mathsf{Sup}_{n+1}(mx+m+m) = \mathsf{Sup}_{n+1}(m(x+1)+m). &
\end{array}$$

We can use the obtained bounds to show the totality of f^{it} by estimating the size of σ that witnesses $\tilde{\varphi}^{\mathsf{it}}(\sigma, x, y)$. We know (outside PRA) that σ is of the form

$$\begin{array}{ll}
[1, f(1), f(f(1)), \ldots, f^x(1)] & \leq \\
[1, m+\mathsf{Sup}_n(1), m+\mathsf{Sup}_n(f(1)), \ldots, m+\mathsf{Sup}_n(f^{x-1}(1))] & \leq \\
[1, m+\mathsf{Sup}_n(1), (m+\mathsf{Sup}_n)^2(1), \ldots, (m+\mathsf{Sup}_n)^2(f^{x-2}(1))] & \leq \\
\vdots & \vdots \\
[1, m+\mathsf{Sup}_n(1), (m+\mathsf{Sup}_n)^2(1), \ldots, (m+\mathsf{Sup}_n)^x(1)] & \leq \\
[(m+\mathsf{Sup}_n)^x(1), \ldots, (m+\mathsf{Sup}_n)^x(1)] & \leq \\
[\mathsf{Sup}_{n+1}(mx+m), \ldots, \mathsf{Sup}_{n+1}(mx+m)] &
\end{array}$$

Every time we used dots here in our informal argument, some $\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$-induction should actually be applied. To neatly formalize our reasoning we choose some $k \in \omega$ large enough for our n and m such that (in $\mathcal{M}$)

$(c.)$ $[1] \leq \mathsf{Sup}_{n+k}(2m)$

$(d.)$ $\mathsf{Sup}_{n+k}(m(x+1)+m) \sqcap [\mathsf{Sup}_{n+1}(m(x+1)+m)] \quad \leq$ $\mathsf{Sup}_{n+k}(m(x+2)+m)$[8]

With these choices for m, n and k it is easy to prove by $\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$-induction that

$$\forall x\, \exists \sigma{\leq}\mathsf{Sup}_{n+k}(m(x+1)+m)\, \exists y{\leq}\mathsf{Sup}_{n+1}(mx+m)\ \tilde{\varphi}^{\mathsf{it}}(\sigma, x, y).$$

If $x = 0$ then $\tilde{\varphi}^{\mathsf{it}}([1], 0, 1)$ and by $(c.)$ we have $[1] \leq \mathsf{Sup}_{n+k}(m(0+1)+m)$. Also $1 \leq \mathsf{Sup}_{n+1}(m)$. Now suppose $\tilde{\varphi}^{\mathsf{it}}(\sigma, x, y)$ with σ and y below their respective bounds. We have by the definition of $\tilde{\varphi}^{\mathsf{it}}$ that $\tilde{\varphi}^{\mathsf{it}}(\sigma \sqcap [f(y)], x+1, f(y))$ (again we do as if we had f available in our language). We need to show that the new values do not grow too fast. But,

$$\begin{array}{rlll}
f(y) & \leq_{\text{I.H.}} & f(\mathsf{Sup}_{n+1}(mx+m)) & \leq_{(a.)} \\
m+\mathsf{Sup}_n(\mathsf{Sup}_{n+1}(mx+m)) & \leq_{(b.)} & f(\mathsf{Sup}_{n+1}(m(x+1)+m)) &
\end{array}$$

[7] This looks like a legitimate induction but remember that $(m+\mathsf{Sup}_n)^{\mathsf{it}}$ has an a priori $\Sigma_1(\{\mathsf{Sup}_i\}_{i\in\omega})$-definition. The argument should thus be encapsulated in a $\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$-induction, for example by proving $\forall z\, \forall \sigma, x, y{\leq}z\ (\widetilde{(m+\mathsf{Sup}_n)}^{\mathsf{it}}(\sigma, x, y) \rightarrow y \leq \mathsf{Sup}_{n+1}(mx+m))$. The essential reasoning though boils down to the argument given here.

[8] It is not hard to convince oneself that under any reasonable coding protocol such a k does exist.

as we have seen before. By $(d.)$ we get that

$$\begin{array}{lll} \sigma \sqcap [f(y)] & \leq_{\text{I.H.}} & \mathsf{Sup}_{n+k}(m(x+1)+m) \sqcap [\mathsf{Sup}_{n+1}(m(x+1)+m)] \\ & \leq_{(d.)} & \mathsf{Sup}_{n+k}(m(x+2)+m). \end{array}$$

$\dashv$

10.1.3 The actual proof of Parsons' theorem

In the setting of this section we formulate Parsons' theorem as follows.

Theorem 10.1.13. $\forall \pi \in \Pi_2 \ (\mathrm{I}\Sigma_1 \vdash \pi \Rightarrow \mathrm{PRA} \vdash \pi)$

Before we give the proof of Parsons' theorem we first agree on some model theoretic notation.

We recall the definition of M' being a 1-elementary extension of M, denoted by $M \prec_1 M'$. This means that $M \subseteq M'$ and that for $\vec{m} \in M$ and $\sigma(\vec{y}) \in \Sigma_1$ we have $M \models \sigma(\vec{m}) \quad \Leftrightarrow M' \models \sigma(\vec{m})$. In this case we also say that M is a 1-elementary submodel of M'. It is easy to see that

$$M \prec_1 M' \ \Leftrightarrow \ [M \models \sigma(\vec{m}) \Rightarrow M' \models \sigma(\vec{m})] \ \text{ for all } \sigma(\vec{y}) \in \Sigma_2.$$

A 1-elementary chain is a sequence $M_0 \prec_1 M_1 \prec_1 M_2 \prec_1 \ldots$. It is well known that the union of a 1-elementary chain is a 1-elementary extension of every model in the chain. It is worthy to note that in a 1-elementary chain the truth of Σ_2-sentences (with parameters) is preserved from left to right and the truth of Π_2-sentences (without parameters) is preserved from right to left.

By $\mathsf{Th}(M, C)$ we denote the first-order theory of M with all constants from C added to the language. This makes sense if we know how to interpret the constants of C in M.
We also recall the definition of the collection principle.

$$\mathrm{B}\Gamma := \{\forall\, x{<}t\, \exists\, y\ \varphi(x,y) \rightarrow \exists\, s\, \forall\, x{<}t\, \exists\, y{<}s\ \varphi(x,y) \mid \varphi \in \Gamma\}$$

together with a minimum of arithmetical axioms, e.g. PA^-. We now come to the actual proof of Theorem 10.1.13.

Proof of Theorem 10.1.13. Let a countable model $M \models \mathrm{PRA}+\sigma$ be given with $\sigma \in \Sigma_2$. We will construct a countable model M' of $\mathrm{I}\Sigma_1 + \sigma$ using Theorem 10.1.12.

Our strategy will be to make any $\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$-definable total function of M that is not bounded by any of the $m+\mathsf{Sup}_n$ $(n \in \omega,\ m \in M)$ either bounded by some $m + \mathsf{Sup}_n$ $(n \in \omega,\ m \in M')$ or not total in the PRA-model M'. The model M' will be the union of a Σ_1-elementary chain of models $M = M_0 \prec_1 M_1 \prec_1 M_2 \ldots \prec_1 M' = \cup_{i\in\omega} M_i$.

At each stage either the boundedness of a total $\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$-definable function is guaranteed (a Π_1-sentence: $\forall\, x, y\ (\varphi(x,y) \rightarrow y \leq m + \mathsf{Sup}_n(x))$) or its non-totality (a Σ_2-sentence: $\exists x \forall y\ \neg\varphi(x,y)$). As we shall work with a 1-elementary chain of models, functions that are dealt with need no more attention further on in the chain. Their interesting properties, that is boundedness or non-totality, are stable. By choosing the order in which functions are dealt with in a good way, eventually all total funtions of all models M_i will be considered. We shall see that as a result of this process every total function in M' that is $\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$-definable is bounded by some $M + \mathsf{Sup}_n$.

To properly order the functions that we shall deal with, we fix a bijective pairing function in this proof satisfying $x, y \leq \langle x,y\rangle$. We do as if the models M_n were already defined and write $f_{n0}, f_{n1}, f_{n2}, \ldots$ for the list of the (countably many) total $\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$-definable functions of M_n. We emphasize that we allow the functions f_{ni} to contain parameters from M_n. Furthermore we define g_n to be f_{ab} for the unique $a, b \in \omega$ such that $\langle a,b\rangle = n$.

We define $M_0 := M$.

We will define M_{n+1} to be such that g_n becomes (or remains) either bounded or non-total in it and $M_n \prec_1 M_{n+1}$. If we can do so, we are done. For suppose $M = M_0 \models \mathrm{PRA} + \sigma$. As PRA is Π_1-axiomatizable in the language containing the $\{\mathsf{Sup}_i\}_{i\in\omega}$ we get that $M' \models \mathrm{PRA}$ and likewise $M' \models \sigma$.

If now $M' \models \mathsf{Tot}(\varphi)$ for some $\varphi \in \Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$, we see that for some n, $M_n \models \mathsf{Tot}(\varphi)$ as soon as M_n contains all the parameters that occur in φ. Thus $f = g_m$ for some $m \geq n$. Thus in M_{m+1} the function f will be surely majorized, for $M_{m+1} \models \neg\mathsf{Tot}(\varphi) \Rightarrow M' \models \neg\mathsf{Tot}(\varphi)$. Consequently $M' \models f \leq m' + \mathsf{Sup}_k$ for some $m' \in M_{m+1} \subseteq M'$, $k \in \omega$. By Theorem 10.1.12 we see that $M' \models \mathrm{I}\Sigma_1$.

If $M_n \models g_n \leq m + \mathsf{Sup}_k$ for some $m \in M_n$ and $k \in \omega$ we set $M_{n+1} := M_n$. Clearly $M_n \prec_1 M_{n+1}$ and g_n is bounded in M_{n+1} (regargless its totality).

So, suppose that g_n is total in M_n and that $M_n \models \neg(g_n \leq m + \mathsf{Sup}_k)$ for all $m \in M_n$ and all $k \in \omega$. We obtain our required model M_{n+1} in two steps.

Step 1.

We go from $M_n \prec_1 M_{n1} \models \mathsf{B}\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})(+\mathrm{PRA})$. To this purpose, we add a fresh constant d to our language and consider

$$T := \mathsf{Th}(M_n, \{m\}_{m\in M_n}) \cup \{d > \mathsf{Sup}_k(m) \mid k \in \omega,\ m \in M_n\}.$$

As T is finitely satisfiable in M_n, we can find a countable model $M_{n0} \models T$. Let M_{n1} be the (initial) submodel of M_{n0} with domain $\{x \in M_{n0} \mid \exists\, k{\in}\omega\, \exists\, m{\in}M_n\ x \leq \mathsf{Sup}_k(m)\}$. Clearly, M_{n1} is indeed a submodel, that is, it is closed under all the Sup_k. For if $x \leq \mathsf{Sup}_l(m)$ then $\mathsf{Sup}_k(x) \leq \mathsf{Sup}_k(\mathsf{Sup}_l(m)) \leq \mathsf{Sup}_{k+l+2}(m)$. We see that M_{n1} is a model of PRA as PRA is Π_1-axiomatized. As $M_n \subseteq M_{n1}$, we get $M_n \prec_1 M_{n1}$. For,

$$\begin{array}{lcll} M_n & \models & \exists x\, \varphi(x) \ \ , \varphi(x) \in \Pi_1 & \Rightarrow \text{for some } {\scriptstyle m \in M_n} \\ M_n & \models & \varphi(m) & \Rightarrow \\ M_{n0} & \models & \varphi(m) & \Rightarrow \\ M_{n1} & \models & \varphi(m) & \Rightarrow \\ M_{n1} & \models & \exists x\, \varphi(x). & \end{array}$$

We now see that $M_{n1} \models \mathsf{B}\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$. So, suppose $M_{n1} \models \forall\, x{<}t\, \exists\, y\, \varphi(x,y)$ for some $t \in M_{n1}$ and $\varphi \in \Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$. Clearly $M_{n0} \models \forall\, x{<}t\, \exists\, y{<}d\, \varphi(x,y)$ for some $d \in M_{n0}$, actually for any $d \in M_{n0} \setminus M_{n1}$. Now by the $\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$ minimal number principle we get a minimal d_0 such that $M_{n0} \models \forall\, x{<}t\, \exists\, y{<}d_0\, \varphi(x,y)$. If d_0 were in $M_{n0} \setminus M_{n1}$, then $d_0 - 1$ would also suffice as a bound on the y's. The minimality of d_0 thus imposes that $d_0 \in M_{n1}$. Consequently $M_{n1} \models \exists\, d_0\, \forall\, x{<}t\, \exists\, y{<}d_0\, \varphi(x,y)$ and $M_{n1} \models \mathsf{B}\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$.

Step 2.

We go from[9] $M_{n1} \models \mathsf{B}\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$(+PRA) to a model $M_{n1} \prec_1 M_{n3} \models$ PRA + $\neg\mathsf{Tot}(\chi_n)$. M_{n+1} will be the reduct of M_{n3} to the original language.

If $M_{n1} \models \neg\mathsf{Tot}(\chi_n)$ nothing has to be done and we take $M_{n3} = M_{n1}$. So, we assume that $M_{n1} \models \mathsf{Tot}(\chi_n)$. We consider the set

$$\Gamma := \mathsf{Th}(M_{n1}, \{m\}_{m\in M_{n1}}) \cup \{g_n(c) > m + \mathsf{Sup}_k(c) \mid m \in M_{n1},\ k \in \omega\}$$

with c a fresh constant symbol. As g_n is not majorizable in M_{n1} we see that any finite subset of Γ is satisfiable whence Γ is satisfiable. Let M_{n2} be a countable model of Γ. Of course, we can naturally embed M_{n1} in M_{n2}.

We will now see that $c > M_{n1}$. For suppose $c \leq m \in M_{n1}$. Then $M_{n1} \models \forall\, x{\leq}m\, \exists\, z\ \ g_n(x){=}z$.[10] By $\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$-collection we get $M_{n1} \models \exists\, d_0\, \forall\, x{\leq}m\, \exists\, z{\leq}d_0\ \ g_n(x){=}z$. But then $M_{n1} \models g_n(c) \leq d_0$ whence $M_{n1} \models \neg(g_n(c) > d_0 + \mathsf{Sup}_k(c))$. A contradiction.

Define M_{n3} to be the (initial) submodel of M_{n2} with domain $\{m \in M_{m2} \mid \exists\, k \in \omega \ \ M_{n2} \models m < \mathsf{Sup}_k(c)\}$. As $c \geq M_{n1}$ we get $M_{n1} \subseteq M_{n3}$. We now see that $M_{n1} \prec_1 M_{n3}$. For suppose $M_{n1} \models \exists\, x\ \varphi(x)$ with $\varphi(x) \in \Pi_1$ then $M_{n1} \models \varphi(m_0)$ for some $m_0 \in M_{n1}$. Consequently $M_{n2} \models \varphi(m_0)$ and as $M_{n3} \subset_e M_{n2}$ and $\varphi(m_0) \in \Pi_1$, also $M_{n3} \models \varphi(m_0)$ whence $M_{n3} \models \exists\, x\ \varphi(x)$. Clearly $M_{n3} \models \neg\mathsf{Tot}(\chi_n)$ as $g_n(c)$ can not have a value in M_{n3}. ⊣

Corollary 10.1.14. $\forall \pi \in \Pi_2\ (\mathrm{B}\Sigma_1 \vdash \pi \Rightarrow \mathrm{PRA} \vdash \pi)$

Proof. A direct proof of this fact is given in Step 1 in the above proof. ⊣

[9] Or from the reduct of M_{n1} to the original language for that matter.

[10] We actually should substitute the $\Delta_0(\{\mathsf{Sup}_i\}_{i\in\omega})$-graph of g_n here.

10.2 Cuts, consistency and total functions

In Section 9.3 an explicit $\mathrm{I}\Sigma_1$-cut J is exposed such that $\mathrm{I}\Sigma_1 \vdash \mathsf{Con}^J(\mathrm{PRA})$. In this section, we shall give an alternative proof of this fact. The interesting difference lies in the concept used in this proof. The cut that we used in Section 9.3 was defined in terms of truth predicates. The cut that we shall expose in this section will be defined in terms of totality statements of recursive functions.

The proof we present here is a simplification of an argument by Visser. In an unpublished note [Vis90b], Visser sketched a modification of a proof of Paris and Wilkie from [WP87] to obtain our Theorem 10.2.3. Lemma 8.10 from [WP87], implies that for every $r{\in}\omega$ there is an $(\mathrm{I}\Delta_0 + \mathsf{exp})$-cut such that for every $\sigma{\in}\Sigma_2$, $\mathrm{I}\Delta_0 + \sigma + \mathsf{exp}$ proves the consistency of $\mathrm{I}\Delta_0 + \sigma + \Omega_r$ on that cut.

10.2.1 Basic definitions

Let us first give a definition of PRA that is useful to us in our proof. Again, we will work with the functions $\mathsf{Sup}_n(x)$ as introduced in Section 10.1. However, this time we will not extend our language. Rather we shall work with arithmetical definitions of the $\mathsf{Sup}_n(x)$. Let us recall the defining equations for the functions $\mathsf{Sup}_n(x)$.

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

We see that $\mathsf{Sup}_z(x) = y$ can be expressed by a Σ_1-formula:[11]

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

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

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

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

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

[11] By close inspection of the defining formula we see that $\mathsf{Sup}_z(x){=}z$ can actually be regarded as a $\Delta_0(\exp)$-formula.

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

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

and so on.

Definition 10.2.1. PRA is the first-order theory in the language $\{+, \cdot, \leq, 0, 1\}$ using only the connectives $\neg, \rightarrow$ and $\forall$, with the following *non-logical axioms*.

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

[B.] Finitely many identity axioms of complexity Π_1.

[C.] For every $\varphi(x, \vec{a}){\in}\Delta_0$ an induction axiom of complexity Π_1 of the form:[12] $\forall x\ \forall z\ (\varphi(0, z) \wedge \forall\, y{<}x\ (\varphi(y, z) \rightarrow \varphi(y{+}1, z)) \rightarrow \varphi(x, z))$.

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

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

We shall need in our proof of Theorem 10.2.3 a formalization of a proof system that has the sub-formula property. Like Paris and Wilkie we shall use a notion of tableaux proofs rather than some sequent calculus. In our discussion below we consider theories T that are formulated using only the connectives $\rightarrow$, $\neg$ and $\forall$. The other connectives will still be used as abbreviations.

Definition 10.2.2. A *tableau proof of a contradiction* from a set of axioms T containing the identity axioms is a finite sequence $\Gamma_0, \Gamma_1, \ldots, \Gamma_r$ where the Γ_i satisfy the following conditions.

- For $0 \leq i \leq r$, Γ_i is a sequence of sequences of labeled formulas. The elements of Γ_i are denoted by Γ_i^j. The elements of the Γ_i^j are denoted by $\varphi_{i,j}^k(l)$ where l is the *label* of $\varphi_{i,j}^k$ and is either 0 or 1. In case $l = 1$ in $\varphi_{i,j}^k(l)$, we call $\varphi_{i,j}^k$ the *active* formula of both Γ_i^j and Γ_i. Only non-atomic formulas can be active.

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

- Γ_0 contains just one finite non-empty sequence of labeled formulas. We require $\varphi^k_{0,0}{\in}T$ for $k < \mathsf{lh}(\Gamma^0_0)$.
- In every Γ^j_r ($j < \mathsf{lh}(\Gamma_r)$) there is an atomic formula that also occurs negated in Γ^j_r.
- Every $0 \leq i < r$ contains exactly one sequence Γ^j_i with an active formula in it. This sequence in its turn contains exactly one active formula.
- For $0 \leq i < r$, we have $\mathsf{lh}(\Gamma_i) \leq \mathsf{lh}(\Gamma_{i+1}) \leq \mathsf{lh}(\Gamma_i) + 1$.
- For $0 \leq i < r$, we have $\mathsf{lh}(\Gamma^j_i) \leq \mathsf{lh}(\Gamma^j_{i+1}) \leq \mathsf{lh}(\Gamma^j_i) + 2$.
- For $0 \leq i < r$, we have $\varphi^k_{i,j} = \varphi^k_{i+1,j}$ for $k < \mathsf{lh}(\Gamma^j_i)$.
- $\mathsf{lh}(\Gamma^j_i) < \mathsf{lh}(\Gamma^j_{i+1})$ iff Γ^j_i contains the active formula of Γ_i. In this case, with $n = \mathsf{lh}(\Gamma^j_i)$ and $\varphi^m_{i,j}$ the active formula, one of the following holds.[13]

(β) $\varphi^m_{i,j}$ is of the form $\neg\neg\theta$ in which case $\Gamma^n_{i+1,j} = \theta$ and $\mathsf{lh}(\Gamma^j_{i+1}) = n+1$.

(γ) $\varphi^m_{i,j}$ is of the form $\theta_1 \rightarrow \theta_2$. In this case $\Gamma^n_{i+1,j} = \neg\theta_1$ and only in this case $\mathsf{lh}(\Gamma_{i+1}) = \mathsf{lh}(\Gamma_i) + 1$. Let $p := \mathsf{lh}(\Gamma_i)$. Γ^p_{i+1} is defined as follows: $\mathsf{lh}(\Gamma^p_{i+1}) = \mathsf{lh}(\Gamma^j_{i+1}) = n + 1$, $\Gamma^k_{i+1,p} = \Gamma^k_{i+1,j}$ for $k < n$ and $\Gamma^n_{i+1,p} = \theta_2$.

(δ) $\varphi^m_{i,j}$ is of the form $\neg(\theta_1 \rightarrow \theta_2)$. Only in this case $\mathsf{lh}(\Gamma^j_{i+1}) = \mathsf{lh}(\Gamma^j_i)+2$ and $\Gamma^n_{i+1,j} = \theta_1$ and $\Gamma^{n+1}_{i+1,j} = \neg\theta_2$.

(ϵ) $\varphi^m_{i,j}$ is of the form $\forall x\ \theta(x)$. In this case $\mathsf{lh}(\Gamma^j_{i+1}) = n+1$ and $\Gamma^n_{i+1,j} = \theta(t)$ for some term t that is freely substitutable for x in $\theta(x)$.

(ζ) $\varphi^m_{i,j}$ is of the form $\neg\forall x\ \theta(x)$. In this case $\mathsf{lh}(\Gamma^j_{i+1}) = n + 1$ and $\Gamma^n_{i+1,j} = \neg\theta(y)$ for some variable y that occurs in no formula of Γ^j_i.

It is well-known that φ is provable from T iff there is a tableau proof of a contradiction from $T \cup \{\neg\varphi\}$. The length of tableaux proofs can grow superexponentially larger than their regular counterparts. A pleasant feature of tableaux proofs is the sub-formula property.

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

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

10.2.2 $\mathrm{I}\Sigma_1$ proves the consistency of PRA on a cut

Theorem 10.2.3. *There exists an $\mathrm{I}\Sigma_1$-cut J such that for all $B{\in}\Sigma_2$ we have* $\mathrm{I}\Sigma_1 + B \vdash \mathsf{Con}^J(\mathrm{PRA} + B)$

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

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

The cut that does the job is the following:[14]

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

First we see that $J(z)$ indeed defines a cut in $\mathrm{I}\Sigma_1$. Obviously $\mathrm{I}\Sigma_1 \vdash J(0)$. We now see $\mathrm{I}\Sigma_1 \vdash J(z) \rightarrow J(z{+}1)$. For, reason in $\mathrm{I}\Sigma_1$ and suppose $J(z)$. In order to obtain $J(z{+}1)$ it is sufficient to show that $\forall x\, \exists y\, \mathsf{Sup}_{z+1}(x) = y$. This follows from an easy Σ_1-induction. As $B \in \Sigma_2$ we may assume that $B = \exists x\, A(x)$ with $A \in \Pi_1$.

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

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

- $\sigma_{0,0} = 0$.
- If Γ_i^j contains no active formula, $\sigma_{i+1,j} = \sigma_{i,j}$.
- If Γ_i^j contains an active formula one of (β)-(ζ) applies. Let $m{=}\mathsf{lh}(\Gamma_i^j)$.
  - (β) $\sigma_{i+1,j} = \sigma_{i,j}$.
  - (γ) $\sigma_{i+1,j} = \sigma_{i+1,m} = \sigma_{i,j}$.
  - (δ) $\sigma_{i+1,j} = \sigma_{i,j}$.
  - (ϵ) $\sigma_{i+1,j} = \sigma_{i,j}$.
  - (ζ) In this case essentially an existential quantifier is eliminated. We treat the three possible eliminations.[15]

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

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

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

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

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

Note that the statement is indeed Σ_1 as in $\mathrm{I}\Sigma_1$ we have the Σ_1-collection principle which tells us that the bounded universal quantifiers can be somehow pushed inside the unbounded existential quantifier of the Sat_{Σ_1}.

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

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

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

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

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

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

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

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

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

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

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

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

1. A subformula (modulo substitution of terms) of an axiom of PRA of the form $[A]$ or $[B]$,

2. A subformula (modulo substitution of terms) of an induction axiom $[C]$,

3. $\exists s\, \exists\, y{\leq}s\ \widetilde{\mathsf{Sup}}(s, \overline{z}, t, y)$ for some (possibly non-standard) term t and some $z{\in}J$.

Our strategy in all cases but 3. will be to show that[16]

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

This is sufficient as

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

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

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

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

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

and thus also

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

We immediately see that

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

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

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

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

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

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

or equivalently

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

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

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

For the inductive step we have to show

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

or equivalently that for arbitrary[17] z

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

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

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

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

and set out to prove

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

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

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

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

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

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

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

$\dashv$

Chapter 11

Modal logics with PRA and $\mathrm{I}\Sigma_1$

In this chapter we shall present two modal logics which, in a sense, fully describe what PRA and $\mathrm{I}\Sigma_1$ have to say about each other in terms of provability and interpretability. The questions that the logics can decide on are questions like $\mathrm{I}\Sigma_1 \vdash^? \mathsf{Con}(\mathrm{PRA})$, $\mathrm{PRA} + \mathsf{Con}(\mathrm{PRA}) \vdash^? \mathrm{I}\Sigma_1$, $\mathrm{PRA} + \mathsf{Con}(\mathrm{PRA}) \rhd^? \mathrm{PRA} + \mathsf{Con}(\mathrm{I}\Sigma_1) + \neg\mathrm{I}\Sigma_1$, $\mathrm{I}\Sigma_1 \rhd^? \mathrm{PRA} + \mathsf{Con}(\mathrm{PRA})$, $\mathrm{I}\Sigma_1 + \mathsf{Con}(\mathrm{I}\Sigma_1) \rhd^? \mathrm{PRA} + \mathsf{Con}(\mathsf{Con}(\mathrm{PRA}))$, etc. In this chapter, PRA shall denote $(\mathrm{EA})^2_\omega$ with the axiomatization as fixed in Section 9.1.

In Section 11.1 we shall first compute the closed fragment of the provability logic of PRA with a constant for $\mathrm{I}\Sigma_1$. The full provability logic of PRA with a constant for $\mathrm{I}\Sigma_1$ actually has already been determined in [Bek96]. We give an elementary proof here so that we can extend it when computing the closed fragment of the interpretability logic of PRA with a constant for $\mathrm{I}\Sigma_1$ in Section 11.2.

11.1 The logic PGL

Inductively we define F, the formulas of **PGL**.

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

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

Throughout this section we shall reserve $B, B_0, B_1, \ldots$ to denote boolean combinations of formulas of the form $\Box^n\bot$ with $n \in \omega + 1$.

Definition 11.1.1 (The logic PGL). The formulas of the logic **PGL** are given by F. The logic **PGL** is the smallest normal extension of **GL** in this language that contains the following two axiom schemes.

$$\begin{array}{ll}\mathsf{S}_1: & \Box(\mathtt{S} \to B) \to \Box B \\ \mathsf{S}_2: & \Box(\neg\mathtt{S} \to B) \to \Box B\end{array}$$

It is good to emphasize that **PGL** is a variable free logic. By our notational convention both in S_1 and in S_2 the B is a boolean combination of formulas of the form $\Box^n\bot$ with $n \in \omega$. Immediate consequences of S_1 and S_2 are that both $\Diamond(\mathtt{S} \wedge B)$ and $\Diamond(\neg\mathtt{S} \wedge B)$ are equivalent in **PGL** to $\Diamond B$.

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

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

Theorem 11.1.2. *For all sentences $A \in F$ we have*

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

Proof. The implication "$\Leftarrow$" is proved in the next subsection in Corollary 11.1.3 and Lemma 11.1.4. The other direction is proved in the Subsection after that, in Lemma 11.1.5. ⊣

11.1.1 Arithmetical Soundness of PGL

To see the arithmetical soundness of **PGL**, we only should check the validity of S_1 and S_2. Axiom S_1 can be seen as a direct consequence of the formalization of Parsons' theorem, Theorem 9.2.3 which can be formalized as soon as the totality of the superexponential function is provable.

Corollary 11.1.3. $\mathrm{PRA} \vdash \Box_{\mathrm{PRA}}(\mathrm{I}\Sigma_1 \to B) \to \Box_{\mathrm{PRA}}B$ *for $B \in \Pi_2$ and thus certainly whenever B is as in S_1.*

Lemma 11.1.4. $\mathrm{EA} \vdash \forall^{\Pi_3}B\ (\Box_{\mathrm{PRA}}(\neg\mathrm{I}\Sigma_1 \to B) \to \Box_{\mathrm{PRA}}B)$

Proof. Theorem 9.1.3 gives us that $\mathrm{I}\Sigma_n \vdash \mathrm{RFN}_{\Pi_{n+2}}(\mathrm{EA})$ ([Lei83]). Consequently, the formalization of $\mathrm{I}\Sigma_1 \vdash \mathrm{RFN}_{\Pi_3}(\mathrm{EA})$ is a true Σ_1-sentence and thus provable in EA. As $\mathrm{EA} \vdash \Box_{\mathrm{I}\Sigma_1}(\mathrm{RFN}_{\Pi_3}(\mathrm{EA}))$ we also have

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

Now we reason in EA, fix some $B{\in}\Pi_3$ and assume $\Box_{\mathrm{PRA}}(\neg\mathrm{I}\Sigma_1 \to B)$. We get

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

But, by simple propositional logic, we also have

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

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

We note that Lemma 11.1.4 actually holds for a wider class of formulas than just boolean combinations of $\Box^{\alpha}\bot$ formulas. For example $\neg(A \rhd B)$ is always Π_3. One can also isolate a set of sentences that is always Π_2 in PRA. (See Subsection 12.1.1.) When we study the logic **PIL** it will become clear why we only need to include these low-complexity instantiations of the above arithmetical facts in our axiomatic systems: In the closed fragment we have simple normal forms.

11.1.2 Arithmetical completeness of PGL

Lemma 11.1.5. *For all A in F we have that if* $\mathrm{PRA} \vdash A$ *then* $\mathbf{PGL} \vdash A$.

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

In Lemma 11.1.7 we will show that $\Box A$ is always equivalent in **PGL** to $\Box^{\alpha}\bot$ for some $\alpha \in \omega{+}1$. Then, in Lemma 11.1.8 we show that if $\mathbf{PGL} \vdash \Box A$ then $\mathbf{PGL} \vdash A$. So, if $\mathbf{PGL} \nvdash A$ then $\mathbf{PGL} \nvdash \Box A$. As $\mathbf{PGL} \vdash \Box A \leftrightarrow \Box^{\alpha}\bot$ for some $\alpha \in \omega$ (not $\omega{+}1$ as we assumed $\mathbf{PGL} \nvdash \Box A$!) and **PGL** is sound we also have $\mathrm{PRA} \vdash \Box A \leftrightarrow \Box^{\alpha}\bot$. Hence $\mathrm{PRA} \nvdash \Box A$ and also $\mathrm{PRA} \nvdash A$. ⊣

We work out the exercise in modal normal forms. Although this is already carried out in the literature (see e.g. Boolos [Boo93], or Visser [Vis92b]) we repeat it here to obtain some subsidiary information which we shall need later on.

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

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

So, if we say that some formula B is in conjunctive normal form we will in the sequel assume that B is written as $\bigwedge\!\!\!\bigwedge_i(\Box^{\alpha_i}\bot \rightarrow \Box^{\beta_i}\bot)$ with $\alpha_i > \beta_i$. The empty conjunction gives $\top$ and if we take $\alpha_0{=}\omega > 0{=}\beta_0$, we get with one conjunct just $\bot$.

Lemma 11.1.6. *If a formula B can be written in the form $\bigwedge_i(\Box^{\alpha_i}\bot \to \Box^{\beta_i}\bot)$ with $\alpha_i{>}\beta_i$, then we have that $\mathbf{PGL} \vdash \Box B \leftrightarrow \Box^{\beta+1}\bot$ where $\beta{=}\min(\{\beta_i\})$.*

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

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

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

Lemma 11.1.7. *For any formula A in F we have that A is equivalent in* **PGL** *to a boolean combination of formulas of the form $\mathtt{S}$ or $\Box^{\beta}\bot$. If, on top of that, A is of the form $\Box C$, then A is equivalent in* **PGL** *to $\Box^{\alpha}\bot$ for some $\alpha \in \omega + 1$.*

Proof. By induction on the complexity of formulas in F. The base cases are trivial. The only interesting case in the induction is where we consider the case that $A = \Box C$. Note that C, by induction being a boolean combination of $\Box^{\alpha}\bot$ formulas and $\mathtt{S}$, can be written as $(\mathtt{S} \to B_0) \wedge (\neg\mathtt{S} \to B_1)$. So, by Lemma 11.1.6 we have that for suitable indices β, β', β'':

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

⊣

Lemma 11.1.8. *If* **PGL** $\vdash \Box A$ *then* **PGL** $\vdash A$.

Proof. By Lemma 11.1.7 we can write A as a boolean combination of formulas of the form $\mathtt{S}$ or $\Box^{\beta}\bot$. Thus let $A \leftrightarrow (\mathtt{S} \to B_0) \wedge (\neg\mathtt{S} \to B_1)$ with B_0 and B_1 in conjunctive normal form and assume $\vdash \Box A$. For appropriate indices $\alpha_i{>}\beta_i$ and $\alpha'_j{>}\beta'_j$ we have $B_0 = \bigwedge_i(\Box^{\alpha_i}\bot \to \Box^{\beta_i}\bot)$ and $B_1 = \bigwedge_j(\Box^{\alpha'_j}\bot \to \Box^{\beta'_j}\bot)$. Using S_1, S_2 and Lemma 11.1.6 we get that $\Box A \leftrightarrow \Box^{\beta+1}\bot$ with $\beta = \min(\{\beta_i, \beta'_j\})$. By assumption $\beta = \omega$, thus all the β_i and β'_j were ω and hence $\vdash A$. ⊣

11.1.3 Modal Semantics for PGL, Decidability

In this subsection we will provide a modal semantics for **PGL**. Actually we will give a model $\mathcal{M}$ as depicted in Figure 11.1 which in some sense displays all there is to know about closed sentences with a constant for $\mathrm{I}\Sigma_1$ in **PGL**.

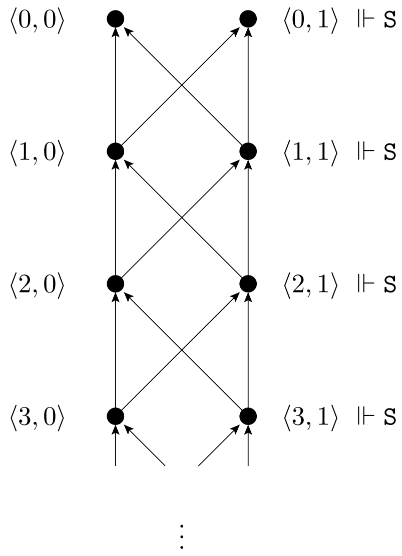

Figure 11.1: The model $\mathcal{M}$

Definition 11.1.9. We define the model $\mathcal{M}$ as follows, $\mathcal{M} := \langle M, R, \Vdash\rangle$. Here $M := \{\langle n,i\rangle \mid n \in \omega,\ i \in \{0,1\}\}$ and $\langle n,i\rangle R\langle m,j\rangle \Leftrightarrow m < n$. Furthermore $\langle n,i\rangle \Vdash \mathtt{S} \Leftrightarrow i = 1$.

Theorem 11.1.10. $\forall \mathbf{m}\ \mathcal{M}, \mathbf{m} \Vdash A \Leftrightarrow \mathbf{PGL} \vdash A$

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

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

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

$\Rightarrow$ Suppose **PGL** $\nvdash A$. By Lemma 11.1.8 **PGL** $\nvdash \Box A$, thus **PGL** $\vdash \Box A \leftrightarrow \Box^{\alpha}\bot$ for a certain $\alpha \in \omega$. By the first part of this proof we may conclude that $\mathbf{m} \Vdash \Box A \leftrightarrow \Box^{\alpha}\bot$ for any $\mathbf{m}$. As $\langle \alpha, i\rangle \nVdash \Box^{\alpha}\bot$, we automatically get $\langle \alpha, i\rangle \nVdash \Box A$. So, for some $\langle \beta, j\rangle$ with $\langle \alpha, i\rangle R\langle \beta, j\rangle$ we have $\langle \beta, j\rangle \Vdash \neg A$ showing the "non-validity" of A.

⊣

The set of theorems of **PGL** is clearly recursively enumerable. If a formula is not provable in **PGL**, then, by Theorem 11.1.10, in some node of the model $\mathcal{M}$ it is refuted. Thus the theoremhood of **PGL** is actually decidable.

11.2 The logic PIL

We shall now present the closed fragment of the interpretability logic of PRA with a constant for $\mathrm{I}\Sigma_1$. This section has some similarities with Visser's paper on exponentiation, [Vis92b].

In that paper the closed fragment of the interpretability logic of the arithmetical theory Ω is presented. (The theory Ω is also known as $\mathrm{I}\Delta_0 + \Omega_1$.) The modal language is enriched with an additional constant exp. The arithmetical translation of this constant is the Π_2-formula stating the totality of the exponential function.

A fundamental difference between Visser's [Vis92b] and this section is that although $\mathrm{I}\Sigma_1$ is a proper extension of PRA, no new recursive functions are proved to be total, as $\mathrm{I}\Sigma_1$ is a Π_2-conservative extension of PRA. In this sense the gap between PRA and $\mathrm{I}\Sigma_1$ is smaller than the gap between Ω and $\Omega + \mathsf{exp}$. This difference is also manifested in the corresponding logics already when we just constrain ourselves to provability. For example we have that

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

whereas

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

Actually, $\Omega + \mathsf{exp} + \mathsf{Con}(\Omega)$ does not even prove $\mathsf{Con}(\Omega + \mathsf{exp})$. It does hold however that $\Omega + \mathsf{Con}(\mathsf{Con}(\Omega)) \vdash \mathsf{Con}(\Omega + \mathsf{exp})$ and there are more similarities. We have that $\mathsf{Con}(\mathrm{PRA})$ is not provable in $\mathrm{I}\Sigma_1$. Similarly $\mathsf{Con}(\Omega)$ is not provable in $\Omega + \mathsf{exp}$. In turn $\mathrm{I}\Sigma_1$ is not provable in PRA together with any iteration of consistency statements and the same holds for exp and Ω.[1]

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

The differences between the pairs of theories is probably best reflected by the corresponding universal models. The interested reader is suggested to compare the universal models from this paper to the ones from [Vis92b].

[1] It is well known that $\mathrm{I}\Sigma_1 \equiv \mathrm{RFN}_{\Pi_3}(\mathrm{EA})$ (Theorem 9.1.3) and that $\mathrm{I}\Sigma_1$ is not contained in any Σ_3-extension of EA (Fact 9.1.5). Consistency statements are all Π_1-sentences. For the case of Ω and exp reason as follows. Take any non-standard model of true arithmetic together with the set $\{2^c{>}\omega_1^k(c) \mid k{\in}\omega\}$. Take the smallest set containing c being closed under the ω_1 function. Consider the initial segment generated by this set. This initial segment is a model of Ω and of all true Π_1 sentences but clearly not closed under exp.

Inductively we define I, the formulas of **PIL**.

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

Again, the constants of the language are $\bot, \top$ and S, and we will reserve the symbols $B, B_0, B_1, \ldots$ to denote boolean combinations of $\Box^\alpha \bot$ formulas. We will write $C \equiv D$ as short for $(C \rhd D) \wedge (D \rhd C)$ and we say that they are equi-interpretable.

Definition 11.2.1 (The logic PIL). The formulas of the logic **PIL** are given by I. The logic **PIL** is the smallest normal extension of **IL**W in this language that contains the following four axiom schemes.

$$\begin{array}{llll} \mathsf{S}_1 : & \Box(\mathsf{S} \to B) \to \Box B & & \\ \mathsf{S}_2 : & \Box(\neg\mathsf{S} \to B) \to \Box B & \mathsf{S}_3 : & \neg\mathsf{S} \wedge B \equiv B \\ \mathsf{S}_4 : & (B \rhd \mathsf{S} \wedge B) \to \Box\neg B & & \end{array}$$

It is good to stress that **PIL** is a variable free logic too. As the interpretability logic **IL**W is a part of **PIL** we have access to all known reasoning in **IL** and **IL**W. In this section, unless mentioned otherwise $\vdash$ refers to provability in **PIL**.

Fact 11.2.2.

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

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

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

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

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

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

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

Theorem 11.2.3. *For all sentences $A \in I$ we have* PRA $\vdash A \Leftrightarrow$ **PIL** $\vdash A$.

Proof. The implication "$\Leftarrow$" is proved in the next subsection in Lemma 11.2.4 and Lemma 11.2.5. The other direction is proved in the Subsection after that, in Lemma 11.2.8. ⊣

11.2.1 Arithmetical soundness of PIL

In [Vis91] it has been shown that **IL**W is sound for any reasonably formulated theory extending $\mathrm{I}\Delta_0 + \Omega_1$. So, to check for soundness of **PIL** with respect to PRA we only need to see that all translations of S_3 and S_4 are provable in PRA.

We shall give two soundness proofs for S_3 and S_4. The first proofs, consisting of Lemma 11.2.4 and 11.2.5 use finite approximations of theories. The second proofs make use of reflection principles and definable cuts. In accordance with Chapter 4, we could call the first proofs P-style, and the second, M-style soundness proofs.

Lemma 11.2.4. $\mathrm{PRA} \vdash B \rhd_{\mathrm{PRA}} B \wedge \neg\mathrm{I}\Sigma_1$ *for* $B {\in} \Sigma_2$*, so, certainly for* B *as in* S_3.

Proof. We want to show inside PRA that $\mathrm{PRA} + B \rhd \mathrm{PRA} + B + \neg\mathrm{I}\Sigma_1$. As we know that every finite Σ_2-extension of PRA is reflexive, we are by Orey-Hájek (Lemma 2.1.1) done if we can prove

$$\mathrm{PRA} \vdash \forall n\ \Box_{\mathrm{PRA}+B}(\Diamond_{\mathrm{PRA}[n]+B+\neg\mathrm{I}\Sigma_1}\top). \tag{11.1}$$

We will set out to prove that

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

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

from which 11.1 immediately follows.

The proof of (i) is just a slight modification of the proof of Lemma 11.1.4. We reason in EA and fix some n:

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

The proof of (ii) is just a formalization of the fact that every finite Σ_2-extension of PRA is reflexive. So, again we reason in EA. Recall that we have $\mathrm{PRA}[n] {=} (\mathrm{EA})^2_n$ in our axiomatization of PRA. Thus, by definition, $\Box_{\mathrm{PRA}[n+1]}(\Box_{\mathrm{PRA}[n]}\pi \to \pi)$ for $\pi {\in} \Pi_2$. Consequently, for our $\neg B {\in} \Pi_2$, we get $\Box_{\mathrm{PRA}[n+1]}(\Box_{\mathrm{PRA}[n]}\neg B \to \neg B)$.

Obviously we also have $\Box_{\mathrm{PRA}[n+1]+B}B$. Combining, we get a proof of (ii):

$$\begin{array}{lll}\Box_{\mathrm{PRA}[n+1]+B} & (& \Box_{\mathrm{PRA}[n]+B}\bot \\ & \rightarrow & \Box_{\mathrm{PRA}[n]}\neg B \\ & \rightarrow & \neg B \\ & \rightarrow & \bot\ ). \end{array}$$

⊣

Lemma 11.2.5. $\mathrm{PRA} \vdash B \rhd_{\mathrm{PRA}} B \wedge \mathrm{I}\Sigma_1 \rightarrow \Box_{\mathrm{PRA}}\neg B$ *for* $B{\in}\Sigma_2$*, so, certainly for* B *as in* $\mathsf{S_4}$

Proof. The theory $\mathrm{PRA}{+}B{+}\mathrm{I}\Sigma_1$ is, verifiably in PRA, equivalent to the finitely axiomatizable theory $\mathrm{I}\Sigma_1 + B$. Now we will reason in PRA.

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

$$\mathrm{PRA} + B + \mathrm{I}\Sigma_1 \rhd \mathrm{PRA}[k] + B + \mathsf{Con}(\mathrm{PRA}[k] + B).$$

Consequently,

$$\mathrm{PRA}[k] + B \rhd \mathrm{PRA}[k] + B + \mathsf{Con}(\mathrm{PRA}[k] + B)$$

and by Feferman's principle we get that $\Box_{\mathrm{PRA}[k]+B}\bot$. Thus $\Box_{\mathrm{PRA}+B}\bot$ and also $\Box_{\mathrm{PRA}}(B \rightarrow \bot)$, i.e., $\Box_{\mathrm{PRA}}\neg B$. ⊣

Lemma 11.2.5 certainly proves the correctness of axiom scheme $\mathsf{S_4}$. The proof also yields the following insights.

Corollary 11.2.6. *A consistent reflexive theory* U *does not interpret any finitely axiomatized theory extending it. In particular* PRA *does not interpret* $\mathrm{I}\Sigma_1$.

Corollary 11.2.7. $\mathrm{PRA} + \neg\mathrm{I}\Sigma_1$ *is not finitely axiomatizable.*

We now give alternative proofs of Lemma 11.2.4 and 11.2.5.

Second Proof of Lemma 11.2.4. We consider $B{\in}\Sigma_2$ and want to show in EA that $\mathrm{PRA} + B \rhd \mathrm{PRA} + B + \neg\mathrm{I}\Sigma_1$. We fix the $\mathrm{I}\Sigma_1$-cut J as given by Corollary 9.3.2 and reason in EA. Clearly

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

So, we are done if we can show that $\mathrm{PRA} + B + \mathrm{I}\Sigma_1 \rhd \mathrm{PRA} + B + \neg\mathrm{I}\Sigma_1$. By Corollary 9.3.2 we get that $\Box_{\mathrm{I}\Sigma_1+B}\mathsf{Con}^J(\mathrm{PRA} + B)$.

Using this cut J to relativize the identity translation, we find an interpretation that witnesses $\mathrm{I}\Sigma_1 + B \rhd \mathsf{S}^1_2 + \Diamond_{\mathrm{PRA}}B$. As $\mathsf{S}^1_2 + \Diamond_{\mathrm{PRA}}B$ is finitely axiomatizable, interpretability and smooth interpretability are in this case the same. We now get

$$\begin{array}{lll}
\mathrm{I}\Sigma_1 + B & \rhd & \\
\mathsf{S}^1_2 + \Diamond_{\mathrm{PRA}} B & \rhd & \text{by } \mathsf{W} \\
\mathsf{S}^1_2 + \Diamond_{\mathrm{PRA}} B + \Box_{\mathrm{I}\Sigma_1 + B}\bot & \rhd & \\
\mathsf{S}^1_2 + \Diamond_{\mathrm{PRA}} B + \Box_{\mathrm{PRA}}(B \rightarrow \neg \mathrm{I}\Sigma_1) & \rhd & \\
\mathsf{S}^1_2 + \Diamond_{\mathrm{PRA}} (B + \neg \mathrm{I}\Sigma_1) & \rhd & \\
\mathrm{PRA} + B + \neg \mathrm{I}\Sigma_1. & &
\end{array}$$

⊣

Second Proof of Lemma 11.2.5. We have $B{\in}\Sigma_2$ and assume in EA that $\mathrm{PRA}+B \rhd \mathrm{PRA}+B+\mathrm{I}\Sigma_1$. We have already seen in the above proof that $\mathrm{PRA}+B+\mathrm{I}\Sigma_1 \rhd \mathsf{S}^1_2 + \Diamond_{\mathrm{PRA}}B$.

Thus, by transitivity $\mathrm{PRA}+B \rhd \mathsf{S}^1_2 + \Diamond_{\mathrm{PRA}}B$, and

$$\begin{array}{lll}
\mathrm{PRA} + B & \rhd & \text{by } \mathsf{W} \\
\mathsf{S}^1_2 + \Diamond_{\mathrm{PRA}} B + \Box_{\mathrm{PRA}+B}\bot & \rhd & \\
\bot. & &
\end{array}$$

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

11.2.2 Arithmetical Completeness of PIL

This subsection is mainly dedicated to prove the next lemma.

Lemma 11.2.8. *For all A in I we have that if* $\mathrm{PRA} \vdash A$ *then* $\mathbf{PIL} \vdash A$*.*

Proof. The reasoning is completely analogous to that in the proof of Lemma 11.1.5. We thus need to prove a Lemma 11.2.15 stating that for any formula A in I we have that $\Box A$ is equivalent over **PIL** to a formula of the form $\Box^\alpha\bot$, and a Lemma 11.2.16 which tells us that $\mathbf{PIL} \vdash A$ whenever $\mathbf{PIL} \vdash \Box A$. ⊣

In a series of rather technical lemmas we will work up to the required lemmata. It is good to recall that in this chapter, B will always denote some boolean combination of formulas of the form $\Box^\alpha\bot$.

Lemma 11.2.9. $\mathbf{PIL} \vdash \mathtt{S} \wedge B \equiv (\mathtt{S} \wedge \Diamond^\beta\top) \vee \Diamond^{\beta+1}\top$ *for some $\beta \in \omega + 1$.*

Proof. $\mathtt{S} \wedge B \equiv (\mathtt{S} \wedge B) \vee \Diamond(\mathtt{S} \wedge B) \equiv \neg(\neg(\mathtt{S} \wedge B) \wedge \Box\neg(\mathtt{S} \wedge B))$, but $\neg(\mathtt{S} \wedge B) \wedge \Box\neg(\mathtt{S} \wedge B) \leftrightarrow (\mathtt{S} \rightarrow \neg B) \wedge \Box(\mathtt{S} \rightarrow \neg B) \leftrightarrow (\mathtt{S} \rightarrow \neg B) \wedge \Box\neg B$. Now we consider a conjunctive normal form of $\neg B$. Thus, $\neg B$ is equivalent to $\bigwedge\!\!\!\bigwedge_i(\Box^{\alpha_i}\bot \rightarrow \Box^{\beta_i}\bot)$ for certain $\alpha_i > \beta_i$ (possibly none). So, by Lemma 11.1.6, $\Box\neg B \leftrightarrow \bigwedge\!\!\!\bigwedge_i \Box^{\beta_i+1}\bot \leftrightarrow \Box^{\beta+1}\bot$ for $\beta = \min(\{\beta_i\})$. So,

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

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

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

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

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

$\dashv$

By a proof similar to that of Lemma 11.2.9 we get the following lemma.

Lemma 11.2.10. **PIL** $\vdash B \equiv \Diamond^{\gamma'}\top$ *for certain* $\gamma' \in \omega + 1$.

In **PIL** we have a substitution lemma in the sense that $\vdash F(C) \leftrightarrow F(D)$ whenever $\vdash C \leftrightarrow D$. We do not have a substitution lemma for equi-interpretable formulas[2] but we do have a restricted form of it.

Lemma 11.2.11. *If (provably in* **PIL***)* $C \equiv C'$, $D \equiv D'$, $E \equiv E'$ *and* $F \equiv F'$, *then* **PIL** $\vdash C \vee D \rhd E \vee F \leftrightarrow C' \vee D' \rhd E' \vee F'$.

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

Lemma 11.2.12. $\mathtt{S} \wedge \Diamond^{\alpha}\top \rhd (\mathtt{S} \wedge \Diamond^{\beta}\top) \vee \Diamond^{\gamma}\top$ *is provably equivalent in* **PIL** *to*

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

Proof. The case when $\alpha \geq \min(\{\beta,\gamma\})$ is trivial as $\Diamond^{\alpha}\top \to \Diamond^{\delta}\top$ whenever $\alpha \geq \delta$. So, we consider the case when $\neg(\alpha \geq \min(\{\beta,\gamma\}))$, that is, $\alpha < \beta, \gamma$.

Then we have $\Diamond^{\beta}\top \rhd \Diamond^{\alpha+1}\top \rhd \Diamond(\Diamond^{\alpha}\top) \rhd \Diamond(\mathtt{S} \wedge \Diamond^{\alpha}\top)$ and likewise for $\Diamond^{\gamma}\top$ in place of $\Diamond^{\beta}\top$. Thus, together with our assumption, we get $\mathtt{S} \wedge \Diamond^{\alpha}\top \rhd (\mathtt{S} \wedge \Diamond^{\beta}\top) \vee \Diamond^{\gamma}\top \rhd \Diamond(\mathtt{S} \wedge \Diamond^{\alpha}\top)$. By Feferman's principle we get $\Box\neg(\mathtt{S} \wedge \Diamond^{\alpha}\top)$ whence $\Box^{\alpha+1}\bot$. The implication in the other direction is immediate by Fact 11.2.2. $\dashv$

Lemma 11.2.13. $\Diamond^{\alpha}\top \rhd (\mathtt{S} \wedge \Diamond^{\beta}\top) \vee \Diamond^{\gamma}\top$ *is provably equivalent in* **PIL** *to*

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

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

Proof. The proof is completely analogous to that of Lemma 11.2.12 with the sole exception in the case that $\alpha = \beta < \gamma$. In this case

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

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

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

Lemma 11.2.14. *If C and D are both boolean combinations of $\mathtt{S}$ and sentences of the form $\Box^{\gamma}\bot$ then we have that* **PIL** $\vdash (C \rhd D) \leftrightarrow \Box^{\delta}\bot$ *for some $\delta \in \omega + 1$.*

Proof. So, let C and D meet the requirements of the lemma and reason in **PIL**. We get that

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

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

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

We will show that each conjunct of $(*)$ is equivalent to a formula of the form $\Box^{\epsilon}\bot$. Starting with the left conjunct we get by repeatedly applying Lemma 11.2.11 that

$$\begin{array}{lcl}
\mathtt{S} \wedge B_0 \rhd (\mathtt{S} \wedge B_2) \vee (\neg\mathtt{S} \wedge B_3) & \leftrightarrow & \text{Lemma 11.2.9} \\
(\mathtt{S} \wedge \Diamond^{\alpha}\top) \vee \Diamond^{\alpha+1}\top \rhd (\mathtt{S} \wedge B_2) \vee (\neg\mathtt{S} \wedge B_3) & \leftrightarrow & \mathsf{S_3} \\
(\mathtt{S} \wedge \Diamond^{\alpha}\top) \vee \Diamond^{\alpha+1}\top \rhd (\mathtt{S} \wedge B_2) \vee B_3 & \leftrightarrow & \text{Lemma 11.2.10} \\
(\mathtt{S} \wedge \Diamond^{\alpha}\top) \vee \Diamond^{\alpha+1}\top \rhd (\mathtt{S} \wedge B_2) \vee \Diamond^{\gamma'}\top & \leftrightarrow & \text{Lemma 11.2.9} \\
(\mathtt{S} \wedge \Diamond^{\alpha}\top) \vee \Diamond^{\alpha+1}\top \rhd (\mathtt{S} \wedge \Diamond^{\beta}\top) \vee \Diamond^{\beta+1}\top \vee \Diamond^{\gamma'}\top & \leftrightarrow & \\
(\mathtt{S} \wedge \Diamond^{\alpha}\top) \vee \Diamond^{\alpha+1}\top \rhd (\mathtt{S} \wedge \Diamond^{\beta}\top) \vee \Diamond^{\gamma}\top & \leftrightarrow & \\
(\mathtt{S} \wedge \Diamond^{\alpha}\top \rhd (\mathtt{S} \wedge \Diamond^{\beta}\top) \vee \Diamond^{\gamma}\top) \qquad \wedge & & \\
(\Diamond^{\alpha+1}\top \rhd (\mathtt{S} \wedge \Diamond^{\beta}\top) \vee \Diamond^{\gamma}\top) & \leftrightarrow & \text{Lemma 11.2.12} \\
\Box^{\mu}\bot \wedge (\Diamond^{\alpha+1}\top \rhd (\mathtt{S} \wedge \Diamond^{\beta}\top) \vee \Diamond^{\gamma}\top) & \leftrightarrow & \text{Lemma 11.2.13} \\
\Box^{\mu}\bot \wedge \Box^{\lambda}\bot & \leftrightarrow & \\
\Box^{\delta}\bot & &
\end{array}$$

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

Lemma 11.2.14 is the only new ingredient needed to prove the next two lemmas in complete analogy to their counterparts 11.1.7 and 11.1.8 in **PGL**.

Lemma 11.2.15. *For any formula A in I we have that A is equivalent in* **PIL** *to a boolean combination of formulas of the form $\mathtt{S}$ or $\Box^{\beta}\bot$. If, on top of that, A is of the form $\Box C$, then A is equivalent in* **PIL** *to $\Box^{\alpha}\bot$ for some $\alpha \in \omega + 1$.*

Lemma 11.2.16. *For all A in I we have that* **PIL** $\vdash A$ *whenever* **PIL** $\vdash \Box A$.

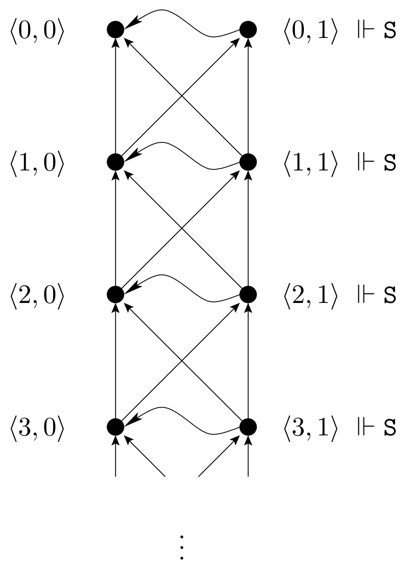
$\langle 0,0\rangle$
$\langle 0,1\rangle \ \Vdash \mathtt{S}$
$\langle 1,0\rangle$
$\langle 1,1\rangle \ \Vdash \mathtt{S}$
$\langle 2,0\rangle$
$\langle 2,1\rangle \ \Vdash \mathtt{S}$
$\langle 3,0\rangle$
$\langle 3,1\rangle \ \Vdash \mathtt{S}$
$\vdots$

Figure 11.2: The (simplified) model $\mathcal{N}$

11.2.3 Modal Semantics for PIL, Decidability

As in the case of **PGL**, we shall define a universal model for the logic **PIL**. **PIL**.

Definition 11.2.17 (Universal model for PIL). The Veltman model $\mathcal{N} = \langle M, R, \{S_m\}_{m\in M}, \Vdash\rangle$ is obtained from the model $\mathcal{M} = \langle M, R, \Vdash\rangle$ as defined in Definition 11.1.9 as follows. We define $\langle m,1\rangle S_{\mathbf{n}} \langle m,0\rangle$ for $\mathbf{n}R\langle m,1\rangle$ and close off as to have the $S_{\mathbf{n}}$ relations reflexive, transitive and containing R the amount it should.

Theorem 11.2.18. $\forall n \ \ \mathcal{N}, n \Vdash A \Leftrightarrow \mathbf{PIL} \vdash A$

Proof. The proof is completely analogous to that of Theorem 11.1.10. We only should check that all the instantiations of S_3 and S_4 hold in all the nodes of $\mathcal{N}$.

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

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

The modal semantics gives us the decidability of the logic **PIL**. In our case it is very easy to obtain a so-called simplified Veltman model. This is a model $\langle M, R, S, \Vdash\rangle$ where S now is a binary relation. Accordingly we define

$$x \Vdash A \rhd B \Leftrightarrow \forall y \ (xRy \Vdash A \Rightarrow \exists z \ (ySz \Vdash B)).$$

Our model model $\mathcal{N}$ is transformed into a simplified Veltman model by defining $\mathbf{n}S\mathbf{m} \Leftrightarrow \exists \mathbf{k}\ \mathbf{n}S_{\mathbf{k}}\mathbf{m}$. A perspicuous picture is readily drawn. The S-relation is depicted in Figure 11.2 by a wavy arrow.

11.2.4 Adding reflection

Just as always, if we want to go from all provable statements to all true statements, we have to only add reflection. As we are in the closed fragment and as we have good normal forms, this reflection only amounts to iterated consistency statements.

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

Theorem 11.2.19. **PGLS** $\vdash A \Leftrightarrow \mathbb{N} \models A$

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

To see the converse, we reason as follows. Consider $A \in F$ such that $\mathbb{N} \models A$. By Lemma 11.1.7 we can find an A' which is a boolean combination of $\mathtt{S}$ and $\Diamond^{\alpha}\top$ $(\alpha \in \omega + 1)$, such that **PGL** $\vdash A \leftrightarrow A'$. Thus PRA $\vdash A \leftrightarrow A'$ and also $\mathbb{N} \models A \leftrightarrow A'$. Consequently $\mathbb{N} \models A'$.

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

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

Theorem 11.2.20. **PILS** $\vdash A \Leftrightarrow \mathbb{N} \models A$

Proof. As the proof of Theorem 11.2.19 ⊣

Clearly, **PILS** is a decidable logic too.

Chapter 12

Remarks on **IL**(PRA)

In this chapter we shall study the interpretability logic of PRA. A modal characterization of **IL**(PRA) is still an open question. The best candidate so far is $\mathbf{IL}\mathsf{BR}^*$, where B is Beklemishev's principle. We shall study this principle in Section 12.1, where amongst others, a frame condition is given for B.

Sections 12.2 and 12.3 make some remarks on upperbounds for **IL**(PRA). Section 12.3 also has an interest independent from **IL**(PRA). We shall study the universal model for the closed fragment of **GLP**.

12.1 Beklemishev's principle

It is possible to write down a valid principle for the full interpretability logic of PRA. This was first done by Beklemishev (see [Vis97]). Beklemishev's principle B exploits the fact that any finite Σ_2-extension of PRA is reflexive, together with the fact that we have a good Orey-Hájek characterization for reflexive theories.

It turns out to be possible to define a class of modal formulae which are under any arithmetical realization provably Σ_2 in PRA. This are the so-called *essentially* Σ_2*-formulas*, we write ES_2. Let us start by defining this class and some related classes. In our definition, $\mathcal{A}$ will stand for the set of all modal interpretability formulae.

$$\begin{array}{lcl}
\mathsf{ED}_2 & := & \Box\mathcal{A} \mid \neg\mathsf{ED}_2 \mid \mathsf{ED}_2 \wedge \mathsf{ED}_2 \mid \mathsf{ED}_2 \vee \mathsf{ED}_2 \\
\mathsf{ES}_2 & := & \Box\mathcal{A} \mid \neg\Box\mathcal{A} \mid \mathsf{ES}_2 \wedge \mathsf{ES}_2 \mid \mathsf{ES}_2 \vee \mathsf{ES}_2 \mid \neg(\mathsf{ES}_2 \rhd \mathcal{A}) \\
\mathsf{EP}_2^{\mathsf{c}} & := & \Box\mathcal{A} \mid \Diamond\mathcal{A} \mid \mathsf{EP}_2^{\mathsf{c}} \vee \mathsf{EP}_2^{\mathsf{c}} \mid \mathsf{EP}_2^{\mathsf{c}} \wedge \mathsf{EP}_2^{\mathsf{c}} \mid \mathcal{A} \rhd \mathcal{A} \\
\mathsf{ES}_3 & := & \Box\mathcal{A} \mid \neg\Box\mathcal{A} \mid \mathcal{A} \rhd \mathcal{A} \mid \mathsf{ES}_3 \wedge \mathsf{ES}_3 \mid \mathsf{ES}_3 \vee \mathsf{ES}_3 \mid \neg(\mathsf{ES}_2 \rhd \mathcal{A}) \\
\mathsf{ES}_4 & := & \Box\mathcal{A} \mid \mathcal{A} \rhd \mathcal{A} \mid \neg\mathsf{ES}_4 \mid \mathsf{ES}_4 \wedge \mathsf{ES}_4 \mid \mathsf{ES}_4 \vee \mathsf{ES}_4 \mid \mathsf{ES}_4 \rightarrow \mathsf{ES}_4
\end{array}$$

We can now formulate Beklemishev's principle B.

$$\mathsf{B} := A \rhd B \rightarrow A \wedge \Box C \rhd B \wedge \Box C \qquad \text{for } A \in \mathsf{ES}_2$$

12.1.1 Arithmetical soundness of B

By Lemma 9.1.7 we know that $\mathrm{PRA} + \sigma$ is reflexive for any $\Sigma_2(\mathrm{PRA})$-sentence σ. Thus, we get by Orey-Hájek, Corollary 2.1.7, that

$$\mathrm{PRA} \vdash \sigma \rhd_{\mathrm{PRA}} \psi \leftrightarrow \forall x\ \Box_{\mathrm{PRA}}(\sigma \rightarrow \mathsf{Con}_x(\mathrm{PRA} + \psi)). \tag{12.1}$$

Consequently, for $\sigma \in \Sigma_2(\mathrm{PRA})$, $\neg(\sigma \rhd_{\mathrm{PRA}} \psi) \in \Sigma_2(\mathrm{PRA})$ and we see that, indeed, $\forall\, A{\in}\mathsf{ES}_2\, \forall *\ A^* \in \Sigma_2(\mathrm{PRA})$. We shall now see the arithmetical soundness of B.

Theorem 12.1.1. *For any formulas B and C we have that $\forall\, A{\in}\mathsf{ES}_2\, \forall * \mathrm{PRA} \vdash (A \rhd B \rightarrow A \wedge \Box C \rhd B \wedge \Box C)^*$.*

Proof. For some $A \in \mathsf{ES}_2$ and arbitrary B and C, we consider some realization $*$ and let $\alpha := A^*$, $\beta := B^*$ and $\gamma := C^*$. We reason in PRA and assume $\alpha \rhd_{\mathrm{PRA}} \beta$. As α is $\Sigma_2(\mathrm{PRA})$, we get by (12.1) that

$$\forall x\ \Box_{\mathrm{PRA}}(\alpha \rightarrow \mathsf{Con}_x(\mathrm{PRA} + \beta)). \tag{12.2}$$

We now consider n large enough (see Lemma 1.2.3 and Remark 1.2.4) such that

$$\Box_{\mathrm{PRA}}(\Box_{\mathrm{PRA}}\gamma \rightarrow \Box_{\mathrm{PRA},n}\Box_{\mathrm{PRA}}\gamma), \tag{12.3}$$

Combining (12.2) and (12.3), we see that for any x, (omitting the subscripts) $\Box(\alpha\wedge\Box\gamma \rightarrow \mathsf{Con}_x(\mathrm{PRA} + \beta \wedge \Box\gamma))$. Clearly, $\alpha\wedge\Box\gamma$ is still a $\Sigma_2(\mathrm{PRA})$-sentence.[1] Again by (12.1) we get $\alpha \wedge \Box\gamma \rhd \beta \wedge \Box\gamma$. ⊣

Let $\mathsf{M}^{\mathsf{ES_n}}$ be the schema $A \rhd B \rightarrow A \wedge \Box C \rhd B \wedge \Box C$ with $A \in \mathsf{ES_n}$.

Corollary 12.1.2. $\mathbf{IL}(\mathrm{I}\Sigma_{\mathrm{n}}^{\mathrm{R}}) \vdash \mathsf{M}^{\mathsf{ES_{n+1}}}$ *for $n = 1, 2, 3$.*

Proof. For $n = 1$ this is just Theorem 12.1.1. The proof can easily be generalized for $n = 2$ and $n = 3$ using Theorem 9.1.4 and realizing that any Σ_{n+1}-extension of $\mathrm{I}\Sigma_n^R$ is reflexive. ⊣

12.1.2 A frame condition

Let us first fix some notation. If $\mathcal{C}$ is a finite set, we write $xR\mathcal{C}$ as short for $\bigwedge_{c\in\mathcal{C}} xRc$. Similar conventions hold for the other relations. The A-critical cone of x, $\mathcal{C}_x^A$ is in this section defined as $\mathcal{C}_x^A := \{y \mid xRy \wedge \forall z\ (yS_xz \rightarrow z \nVdash A)\}$.

We define $xR^*y :\Leftrightarrow y = x \vee xRy$. By $x{\uparrow}$ we denote the set of worlds that lie above x w.r.t. the R relation. That is, $x{\uparrow} := \{y \mid xRy\}$. With $yS_x{\uparrow}$ we denote the set of those z for which yS_xz.

We will consider frames both as modal models without a valuation and as structures for first- (or sometimes second) order logic. We say that a model M is based on a frame F if F is precisely M with the $\Vdash$ relation left away. From now on we will write $A \equiv B$ instead of $(A \rhd B) \wedge (B \rhd A)$.

[1] If we use Lemma 2.1.1, this observation is not necessary.

In this subsection we give the frame condition of Beklemishev's principle. Our frame condition holds on the class of finite frames. At first sight, the condition might seem a bit awkward. On second sight it is just the frame condition of M with some simulation built in. First we approximate the class $\mathsf{ES_2}$ by stages.

Definition 12.1.3.

$$\begin{array}{lll} \mathsf{ES_2^0} & := & \mathsf{ED_2} \\ \mathsf{ES_2^{n+1}} & := & \mathsf{ES_2^n} \mid \mathsf{ES_2^{n+1}} \wedge \mathsf{ES_2^{n+1}} \mid \mathsf{ES_2^{n+1}} \vee \mathsf{ES_2^{n+1}} \mid \neg(\mathsf{ES_2^n} \rhd \mathsf{Form}) \end{array}$$

It is clear that $\mathsf{ES_2} = \cup_i \mathsf{ES_2^i}$. We now define some first order formulas $\mathcal{S}_i(b,u)$ that say that two nodes in a frame b and u look a like. The larger i is, the more the two points look alike. We use the letter $\mathcal{S}$ as to hint at a simulation. For $i \geq 1$ the relation $\mathcal{S}_i(b,u)$ is in general not symmetric.

Definition 12.1.4.

$$\begin{array}{lll} \mathcal{S}_0(b,u) & := & b{\uparrow}{=}u{\uparrow} \\ \mathcal{S}_{n+1}(b,u) & := & \mathcal{S}_n(b,u) \wedge \\ & & \forall c\ (bRc \rightarrow \exists c'\ (uRc' \wedge \mathcal{S}_n(c,c') \wedge cS_bc' \wedge c'S_u{\uparrow} \subseteq cS_b{\uparrow})) \end{array}$$

By induction on n we easily see that $\forall n\ F \models \mathcal{S}_n(b,b)$ for all frames F and all $b{\in}F$.

Lemma 12.1.5. *Let F be a model. For all n we have the following. If $F \models \mathcal{S}_n(b,u)$ then $b \Vdash A \Rightarrow u \Vdash A$ for all $A{\in}\mathsf{ES_2^n}$.*

Proof. We proceed by induction on n. If $n{=}0$, $A{\in}\mathsf{ES_2^0}$ can be written as $\bigvee_i(\Box A_i \wedge \bigwedge_j \Diamond A_{ij})$. Clearly, if $b{\uparrow} = u{\uparrow}$ then $b \Vdash A \Rightarrow u \Vdash A$.

Now consider $A{\in}\mathsf{ES_2^{n+1}}$ and b and u such that $F \models \mathcal{S}_{n+1}(b,u)$. We can write

$$A = \bigvee_i (A_{i0} \wedge \bigwedge_{j\neq 0} \neg(A_{ij} \rhd B_{ij})),$$

with A_{ij} in $\mathsf{ES_2^n}$. If $b \Vdash A$, then for some i, $b \Vdash A_{i0} \wedge \bigwedge_{j\neq 0} \neg(A_{ij} \rhd B_{ij})$. As $\mathcal{S}_{n+1}(b,u) \rightarrow \mathcal{S}_n(b,u)$, and by the induction hypothesis we see that $u \Vdash A_{i0}$. So, we only need to see that $u \Vdash \neg(A_{ij} \rhd B_{ij})$ for $j{\neq}0$. As $b \Vdash \neg(A_{ij} \rhd B_{ij})$, for some $c{\in}\mathcal{C}_b^{B_{ij}}$ we have $c \Vdash A_{ij}$. By $\mathcal{S}_{n+1}(b,u)$ we find a c' such that uRc', cS_bc', and $c'S_u{\uparrow} \subseteq cS_b{\uparrow}$. This guarantees that $c'{\in}\mathcal{C}_u^{B_{ij}}$. Moreover we know that $\mathcal{S}_n(c,c')$, thus by the induction hypothesis, as $c \Vdash A_{ij}$, we get that $c' \Vdash A_{ij}$. Consequently $u \Vdash \neg(A_{ij} \rhd B_{ij})$.

⊣

Lemma 12.1.6. *Let F be a finite frame. For all i, and any $b{\in}F$, there is a valuation V_i^b on F and a formula $A_i^b{\in}\mathsf{ES_2^i}$ such that $F \models \mathcal{S}_i(b,u) \Leftrightarrow u \Vdash A_i^b$.*

Proof. The proof proceeds by induction on i. First consider the basis case, that is, $i{=}0$. Let $b{\uparrow}$ be given by the finite set $\{x_j\}_{j\in J}$. We define

$$\begin{array}{lll} y \Vdash p_j & \leftrightarrow & y{=}x_j \\ y \Vdash r & \leftrightarrow & bRy \end{array}$$

Let A^b_0 be $\Box r \wedge \bigwedge_j \Diamond p_j$. It is now obvious that $u \Vdash A_0 \Leftrightarrow u{\uparrow}{=}b{\uparrow}$.

For the inductive step, we fix some b and reason as follows. First, let V^b_i and A^b_i be given by the induction hypothesis such that $u \Vdash A^b_i \Leftrightarrow F \models \mathcal{S}_i(b,u)$. We do not specify the variables in A_i but we suppose they do not coincide with any of the ones mentioned below. Let $b{\uparrow} = \{x_j\}_{j \in J}$. The induction hypothesis gives us sentences A^j_i (no sharing of variables) and valuations V^j_i such that $F, u \Vdash A^j_i \Leftrightarrow F \models \mathcal{S}_i(x_j, u)$.

Let $\{q_j\}_{j \in J}$ be a set of fresh variables. V^b_{i+1} will be V^b_i and V^j_i on the old variables. For the $\{q_j\}_{j \in J}$ we define V^b_{i+1} to act as follows:

$$y \Vdash q_j \Leftrightarrow y \notin x_j S_b {\uparrow}.$$

Moreover we define

$$A^b_{i+1} := A^b_i \wedge \bigwedge_j \neg (A^j_i \rhd q_j).$$

Now we will see that under the new valuation V^b_{i+1},

(*i*) $u \Vdash A^b_{i+1} \Rightarrow F \models \mathcal{S}_{i+1}(b,u)$,

(*ii*) $F \models \mathcal{S}_{i+1}(b,u) \Rightarrow u \Vdash A^b_{i+1}$.

For (*i*) we reason as follows. Suppose $u \Vdash A^b_{i+1}$. Then also $u \Vdash A^b_i$ and thus $F \models \mathcal{S}_i(b,u)$. It remains to show that

$$F \models \forall c\, (bRc \to \exists c'\ (uRc' \wedge \mathcal{S}_i(c,c') \wedge cS_bc' \wedge c'S_u{\uparrow} \subseteq cS_b{\uparrow})).$$

To this purpose we consider and fix some x_j in $b{\uparrow}$. As $u \Vdash A^b_{i+1}$, we get that $u \Vdash \neg(A^j_i \rhd q_j)$. Thus, for some $c' {\in} \mathcal{C}^{q_j}_u$, $c' \Vdash A^j_i$. Clearly $c' \Vdash \neg q_j$ whence $x_j S_b c'$. Also $\forall t\ (c' S_u y \Rightarrow y \Vdash \neg q_j)$ which, by the definition of V^b_{i+1} translates to $c' S_u {\uparrow} \subseteq x_j S_b {\uparrow}$. Clearly also uRc'. By $c' \Vdash A^j_i$ and the induction hypothesis we get that $\mathcal{S}_i(x_j, c')$. Indeed we see that $F \models \mathcal{S}_{i+1}(b,u)$.

For (*ii*) we reason as follows. As $F \models \mathcal{S}_{i+1}(b,u)$, also $F \models \mathcal{S}_i(b,u)$ and by the induction hypothesis, $u \Vdash A^b_i$. It remains to show that $u \Vdash \neg(A^j_i \rhd q_j)$ for any j. So, let us fix some j. Then, by the second part of the $\mathcal{S}_{i+1}$ requirement we find a c' such that

$$uRc' \wedge \mathcal{S}_i(x_j, c') \wedge x_j S_b c' \wedge c' S_u {\uparrow} \subseteq x_j S_b {\uparrow}.$$

Now, $uRc' \wedge x_j S_b c' \wedge c' S_u {\uparrow} \subseteq x_j S_b {\uparrow}$ gives us that $c' {\in} \mathcal{C}^{q_j}_u$. By $\mathcal{S}_i(x_j, c')$ and the induction hypothesis we get that $c' \Vdash A^j_i$. Thus indeed $u \Vdash \neg(A^j_i \rhd q_j)$. $\dashv$

Notice that in the proof of this lemma, we have only used conjunctions to construct the formulas A^b_i.

Definition 12.1.7. For every i we define the frame condition $\mathcal{C}_i$ to be

$$\forall\, a, b\ (aRb \to \exists u\ (bS_a u \wedge \mathcal{S}_i(b,u) \wedge \forall\, d, e\ (uS_a dRe \to bRe))).$$

Definition 12.1.8. Let F be a finite frame. For all i, we have that

$$\begin{gathered}\text{for all } A{\in}\mathsf{ES}_2^{\mathsf{i}},\ F \models A \rhd B \to A \wedge \Box C \rhd B \wedge \Box C,\\ \text{if and only if}\\ F \models \mathcal{C}_i.\end{gathered}$$

Proof. First suppose that $F \models \mathcal{C}_i$ and that $a \Vdash A \rhd B$ for some $A{\in}\mathsf{ES}_2^{\mathsf{i}}$ and some valuation on F. We will show that $a \Vdash A \wedge \Box C \rhd B \wedge \Box C$ for any C. Consider therefore some b with aRb and $b \Vdash A \wedge \Box C$. The $\mathcal{C}_i$ condition provides us with a u such that

$$bS_a u \wedge \mathcal{S}_i(b,u) \wedge \forall\, d,e\ (uS_a dRe \to bRe) \quad (*)$$

As $F \models \mathcal{S}_i(b,u)$, we get by Lemma 12.1.5 that $u \Vdash A$. Thus, as aRu and $a \Vdash A \rhd B$, we know that there is some d with $uS_a d$ and $d \Vdash B$. If now dRe, by $(*)$, also bRe and hence $e \Vdash C$. Thus, $d \Vdash B \wedge \Box C$. Clearly $bS_a d$ and thus $a \Vdash A \wedge \Box C \rhd B \wedge \Box C$.

For the opposite direction we reason as follows. Suppose that $F \not\models \mathcal{C}_i$. Thus, we can find a, b with

$$aRb \wedge \forall u\ (bS_a u \wedge \mathcal{S}_i(b,u) \to \exists\, d,e\ (uS_a dRe \wedge \neg bRe)) \quad (**).$$

By Lemma 12.1.6 we can find a valuation V_i^b and a sentence $A_i^b{\in}\mathsf{ES}_2^{\mathsf{i}}$ such that $u \Vdash A_i^b \Leftrightarrow F \models \mathcal{S}_i(b,u)$. Let q and s be fresh variables. Moreover, let $\mathcal{D}$ be the following set.

$$\mathcal{D} := \{d{\in}F \mid bS_a dRe \wedge \neg bRe \text{ for some } e\ \}.$$

We define a valuation V that is an extension of V_i^b by stipulating that

$$\begin{array}{lcl} y \Vdash q & \leftrightarrow & (y{\in}\mathcal{D}) \vee \neg(bS_a y),\\ y \Vdash s & \leftrightarrow & bRy. \end{array}$$

We now see that

(i) $a \Vdash A_i^b \rhd q$,

(ii) $a \Vdash \neg(A_i^b \wedge \Box s \rhd q \wedge \Box s)$.

For (i) we reason as follows. Suppose that aRb' and $b' \Vdash A_i^b$. If $\neg(bS_a b')$, $b' \Vdash q$ and we are done. So, we consider the case in which $bS_a b'$. As $\mathcal{S}_i(b,b')$, $(**)$ now yields us a $d{\in}\mathcal{D}$ such that $b'S_a d$. Clearly $bS_a d$ and thus, by definition, $d \Vdash q$.

To see (ii) we notice that $b \Vdash A_i^b \wedge \Box s$. But if $bS_a y$ and $y \Vdash q$, by definition $y{\in}\mathcal{D}$ and thus $y \Vdash \neg\Box s$. Thus $b{\in}\mathcal{C}_a^{q\wedge\Box s}$ and $a \Vdash \neg(A_i \wedge \Box s \rhd q \wedge \Box s)$. ⊣

The following theorem is now an immediate corollary of the above reasoning.

Theorem 12.1.9. *A finite frame F validates all instances of Beklemishev's principle if and only if* $\forall i\ F \models \mathcal{C}_i$.

Definition 12.1.10. Let $\mathsf{B_i}$ be the principle $A \rhd B \to A \wedge \Box C \rhd B \wedge \Box C$ for $A \in \mathsf{ES}_2^{\mathsf{i}}$.

Corollary 12.1.11. *For a finite frame we have $F \models \mathsf{B_i} \Leftrightarrow F \models \mathcal{C}_i$.*

For the class of finite frames, we can get rid of the universal quantification in the frame condition of Beklemishev's principle. Remember that $\mathsf{depth}(x)$, the depth of a point x, is the length of the longest chain of R-successors starting in x.

Lemma 12.1.12. *If $\mathcal{S}_n(x,x')$, then $\mathsf{depth}(x) = \mathsf{depth}(x')$.*

Proof. $\mathcal{S}_n(x,x') \Rightarrow \mathcal{S}_0(x,x') \Rightarrow x{\uparrow} = x'{\uparrow}$. ⊣

Lemma 12.1.13. *If $\mathcal{S}_n(x,x')$ & $\mathsf{depth}(x) \leq n$, then $\mathcal{S}_m(x,x')$ for all m.*

Proof. The proof goes by induction on n. For $n = 0$, the result is clear. So, we consider some x, x' with $\mathcal{S}_{n+1}(x,x')$ & $\mathsf{depth}(x) \leq n+1$. We are done if we can show $\mathcal{S}_{m+1}(x,x')$ for $m \geq n+1$.

This, we prove by a subsidiary induction on m. The basis is trivial. For the inductive step, we assume $\mathcal{S}_m(x,x')$ for some $m \geq n+1$ and set out to prove $\mathcal{S}_{m+1}(x,x')$, that is

$$\mathcal{S}_m(x,x') \wedge \forall y\ (xRy \to \exists y'\ (yS_xy' \wedge \mathcal{S}_m(y,y') \wedge y'S_{x'}{\uparrow} \subseteq yS_x{\uparrow}))$$

The first conjunct is precisely the induction hypothesis. For the second conjunct we reason as follows. As $m \geq n+1$, certainly $\mathcal{S}_{n+1}(x,x')$. We consider y with xRy. By $\mathcal{S}_{n+1}(x,x')$, we find a y' with

$$yS_xy' \wedge \mathcal{S}_n(y,y') \wedge y'S_{x'}{\uparrow} \subseteq yS_x{\uparrow}.$$

As xRy and $\mathsf{depth}(x) \leq n+1$, we see $\mathsf{depth}(y) \leq n$. Hence by the main induction, we get that $\mathcal{S}_m(y,y')$ and we are done. ⊣

Definition 12.1.14. A B-simulation on a frame is a binary relation $\mathcal{S}$ for which the following holds.

1. $\mathcal{S}(x,x') \to x{\uparrow} = x'{\uparrow}$
2. $\mathcal{S}(x,x')$ & $xRy \to \exists y'(yS_xy' \wedge \mathcal{S}(y,y') \wedge y'S_{x'}{\uparrow} \subseteq yS_x{\uparrow})$

If F is a finite frame that satisfies $\mathcal{C}_i$ for all i, we can consider $\bigcap_{i\in\omega} \mathcal{S}_i$. This will certainly be a B-simulation.

Definition 12.1.15. The frame condition $\mathcal{C}_{\mathsf{B}}$ is defined as follows. $F \models \mathcal{C}_{\mathsf{B}}$ if and only if there is a B-simulation $\mathcal{S}$ on F such that for all x and y,

$$xRy \to \exists y'(yS_xy' \wedge \mathcal{S}(y,y') \wedge \forall d,e\ (y'S_xdRe \to yRd)).$$

An immediate consequence of Lemma 12.1.13 is the following theorem.

Theorem 12.1.16. *For F a finite frame, we have*

$$F \models \mathsf{B} \quad \Leftrightarrow \quad F \models \mathcal{C}_{\mathsf{B}}.$$

Note that the M-frame condition can be seen as a special case of the frame condition of B: we demand that $\mathcal{S}$ be the identity relation.

It is not hard to see that the frame condition of M_0 follows from $\mathcal{C}_0$. And indeed, $\mathbf{IL}\mathsf{B} \vdash \mathsf{M}_0$ as $\Diamond A \in \mathsf{ES}_2$ and $A \rhd B \to \Diamond A \rhd B$. Actually, we have that $\mathbf{IL}\mathsf{B}_1 \vdash \mathsf{M}_0$.

12.1.3 Beklemishev and Zambella

Zambella proved in ([Zam94]) a fact concerning Π_1-consequences of theories with a Π_2 axiomatization. As we shall see, his result has some repercussions on the study of the interpretability logic of PRA.

Lemma 12.1.17 (Zambella). *Let T and S be two theories axiomatized by Π_2-axioms. If T and S have the same Π_1-consequences then $T + S$ has no more Π_1-consequences than T or S.*

In [Zam94], Zambella gave a model-theoretic proof of this lemma. As was sketched by G. Mints (see [BV04]), also a finitary proof based on Herbrand's theorem can be given. This proof can certainly be formalized at the presence of the superexponentiation function where it yields a principle for the Π_1-conservativity logic of Π_2-axiomatized theories. We denote it here Z^{c}. In the formulation, $\rhd_c$ denotes formalized Π_1-conservativity.

$$\mathsf{Z}^{\mathsf{c}} \quad (A \equiv_c B) \to A \rhd_c A \wedge B \quad \text{for } A \text{ and } B \text{ in } \mathsf{EP}^{\mathsf{c}}_2.$$

For reflexive theories we know that Π_1-conservativity coincides with interpretability. Thus, we see that Zambella's lemma has its repercussions on the interpretability logic of PRA.

As PRA is Π_2-axiomatized, and as any Σ_2-extension of PRA is reflexive, we see that we have Z^{c} for extensions of PRA that are both Σ_2 and Π_2. Put differently, we are interested in Δ_2-extensions of PRA. Thus, we can formulate Zambella's principle for interpretability logic.

$$\mathsf{Z} \quad (A \equiv B) \to A \rhd A \wedge B \quad \text{for } A \text{ and } B \text{ in } \mathsf{ED}_2$$

For the Π_1-conservativity logic of PRA, the principle Z^{c} is really informative (see [BV04]). However, we shall now see that Zambella's principle does not give us additional information for $\mathbf{IL}$(PRA) as Z is provable in $\mathbf{IL}\mathsf{B}$. We also include a direct proof that Z follows semantically from $\mathbf{IL}\mathsf{B}$ on finite frames. In our proofs it becomes clear that we actually only need B_0 and $\mathcal{C}_0$.

Lemma 12.1.18. *Let F be a finite frame with $F \models \mathcal{C}_0$, then $F \models \mathsf{Z}$.*

Proof of Lemma 12.1.18. Let M be any model based on F. Consider $\alpha, \beta \in \mathsf{ED}_2$ with

$$\vdash \alpha \leftrightarrow \bigvee_{i \leq n} (\bigwedge_j \Box A_i \wedge \Diamond B_{ij}). \tag{12.4}$$

Next, consider any $a \in M$ with $a \Vdash \alpha \equiv \beta$. The aim is to show that $a \Vdash \alpha \rhd \alpha \wedge \beta$. To this purpose, we consider any b with $aRb \Vdash \alpha$. We define a sequence of worlds $b_0^m, b_1^m, b_2^m, b_3^m$, $m \in \omega$ satisfying the following properties. (Recall that $\mathcal{S}_0(x,y)$ is just $x{\uparrow} = y{\uparrow}$.)

1. $b_0^m S_a b_1^m S_a b_2^m S_a b_3^m S_a b_0^{m+1}$
2. $b_0^m, b_1^m \Vdash \alpha$ & $\mathcal{S}_0(b_0^m, b_1^m)$
3. $\forall d, e\ (b_1^m S_a dRe \rightarrow b_0^m Re)$
4. $b_2^m, b_3^m \Vdash \beta$ & $\mathcal{S}_0(b_2^m, b_3^m)$
5. $\forall d, e\ (b_3^m S_a dRe \rightarrow b_2^m Re)$

Then, for some $m \in \omega$, we will have that $b_2^m \Vdash \alpha \wedge \beta$. Let us see that this follows from properties 1.-5. For any m, we have by 2. that $b_1^m \Vdash \alpha$. Thus, in any b_1^m some disjunct of (12.4) should hold.

Let k be the first time that $b_1^k \Vdash \alpha$, by satisfying some disjunct $\bigwedge_j (\Box A_i \wedge \Diamond B_{ij})$ that was satisfied by some b_1^l for some $l < k$. Note that $k \geq 1$. We now claim that $b_2^{k-1} \Vdash \alpha \wedge \beta$.

By 4., clearly $b_2^{k-1} \Vdash \beta$. As $l \leq k-1$, we get by 1., 2. and 3. that $\forall e\ (b_2^{k-1} Re \rightarrow b_1^l Re)$ and thus, $b_2^{k-1} \Vdash \Box A_i$.

By 1., 4. and 5. we see that $\forall e\ (b_1^k Re \rightarrow b_2^{k-1} Re)$, whence $b_2^{k-1} \Vdash \bigwedge_j \Diamond B_{ij}$. Thus, indeed $b_2^{k-1} \Vdash \bigwedge_j (\Box A_i \wedge \Diamond B_{ij})$ and $b_2^{k-1} \Vdash \alpha \wedge \beta$.

The proof is thus finished if we can properly define our sequence of worlds. The sequence $b_0^0, b_1^0, \cdots$ is defined in the obvious way.

- $b_0^0 = b$
- As aRb_0^m, by $\mathcal{C}_0$, we can find b_1^m with $b_0^m S_a b_1^m$, $\mathcal{S}_0(b_0^m, b_1^m)$ and $\forall d, e\ (b_1^m S_a dRe \rightarrow b_0^m Re)$.
- As $b_0^m \Vdash \alpha$ and $\mathcal{S}_0(b_0^m, b_1^m)$, also $b_1^m \Vdash \alpha$. As $a \vdash \alpha \rhd \beta$, we can find b_2^m with $b_1^m S_a b_2^m \Vdash \beta$.
- Again, as aRb_2^m, by $\mathcal{C}_0$, we can find b_3^m with $b_2^m S_a b_3^m$, $\mathcal{S}_0(b_2^m, b_3^m)$ and $\forall d, e\ (b_3^m S_a dRe \rightarrow b_2^m Re)$.
- As $b_2^m \Vdash \beta$ and $\mathcal{S}_0(b_2^m, b_3^m)$, also $b_3^m \Vdash \beta$. As $a \vdash \beta \rhd \alpha$, we can find b_0^{m+1} with $b_3^m S_a b_0^{m+1} \Vdash \beta$.

$\dashv$

We shall now give a purely syntactical proof of $\mathbf{IL}\mathsf{B}_0 \vdash \mathsf{Z}$.

Lemma 12.1.19. *$\mathbf{\mathit{IL}}B \vdash \mathsf{B}'$, where $\mathsf{B}' : A \rhd B \to A \wedge C \rhd B \wedge C$ with $A \in ES_2$ and C a CNF of boxed formulas.*

Proof. Easy. ⊣

Theorem 12.1.20. $\mathbf{IL}\mathsf{B}_0 \vdash \mathsf{Z}$

Proof. First we notice that in **IL**, every ED_2-formula is equivalent to its disjunctive normal form. Thus, we need to prove the statement only for formulas in DNF.

Let $A, B \in \mathsf{ED}_2$ where a DNF of A is given by $\bigvee_{i=1}^{n}(\Box A_i \wedge \mathbf{C}_i)$. Here $\mathbf{C}_i$ is short for $\bigwedge_{j=1}^{k_i} \Diamond C_{ij}$ for some suitable indices. From now on we shall only give the range of the indices if necessary.

We show that for our $B \in \mathsf{ED}_2$ we have

$$(\bigvee(\Box A_i \wedge \mathbf{C}_i) \equiv B) \to \bigvee(\Box A_i \wedge \mathbf{C}_i) \rhd (\bigvee(\Box A_i \wedge \mathbf{C}_i)) \wedge B$$

Thus, our two assumptions are:

$$\bigvee(\Box A_i \wedge \mathbf{C}_i) \rhd B, \tag{12.5}$$

$$B \rhd \bigvee(\Box A_i \wedge \mathbf{C}_i). \tag{12.6}$$

By (12.6) and B' on $\bigwedge \overline{\mathbf{C}_i}$, where $\overline{\mathbf{C}_i}$ stands for $\bigvee \Box\neg C_{ij}$, we get $B \wedge \bigwedge \overline{\mathbf{C}_i} \rhd \bot$, i.e.,

$$\Box(B \to \bigvee \mathbf{C}_i). \tag{12.7}$$

Now we consider any disjunct of A, say $\Box A_n \wedge \mathbf{C}_n$ and prove that $\Box A_n \wedge \mathbf{C}_n \rhd A \wedge B$.

Clearly $\Box A_n \wedge \mathbf{C}_n \rhd B$. Together with (12.7) and using B' this yields $\Box A_n \wedge \mathbf{C}_n \rhd B \wedge \Box A_n \wedge (\bigvee \mathbf{C}_i)$. By propositional logic we see conclude the following.

$$\begin{array}{rcl} B \wedge \Box A_n \wedge (\bigvee \mathbf{C}_i) & \leftrightarrow & \bigvee_i (B \wedge \mathbf{C}_i \wedge \Box A_n) \\ & \leftrightarrow & \bigvee_i (B \wedge \mathbf{C}_i \wedge \Box A_n \wedge \bigwedge_{j \neq i} \overline{\mathbf{C}_j}) \end{array}$$

Now we consider any disjunct with index m of the latter, and show that

$$B \wedge \mathbf{C}_m \wedge \Box A_n \wedge \bigwedge_{j \neq m} \overline{\mathbf{C}_j} \rhd A \wedge B.$$

Clearly $B \wedge \mathbf{C}_m \wedge \Box A_n \wedge \bigwedge_{j \neq m} \overline{\mathbf{C}_j} \rhd A$. By B' we get

$$B \wedge \mathbf{C}_m \wedge \Box A_n \wedge \bigwedge_{j \neq m} \overline{\mathbf{C}_j} \rhd A \wedge \Box A_n \wedge \bigwedge_{j \neq m} \overline{\mathbf{C}_j}.$$

But, $A \wedge \bigwedge_{j \neq m} \overline{\mathbf{C}_j}$ gives us $\Box A_m \wedge \mathbf{C}_m \wedge \bigwedge_{j \neq m} \overline{\mathbf{C}_j}$. We can conclude with the following argument.

$$\begin{array}{rcll} A \wedge \bigwedge_{j \neq m} \overline{\mathbf{C}_j} & \rhd & \Box A_m \wedge \mathbf{C}_m \wedge \bigwedge_{j \neq m} \overline{\mathbf{C}_j} & \text{by (12.5) and } \mathsf{B}' \\ & \rhd & B \wedge \Box A_m \wedge \bigwedge_{j \neq m} \overline{\mathbf{C}_j} & \text{by (12.7)} \\ & \rhd & B \wedge \Box A_m \wedge (\bigvee \mathbf{C}_i) \wedge \bigwedge_{j \neq m} \overline{\mathbf{C}_j} & \\ & \rhd & B \wedge \Box A_m \wedge \mathbf{C}_m & \\ & \rhd & A \wedge B & \end{array}$$

$\dashv$

We see that Zambella's principle does not add any information to **IL**B. The best candidate for **IL**(PRA) is thus **IL**BR^*. The principles B, R and W seem to be as independent from each other as can be.

12.2 Upperbounds

Let $\mathsf{Sub}(\Gamma)$ be the set of realizations that take their values in Γ. We define the Γ-interpretability logic of T to be set of all formulas in $\mathsf{Form}_{\mathbf{IL}}$ that are provable in T under any realization in $\mathsf{Sub}(\Gamma)$. We denote this logic by $\mathbf{IL}_\Gamma(\mathrm{T})$. Clearly we have that $\mathbf{IL}_\Delta(\mathrm{T}) \subseteq \mathbf{IL}_\Gamma(\mathrm{T})$ whenever $\Gamma \subseteq \Delta$. This observation can be used to obtain a rough upperbound for **IL**(PRA). In order to do so, we first calculate the Γ-*provability* logic of PRA for a specific Γ. This is defined completely analogously to its interpretability variant and is denoted by $\mathbf{PL}_\Gamma(\mathrm{PRA})$.

First we define the set $\mathcal{B}$ of arithmetical sentences as follows.

$$\mathcal{B} := \bot \mid \top \mid \Box(\mathcal{B}) \mid \Diamond(\mathcal{B}) \mid \mathcal{B} \to \mathcal{B} \mid \mathcal{B} \vee \mathcal{B} \mid \mathcal{B} \wedge \mathcal{B}$$

Definition 12.2.1. The logic **RGL** is obtained by adding the linearity axiom schema $\Box(\Box A \to B) \vee \Box(\boxdot B \to A)$ to **GL**. Here $\boxdot B$ is an abbreviation of $B \wedge \Box B$.

Theorem 12.2.2. $\mathbf{PL}_{\mathcal{B}}(\mathrm{PRA}) = \mathbf{RGL}$

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

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

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

$$\varphi_m := \Box^{m+1}_{\mathrm{PRA}}\bot \wedge \Diamond^{m}_{\mathrm{PRA}}\top.$$

From now on we will omit the subscript PRA. It is easy to see that

1. $\mathrm{PRA} \vdash \varphi_l \to \neg\varphi_m \quad$ if $l \neq m$,
2. $\mathrm{PRA} \vdash \varphi_l \to \Box(\bigvee_{m<l}\varphi_m)$,
3. $\mathrm{PRA} \vdash \varphi_l \to \bigwedge_{m<l}\Diamond\varphi_m$.

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

$$\begin{array}{ll}\mathbf{L}_n, m \Vdash C \Rightarrow \mathrm{PRA} \vdash \varphi_m \to C^* & \text{and}\\ \mathbf{L}_n, m \nVdash C \Rightarrow \mathrm{PRA} \vdash \varphi_m \to \neg C^*. & \end{array}$$

By this truth lemma, $\mathbf{L}_n, n{-}1 \Vdash \neg A \Rightarrow \mathrm{PRA} \vdash \varphi_{n-1} \to (\neg A)^*$ and consequently $\mathrm{PRA} \vdash \Diamond\varphi_{n-1} \to \neg\Box A^*$. Thus $\mathbb{N} \models \Diamond\varphi_{n-1} \to \neg\Box A^*$. As φ_{n-1} is consistent with PRA we see that $\mathbb{N} \models \Diamond\varphi_{n-1}$ whence $\mathbb{N} \models \neg\Box A^*$ and thus $\mathrm{PRA} \nvdash A^*$. $\dashv$

Definition 12.2.3. The logic **RIL** is obtained by adding the linearity axiom schema $\Box(\Box A \to B) \vee \Box(\boxdot B \to A)$ to $\mathbf{IL}\mathsf{W}$.

Theorem 12.2.4. $\mathbf{RIL} = \mathbf{IL}_{\mathcal{B}}(\mathrm{PRA})$

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

$$\begin{array}{c}\mathbf{RIL} \vdash \varphi \Leftrightarrow \mathbf{RGL} \vdash \varphi^{\mathsf{tr}} \quad (*)\\ \text{and}\\ \mathbf{RIL} \vdash \varphi \leftrightarrow \varphi^{\mathsf{tr}}. \qquad (**)\end{array}$$

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

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

and consequently

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

We first see that $(***)$ holds. Certainly $\mathbf{IL}\mathsf{W} \subseteq \mathbf{IL}_{\mathcal{B}}(\mathrm{PRA})$. Thus it remains to show that $\mathrm{PRA} \vdash \Box(\Box A^* \to B^*) \vee \Box(\boxdot B^* \to A^*)$ for any formulas A and B

in $\mathsf{Form}_{\mathbf{IL}}$ and any $*{\in}\mathsf{Sub}(\mathcal{B})$. As any formula in the closed fragment of **IL**W is equivalent to a formula in the closed fragment of **GL** (see [Hv91]), Theorem 12.2.2 gives us that indeed the linearity axiom holds for the closed fragment of **GL**.

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

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

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

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

We now prove $(*)$. By induction on $\mathbf{RIL} \vdash \varphi$ we see that $\mathbf{RGL} \vdash \varphi^{\mathsf{tr}}$. All the specific interpretability axioms turn out to be provable under our translation in **GL**. The only axioms where the $\Box A \to \Box\Box A$ axiom scheme is really used is in J_2 and J_4. To prove the translation of W we also need L_3.

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

We thus see that **RIL** is an upperbound for **IL**(PRA). Using the translation from the proof of Theorem 12.2.4, it is not hard to see that both the principles P and M are provable in **RIL**. This tells us that the upperbound is actually not very informative as we know that $\mathbf{IL}(\mathrm{PRA}) \nvdash \mathsf{M}$. Choosing larger Γ will generally yield a smaller $\mathbf{IL}_{\Gamma}(\mathrm{PRA})$ and thus a sharper upperbound.

In [Vis97] it is shown that $\mathbf{IL}(\mathrm{PRA}) \nvdash A \rhd \Diamond B \to \Box(A \rhd \Diamond B)$, which implies that M is certainly not derivable. We can also find explicit realizations that violate M, as the following lemma tells us.

Lemma 12.2.5. *For $n \geq 1$, we have that* $\mathbf{IL}(\mathrm{I}\Sigma_{\mathrm{n}}^{\mathrm{R}}) \nvdash \mathsf{M}$.

Proof. We will expose a realization $*$ such that $\mathrm{I}\Sigma_n^R \nvdash (p \rhd q \to p \wedge \Box r \rhd q \wedge \Box r)^*$.

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

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

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

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

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

We see that the realizations used in the proof of Lemma 12.2.5 get higher and higher complexities. The complexity is certainly higher than Σ_2. By Theorem 12.1.1 we know that this is necessarily so. This observation also indicates that an arithmetical completeness proof can not work with only Σ_2-realizations.

It is an open question if $\mathbf{IL}(\mathrm{PRA}) \subseteq \mathbf{IL}\mathsf{M}$. For $\mathrm{I}\Sigma^R_n$, $n \geq 2$ we know that $\mathbf{IL}(\mathrm{I}\Sigma^{\mathrm{R}}_{\mathrm{n}}) \subset \mathbf{IL}\mathsf{M}$. This follows from the next lemma.

Lemma 12.2.6. $\mathbf{IL}_{\Sigma_2}(\mathrm{I}\Sigma^{\mathrm{R}}_{\mathrm{n}}) = \mathbf{IL}_{\Delta_{n+1}}(\mathrm{I}\Sigma^{\mathrm{R}}_{\mathrm{n}}) = \mathbf{IL}\mathsf{M}$ *whenever* $n \geq 2$.

Proof. We shall use that the logic of Π_1-conservativity for theories containing $\mathrm{I}\Sigma_1$ is $\mathbf{IL}\mathsf{M}$ ([HM90], [HM92]). By $U \rhd_{\Pi_1} V$ we denote the formalization of the statement "U is Π_1-conservative over V".

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

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

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

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

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

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

12.3 Encore: graded provability algebras

In this final subsection, we shall make some remarks on the universal model of the closed fragment of **GLP** as introduced by Ignatiev in [Ign93a]. We shall see that upperbounds as provided in Theorem 12.2.2 are not easily improved by switching to the closed fragment of **GLP**.

12.3.1 The logic GLP

Japaridze's logic **GLP**, as defined below ([Dzh86]) has for each $n \in \omega$ a modality $[n]$. An arithmetical reading of $[n]\varphi$ is "φ is provable in T together with all true Π_n-statements". **GLP** is known to be sound and complete with respect to this reading for sound arithmetical theories T containing EA.

Definition 12.3.1. The axioms of **GLP** are

1. Boolean tautologies,
2. $[n]([n]\varphi \to \varphi) \to [n]\varphi$ for all n,
3. $[m]\varphi \to [n]\varphi$ for $m \leq n$,
4. $\langle m\rangle\varphi \to [n]\langle m\rangle\varphi$ for $m < n$.

The rules of **GLP** are modus ponens and necessitation.

It is easy to see that **GLP** does not allow for a natural Kripke semantics. Ignatiev however, gave a nice universal model for the closed fragment of **GLP**. The closed fragment of **GLP** is related to ordinal notation systems and combinatoric principles independent from PA, like "every worm dies" (see [Bek04], [Bek03b]).

The universal model $\mathcal{U}$ for **GLP** should satisfy two main properties.

- **GLP** $\vdash \varphi \Rightarrow \forall x\, \mathcal{U}, x \Vdash \varphi$
- **GLP** $\nvdash \varphi \Rightarrow \exists x\, \mathcal{U}, x \Vdash \neg\varphi$

From the arithmetical soundness of **GLP**, we know that **GLP** $\nvdash [m_0]\cdots[m_n]\bot$ for any sequence $m_0 \cdots m_n$. Thus, we should be able to find an $x \in \mathcal{U}$ with $x \Vdash \langle m_0\rangle \cdots \langle m_n\rangle\top$. It is easy to see that **GLP** $\vdash \langle 1\rangle\top \to \langle 0\rangle^n\top$ for any n. Thus, for example, any time the model contains a transition to witness $\langle 1\rangle\top$, there should be chains of transitions of arbitrary length witnessing the $\langle 0\rangle^n\top$.

If φ is a formula in the language of **GLP**, we denote by φ^+ the formula that arises by making all the modalities one higher. Thus, $(\langle 0\rangle\top \wedge [1]\bot)^+ = \langle 1\rangle\top \wedge [2]\bot$. It is clear that **GLP** $\vdash \varphi \Rightarrow$ **GLP** $\vdash \varphi^+$. This phenomenon is also nicely reflected in the universal model. The R_n-transitions, corresponding to the $\langle n\rangle$-modality, repeat the behavior of the lower modalities. All these considerations combine to yield Ignatiev's model as depicted in figure 12.1. (We have not depicted the arrows that should be there by transitivity.)

The depth of $\mathcal{U}$ will be ϵ_0. In the next subsection we shall give a formal definition.

12.3.2 A universal model for GLP$_0$

We shall make extensive use of ordinals to describe our universal model $\mathcal{U}$. In this section, all ordinals denote ordinals below ϵ_0. We will denote them by lower case Greek letters.

Definition 12.3.2. If $\alpha = \alpha' + \omega^\gamma$ with $\alpha' + \omega^\gamma$ in Cantor Normal Form, (α' might be 0 in the case of the empty sum) then $d(\alpha) := \gamma$ and $\alpha^- := \alpha'$.

For simplicity we set $d(0) = 0$. The intuition behind the construction of $\mathcal{U}_\alpha$ (the universal model up to stage α) is as follows. For $\mathcal{U}_0$ we just take one irreflexive point. If we wish to construct $\mathcal{U}_\alpha$, we first take the union of all the

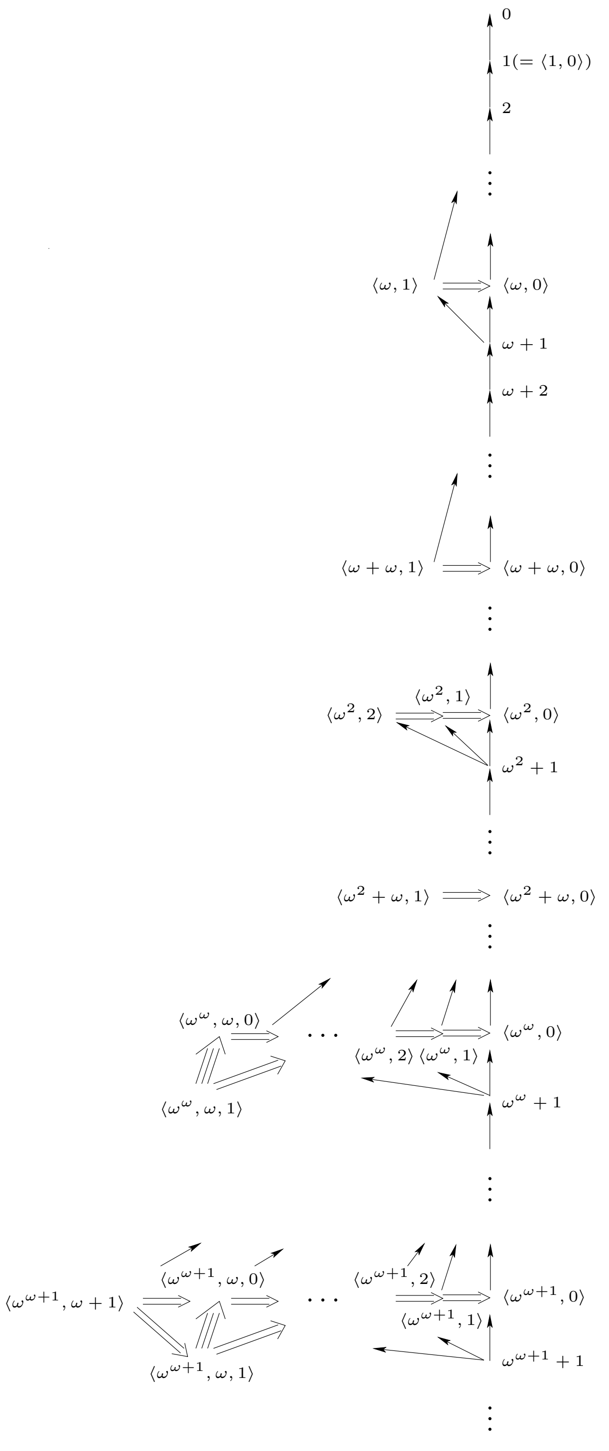
0
$1(=\langle 1,0\rangle)$
2
$\langle\omega,1\rangle$
$\langle\omega,0\rangle$
$\omega+1$
$\omega+2$
$\langle\omega+\omega,1\rangle$
$\langle\omega+\omega,0\rangle$
$\langle\omega^2,1\rangle$
$\langle\omega^2,2\rangle$
$\langle\omega^2,0\rangle$
ω^2+1
$\langle\omega^2+\omega,1\rangle$
$\langle\omega^2+\omega,0\rangle$
$\langle\omega^\omega,\omega,0\rangle$
$\langle\omega^\omega,0\rangle$
$\langle\omega^\omega,2\rangle$
$\langle\omega^\omega,1\rangle$
$\omega^\omega+1$
$\langle\omega^\omega,\omega,1\rangle$
$\langle\omega^{\omega+1},\omega,0\rangle$
$\langle\omega^{\omega+1},2\rangle$
$\langle\omega^{\omega+1},\omega+1\rangle$
$\langle\omega^{\omega+1},0\rangle$
$\langle\omega^{\omega+1},1\rangle$
$\omega^{\omega+1}+1$
$\langle\omega^{\omega+1},\omega,1\rangle$

Figure 12.1: The universal model for $\mathbf{GLP}_0$

$\mathcal{U}_\beta$ with $\beta < \alpha$ (the $\mathcal{U}_\beta$ come with natural embeddings). Then, we consider $\mathcal{U}_{d(\alpha)}$. We make all the relations in $\mathcal{U}_{d(\alpha)}$ one higher (that is, an R_n becomes an R_{n+1}-relation) and then place it R_0-below $\cup_{\beta<\alpha}\mathcal{U}_\beta$. In this process we see that the maximal length of an R_{n+1}-chain is always certainly a power of ω behind the maximal length of an R_n-chain. Of course, in the limit all the lengths catch up.

Definition 12.3.3 (Universal model). $\mathcal{U} := \{x \in \prod_{i<\omega}[0, \epsilon_0) \mid \forall i \ (x)_{i+1} \leq d((x)_i)\}$ and xR_ny iff $((x)_n > (y)_n$ and $(x)_i = (y)_i$ for all $i < n$.

We will write elements of $\mathcal{U}$ as finite sequences, omitting all the zero elements. The sole exception of course, is the sequence $\vec{0}$ which we shall denote by 0.

We can also define the universal model up to stage α for $\alpha < \epsilon_0$.

$$\mathcal{U}_\alpha := \{x \in \prod_{i<\omega}[0, \epsilon_0) \mid (x)_0 \leq \alpha \wedge \forall i \ (x)_{i+1} \leq d((x)_i)\}$$

Lemma 12.3.4. *The formal definition of $\mathcal{U}$ captures the intuitive construction of it.*

Proof. One just has to carefully describe the process in the intuitive picture. Every model comes with a set of embeddings into larger models. As an intermediate coding of the process one can use the definition of $\mathcal{I}_\alpha$ as given in Definition 12.3.5. It is not hard to show that $\forall\, \alpha{<}\epsilon_0 \ \mathcal{I}_\alpha = \mathcal{U}_\alpha$ and that $\bigcup_{\alpha<\epsilon_0} \mathcal{I}_\alpha = \mathcal{U}$. $\dashv$

Definition 12.3.5.
$\mathcal{I}_0 := \{\vec{0} \in \prod_{i<\omega}[0, \epsilon_0)\}$, that is, just one irreflexive point,
$(\mathcal{I}_\alpha)^+ := \{x \in \prod_{i<\omega}[0, \epsilon_0) \mid \exists\, y{\in}\mathcal{I}_\alpha \ ((x)_0 = 0 \wedge \forall i \ (x)_{i+1} = (y)_i)\}$,
$\mathcal{I}_\alpha := (\mathcal{I}_{d(\alpha)})^+ \oplus \vec{\bigcup}_{\beta<\alpha}\mathcal{I}_\beta$.
Here the $\vec{\bigcup}$ is the familiar direct limit, and $\oplus$ is the relation defined by the following.
$B^+ \oplus A := \{y \in \prod_{i<\omega}[0, \epsilon_0) \mid \exists\, z{\in}B^+ \ ((y)_0{=}\sup\{(x)_0 | x{\in}A\} \wedge \forall i \ (z)_{i+1} = (y)_{i+1})\} \cup A$

In [Ign93a] it is shown that, indeed, $\mathcal{U}$ is a universal model for the closed fragment of **GLP**. His proofs make at various points reference to the arithmetical interpretation. In the next subsection we shall prove that the model is sound for **GLP**, by a proof that uses modal considerations only.

12.3.3 Bisimulations and soundness

In this subsection, we shall make extensive use of bisimulations.

Definition 12.3.6 (n-bisimilarity).

- x, x' are 0-bisimilar, we write $x \simeq_0 x'$, if $\forall p \ [x \Vdash p \Leftrightarrow x' \Vdash p]$
- x, x' are $n+1$-bisimilar, we write $x \simeq_{n+1} x'$, if

– x, x' are n-bisimilar
– $\forall\, i{<}\omega\, \forall y\ (xR_iy \to \exists y'\ (x'R_iy'\ \&\ y$ and y' are n-bisimilar))
– $\forall\, i{<}\omega\, \forall y'\ (x'R_iy' \to \exists y\ (xR_iy\ \&\ y'$ and y are n-bisimilar))

Lemma 12.3.7. *Let M and N be models of the same modal signature. If $m \in M$ and $n \in N$ are l-bisimilar, then*

$$M, m \Vdash \varphi \Leftrightarrow N, n \Vdash \varphi$$

for every φ with $\mathsf{rk}(\varphi) \leq l$. Here $\mathsf{rk}(\varphi)$ is the modality rank of φ, that is, the maximal number of nested modalities in φ.

Proof. The lemma is well-known and can easily be proven for example with a variant of Ehrenfeucht-Fraïssé games and the first order translation of modal logic. (See also [BRV01], [CZ97].) ⊣

Lemma 12.3.8.

$$\begin{array}{ll} (i) & \beta < \alpha\ \&\ \gamma < d(\alpha) \Rightarrow \beta + \omega^\gamma < \alpha \\ (ii) & \beta < \alpha\ \&\ \gamma \leq d(\alpha) \Rightarrow \beta + \omega^\gamma \leq \alpha \end{array}$$

Proof. By elementary observations on the Cantor Normal Forms (CNF's) of α and β. ⊣

Definition 12.3.9 (Nested width). An ordinal α has *nested width* at least n, we write $\mathsf{NW}(\alpha) \leq n$, iff the CNF of α contains at most n terms and each exponent of each term has nested width at least n.

We note that for each α, there is an n with $\mathsf{NW}(\alpha) \leq n$, and that $\mathsf{NW}(\alpha) \leq n \to \mathsf{NW}(\alpha) \leq n + 1$. As all our ordinals are below ϵ_0, we see that for any p and α, there are only finitely many $\beta < \alpha$ with $\mathsf{NW}(\beta) \leq p$.

Definition 12.3.10. We define an alternative fundamental sequence as follows. $\alpha\langle p\rangle := \max\{\xi < \alpha \mid \mathsf{NW}(\xi) \leq p\}$

It is immediate that $\forall p\ \alpha\langle p\rangle < \alpha$. If α is a limit ordinal, we have $\cup_{p\in\omega}\alpha\langle p\rangle = \alpha$. We also note that $\alpha\langle p\rangle$ is monotone in both α and p.

Lemma 12.3.11 (Bisimulation lemma). *Let $\vec{\alpha}, \vec{\beta} \in \mathcal{U}$ with for all i,*

$$\begin{array}{c} \alpha_i > \beta_i\langle p\rangle \\ \beta_i > \alpha_i\langle p\rangle \end{array}$$

then, $\vec{\alpha} \simeq_p \vec{\beta}$.

Proof. By induction on p. The case $p = 0$ is trivial, so let us consider the induction step. We assume that for all i, $\alpha_i > \beta_i\langle p + 1\rangle$ and $\beta_i > \alpha_i\langle p + 1\rangle$. We have to see that we can reply any R_n-step from $\vec{\alpha}$ to $\vec{\alpha'}$ with an R_n-step from $\vec{\beta}$ to a $\vec{\beta'}$ so that $\alpha' \simeq_p \beta'$ and vice versa.

So, we suppose that we make some R_n-step in $\vec{\alpha}$, that is, we go from

$$\begin{array}{lll} \vec{\alpha} = & \alpha_0, \cdots, \alpha_{n-1}, \alpha_n, \cdots & \text{to} \\ \vec{\alpha'} := & \alpha_0, \cdots, \alpha_{n-1}, \alpha'_n, \cdots, \alpha'_m. & \end{array}$$

We now reply this step in $\vec{\beta}$ by going to a $\vec{\beta'}$ with the same length as $\vec{\alpha'}$ defined as follows.

$$\begin{array}{lll} \beta'_m := & \alpha'_m\langle p\rangle + 1 & \\ \beta'_k := & \alpha'_k\langle p\rangle + \omega^{\beta'_{k+1}} & \text{for } n \leq k < m \\ \beta'_k := & \beta_k & \text{for } 0 \leq k < n \end{array}$$

We now claim that for all i, we have

$$\begin{array}{ll} (a) & \alpha'_i > \beta'_i\langle p\rangle, \\ (b) & \beta'_i > \alpha'_i\langle p\rangle. \end{array}$$

Let us first see (a). By induction and using Lemma 12.3.8, we see that $\alpha'_i \geq \beta'_i$ for $n \leq i \leq m$. For $i < n$ we have $\alpha'_i = \alpha_i > \beta_i\langle p+1\rangle \geq \beta_i\langle p\rangle = \beta'_i\langle p\rangle$.

For $n \leq i \leq m$, (b) follows by the definition of $\vec{\beta'}$ and an easy induction. For $i < n$ we reason as in (a).

The induction hypothesis now gives us that $\vec{\alpha'} \simeq_p \vec{\beta'}$. Thus, we only need to see that going from $\vec{\beta}$ to $\vec{\beta'}$ is indeed a transition in $\mathcal{U}$. That is, we need to see that $\beta'_n < \beta_n$ (with this it is also clear that for all i, $d(\beta'_i) \geq \beta'_{i+1}$).

We know that $\beta'_n \leq \alpha'_n < \alpha_n$. By an easy induction we see that $\mathsf{NW}(\beta'_i) \leq p+1$ for $n \leq i \leq m$. Thus, $\alpha_n\langle p+1\rangle \geq \beta'_n$. Combining this with our assumption, we get $\beta_n > \alpha_n\langle p+1\rangle \geq \beta'_n$. ⊣

Theorem 12.3.12. **GLP** *is sound with respect to* $\mathcal{U}$.

Proof. By induction on **GLP**-proofs. Löb's axioms follow from the fact that the model is transitive and conversely well-founded. The only axioms that need some special attention are

$$\begin{array}{ll} \langle n\rangle\varphi \to [m]\langle n\rangle\varphi & m > n \ \text{ and} \\ \langle n\rangle\varphi \to \langle m\rangle\varphi & m \leq n. \end{array}$$

The first follows from elementary observations on $\mathcal{U}$. For the second, we reason as follows. It suffices to show that $\langle n+1\rangle\varphi \to \langle n\rangle\varphi$. So, suppose that for some $\vec{\alpha} \in \mathcal{U}$ we have $\vec{\alpha} \Vdash \langle n+1\rangle\varphi$. Thus, for some $\vec{\alpha'}$ with $\vec{\alpha}R_{n+1}\vec{\alpha'}$ we have $\vec{\alpha'} \Vdash \varphi$.

Let $p := \mathsf{rk}(\varphi)$. By Lemma 12.3.7 it suffices to find some $\vec{\beta}$ with $\vec{\alpha'}R_n\vec{\beta}$ and $\vec{\alpha'} \simeq_p \vec{\beta}$. Now, if

$$\begin{array}{lll} \vec{\alpha}: & \alpha_0, \cdots, \alpha_{n-1}, \alpha_n, \alpha_{n+1}, \cdots & \text{goes to} \\ \vec{\alpha'}: & \alpha_0, \cdots, \alpha_{n-1}, \alpha_n, \alpha'_{n+1}, \cdots, \alpha'_m, & \end{array}$$

we define $\vec{\beta}$ of the same length as $\vec{\alpha'}$ as follows.

$$\begin{array}{lll} \beta_m := & \alpha'_m\langle p\rangle + 1 & \\ \beta_k := & \alpha'_k\langle p\rangle + \omega^{\beta_{k+1}} & \text{for } n \leq k < m \\ \beta_k := & \alpha_k & \text{for } 0 \leq k < n \end{array}$$

By induction and Lemma 12.3.8 we see that $\beta_k \leq \alpha'_k$ for $n+1 \leq k \leq m$. As $\alpha'_{n+1} < \alpha_{n+1}$, we see by Lemma 12.3.8 that $\beta_n < \alpha_n$. Moreover, $\vec{\alpha'}$ and $\vec{\beta}$ satisfy the requirements of Lemma 12.3.11 and we conclude that $\vec{\alpha'} \simeq_p \vec{\beta}$. $\dashv$

Definition 12.3.13. The Main Axis MA on $\mathcal{U}$ is defined as follows.

$$\mathsf{MA} := \{x \in \mathcal{U} \mid \forall i \ (x)_{i+1} = d((x)_i)\}$$

Corollary 12.3.14. *Any point $\vec{\alpha} \in \mathcal{U}$ is p-bisimilar to some $\vec{\beta} \in \mathsf{MA}$.*

Proof. Given α and p, we consider $\vec{\beta}$ with $\beta_k = \alpha_k\langle p\rangle + \omega^{\beta_{k+1}}$. $\dashv$

Let $\mathcal{D}$ denote the arithmetical translation of the closed fragment of **GLP**, where [0] is translated to "provable in PRA ".

Lemma 12.3.15. $\mathbf{PL}_{\mathcal{D}}(\mathrm{PRA}) = \mathbf{RGL}$

Proof. By Lemma 12.2.2, it is clear that $\mathbf{PL}_{\mathcal{D}}(\mathrm{PRA}) \subseteq \mathbf{RGL}$. For the other inclusion, we only need to see that

$$\mathrm{PRA} \vdash \Box(\Box A \to B) \vee \Box(\boxdot B \to A) \tag{12.8}$$

for any $A, B \in \mathcal{D}$. With the arithmetical completeness and the modal semantics at hand, it is easy to see (12.8).

For, suppose for a contradiction that at some x, we have $x \Vdash \Diamond(\Box A \wedge \neg B) \wedge \Diamond(\boxdot B \wedge \neg A)$. By Corollary 12.3.14 we can find y_0 and y_1 on the main axis with $y_0 \Vdash \Box A \wedge \neg B$ and $y_1 \Vdash \boxdot B \wedge \neg A$. As y_0 and y_1 lie on the main axis, we have $y_0 = y_1$, $y_0 R y_1$ or $y_1 R y_0$. All of these possibilities lead to a contradiction. $\dashv$

12.3.4 Finite approximations

The soundness proof we gave in Theorem 12.3.12 used the observation that $\mathcal{U}$ is a conversely well-founded model. This implies that our proof is not even formalizable in PA, as the depth of $\mathcal{U}$ is ϵ_0. Yet, we have the idea that this much induction is not needed to reason about the closed fragment of the decidable logic **GLP**.

One way to get some sort of soundness available in weaker theories is by means of finite approximations of $\mathcal{U}$.

Definition 12.3.16. $\alpha \prec \beta$ iff $\alpha < \beta$ and $\mathsf{NW}(\alpha) \leq \min\{n \mid \mathsf{NW}(\beta) \leq n\}$

Definition 12.3.17 (Finite approximations). $\mathcal{F}_0 := \{\vec{0} \in \prod_{i<\omega}[0,\epsilon_0)\}$, that is, just one irreflexive point,
$\mathcal{F}_\alpha := (\mathcal{F}_{d(\alpha)})^+ \oplus \vec{\bigcup}_{\beta \prec \alpha}\mathcal{F}_\beta$.
Here the $\oplus$ and the $(\cdot)^+$ are as in Definition 12.3.5.

Lemma 12.3.18. *There exists increasing sequences α_i with $\vec{\bigcup}_{i<\omega}\mathcal{F}_{\alpha_i} = \mathcal{U}$.*

Proof. An example is $\alpha_i := \omega_i \cdot i$. $\dashv$

Lemma 12.3.19. *If φ is a closed formula that is provable in* **GLP***, then there is a proof of φ in which only closed formulas occur.*

Proof. If such a proof contains propositional variables, we may substitute $\top$ for them and obtain the desired proof. $\dashv$

Theorem 12.3.20. *If a closed formula φ is provable in* **GLP** *using only closed formulas of complexity $\leq n$, then $\mathcal{F}_\alpha \models \varphi$ for any α with* $\min\{p \mid \mathsf{NW}(\alpha) \leq p\} \geq n$.

Proof. By induction on such a proof. We note that all the points that are needed to repeat the proof of Theorem 12.3.12 are available in $\mathcal{F}_\alpha$. $\dashv$

Corollary 12.3.21. $\mathrm{EA} + \mathsf{supexp} \vdash \mathsf{Con}(\mathbf{GLP}_0 + \{\langle\ \rangle\top, \langle 1\rangle\top, \langle 2\rangle\top, \dots\})$

Proof. We reason in $\mathrm{EA} + \mathsf{supexp}$. Suppose for a contradiction that for some m, $\mathbf{GLP}_0 \vdash [m]\bot$. Then, for some $n{\geq}m$ we have $\mathbf{GLP}_0 \vdash_n [m]\bot$. That is, $[m]\bot$ is is provable in $\mathbf{GLP}_0$ using only formulas of complexity $\leq n$.

The number of points in $\mathcal{F}_{\omega_n \cdot n}$ is certainly bounded by some term using $\mathsf{supexp}(n)$. Sharper bounds may be obtained by analyzing in more detail the number of β for which $\beta \prec \alpha$ for a given α. As $\mathcal{F}_{\omega_n \cdot n} \not\models \varphi$, we get a contraction by Theorem 12.3.20. $\dashv$

Bibliography

[Avi02] J. Avigad. Saturated models of universal theories. *Annals of Pure and Applied Logic*, 118(3):219–234, 2002.

[Bek93] L.D. Beklemishev. On the complexity of arithmetic interpretations of modal formulae. *Archive for Mathematical Logic*, 32:229–238, 1993.

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

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

[Bek03a] L.D. Beklemishev. Proof-theoretic analysis by iterated reflection. *Archive for Mathematical Logic*, 42(6):515–552, 2003.

[Bek03b] L.D. Beklemishev. The worm principle. Logic Group Preprint Series 219, University of Utrecht, 2003.

[Bek04] L. D. Beklemishev. Provability algebras and proof-theoretic ordinals, I. *Annals of Pure and Applied Logic*, 128:103–123, 2004.

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

[BGJ04] M. Bilkova, E. Goris, and J.J. Joosten. Smart labels. To appear, 2004.

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

[BRV01] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic.* Camebridge University Press, 2001.

[Bus98] S.R. Buss. First-order proof theory of arithmetic. In S.R. Buss, editor, *Handbook of Proof Theory*, pages 79–148, Amsterdam, 1998. Elsevier, North-Holland.

[BV04] L.D. Beklemishev and A. Visser. On the limit existence principles in elementary arithmetic and related topics. Logic Group Preprint Series 224, University of Utrecht, 2004.

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

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

[dJP96] D. de Jongh and D. Pianigiani. Solution of a problem of David Guaspari. *Studia Logica*, 1996.

[dJV90] D.H.J. de Jongh and F. Veltman. Provability logics for relative interpretability. In *[Pet90]*, pages 31–42, 1990.

[dJV99] D.H.J. de Jongh and F. Veltman. Modal completeness of ILW. In J. Gerbrandy, M. Marx, M. Rijke, and Y. Venema, editors, *Essays dedicated to Johan van Benthem on the occasion of his 50th birthday.* Amsterdam University Press, Amsterdam, 1999.

[dR92] M. de Rijke. Unary interpretability logic. *The Notre Dame Journal of Formal Logic*, 33:249–272, 1992.

[Dzh86] G.K. Dzhaparidze. *Modal Logical Means of Investigation of Provability.* PhD thesis, Tbilisi State University, 1986. In Russian.

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

[Fef88] S. Feferman. Hilbert's program relativized: proof-theoretical and foundational reductions. *Journal of Symbolic Logic*, 53:364–384, 1988.

[Fer02] F. Ferreira. Yet another proof of Parsons' theorem. Not yet published, 2002.

[GD82] H. Gaifman and C. Dimitracopoulos. Fragments of Peano's arithmetic and the MDRP theorem. In *Logic and Algorithmic*, pages 319–329. l' Enseignement Mathématique, monographie nr. 30, 1982.

[Ger03] P. Gerhardy. Refined Complexity Analysis of Cut Elimination. In Matthias Baaz and Johann Makovsky, editors, *Proceedings of the 17th International Workshop CSL 2003*, volume 2803 of *LNCS*, pages 212–225. Springer-Verlag, Berlin, 2003.

[GJ04] E. Goris and J.J. Joosten. Modal matters in interpretability logics. Logic Group Preprint Series 226, University of Utrecht, March 2004.

[Göd31] K. Gödel. Über formal unentscheidbare Sätze der Principia Mathematica und verwnadter Systeme I. *Monatsh. Math. Physik*, 38:173–198, 1931.

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

[Gor03] E. Goris. Extending ILM with an operator for Σ_1–ness. Illc prepublication series, University of Amsterdam, 2003.

[Gre96] M.J. Greenberg. *Euclidean and Non-Euclidean Geometries, 3d edition.* Freeman, 1996.

[Gua79] D. Guaspari. Partially conservative sentences and interpretability. *Transactions of AMS*, 254:47–68, 1979.

[Gua83] D. Guaspari. Sentences implying their own provability. *Journal of Symbolic Logic*, 48:777–789, 1983.

[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.

[Háj81] P. Hájek. Interpretability in theories containing arithmetic II. *Comm. Math. Univ. Carolinae*, 22:667–688, 1981.

[Han65] W. Hanf. Model-theoretic methods in the study of elementary logic. In J.W. Addison, L. Henkin, and A. Tarski, editors, *The Theory of Models, Proceedings of the 1963 International Symposium at Berkeley*, pages 132–145. North Holland, Amsterdam, 1965.

[HB68] D. Hilbert and P. Bernays. *Grundlagen der Mathematik, Vols. I and II, 2d ed.* Springer-Verlag, Berlin, 1968.

[HH02] R. Hirsch and I. Hodkinson. *Relation Algebras by Games*, volume 147 of *Studies in Logic.* Elsevier, North-Holland, 2002.

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

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

[HMV01] I. Hodkinson, S. Mikulás, and Y. Venema. Axiomatizing complex algebras by games. *Algebra Universalis*, 46:455–478, 2001.

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

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

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

[Ign93a] K.N. Ignatiev. On strong provability predicates and the associated modal logics. *Journal of Symbolic Logic*, 58:249–290, 1993.

[Ign93b] K.N. Ignatiev. The provability logic of Σ_1-interpolability. *Annals of Pure and Applied Logic*, 64:1–25, 1993.

[Jap94] G.K. Japaridze. The logic of the arithmetical hiearchy. *Annals of Pure and Applied Logic*, 66:89–112, 1994.

[Joo98] J.J. Joosten. Towards the interpretability logic of all reasonable arithmetical theories. Master's thesis, University of Amsterdam, 1998.

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

[Joo03a] J.J. Joosten. The closed fragment of the interpretability logic of PRA with a constant for $\mathrm{I}\Sigma_1$. Logic Group Preprint Series 128, University of Utrecht, February 2003.

[Joo03b] J.J. Joosten. Formalized interpretability in primitive recursive arithmetic. In *Proceedings of the ESSlLI student session*, 2003.

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

[JV04a] J.J. Joosten and A Visser. Characterizations of interpretabilty. Logic group preprint series, University of Utrecht, 2004.

[JV04b] J.J. Joosten and A Visser. How to derive principles of interpretability logic, A toolkit. To appear, 2004.

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

[Ken73] C.F Kent. The relation of A to $Prov\ulcorner !A\urcorner$ in the Lindenbaum sentence algebra. *Journal of Symbolic Logic*, 38:359–367, 1973.

[KL68] G. Kreisel and A. Lévy. Reflection principles and their use for establishing the complexity of axiomatic systems. *Zeitschrift f. math. Logik und Grundlagen d. Math.*, 14:97–142, 1968.

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

[Lin79] P. Lindström. Some results on interpretability. In *Proceedings of the 5th Scandinavian Logic Symposium*, pages 329–361. Aalborg University press, 1979.

[Lin84] P. Lindström. On partially conservative sentences and interpretability. *Proceedings of the AMS*, 91(3):436–443, 1984.

[Min72] G Mints. Quantifier-free and one-quantifier systems. *Journal of Soviet Mathematics*, 1:71–84, 1972. First published in Russian in 1971.

[Mon58] R. Montague. The continuum of relative interpretability types. *Journal of Symbolic Logic*, 23, 1958.

[MPS90] J. Mycielski, P. Pudlák, and A.S. Stern. *A lattice of chapters of mathematics (interpretations between theorems)*, volume 426 of *Memoirs of the American Mathematical Society.* AMS, Providence, Rhode Island, 1990.

[Myc77] J. Mycielcski. A lattice of interpretability types of theories. *Journal of Symbolic Logic*, 42:297–305, 1977.

[Nel86] E. Nelson. *Predicative arithmetic.* Princeton University Press, Princeton, 1986.

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

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

[Par71] R. Parikh. Existence and feasibility in arithmetic. *Journal of Symbolic Logic*, 36:494–508, 1971.

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

[Pet90] P.P. Petkov, editor. *Mathematical logic, Proceedings of the Heyting 1988 summer school in Varna, Bulgaria.* Plenum Press, Boston, 1990.

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

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

[Sch77] H. Schwichtenberg. Some applications of cut-elimination. In J. Barwise, editor, *Handbook of Mathematical Logic*, pages 867–896. North Holland, Amsterdam, 1977.

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

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

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

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

[Sie91] W. Sieg. Herbrand analyses. *Archive for Mathematical Logic*, 30:409–441, 1991.

[Sim88] S. G. Simpson. Partial realizations of Hilbert's program. *Journal of Symbolic Logic*, 53:349–363, 1988.

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

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

[Sla04] T Slaman. Σ_n-bounding and Δ_n-induction. *Proceedings of the AMS*, 132:2449–2456, 2004.

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

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

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

[Šve78] V. Švejdar. Degrees of interpretability. *Commentationes Mathematicae Universitatis Carolinae*, 19:789–813, 1978.

[Šve83] V. Švejdar. Modal analysis of generalized Rosser sentences. *Journal of Symbolic Logic*, 48:986–999, 1983.

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

[Tak75] G. Takeuti. *Proof Theory.* North–Holland, Amsterdam, 1975.

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

[Vis88] A. Visser. Preliminary notes on interpretability logic. Technical Report LGPS 29, Department of Philosophy, Utrecht University, 1988.

[Vis90a] A. Visser. Interpretability logic. In *[Pet90]*, pages 175–209, 1990.

[Vis90b] A. Visser. Notes on IΣ_1. Unpublished manuscript, 1990?

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

[Vis92a] A. Visser. An inside view of exp. *Journal of Symbolic Logic*, 57:131–165, 1992.

[Vis92b] A. Visser. An inside view of EXP. the closed fragment of the provability logic of $I\Delta_0 + \Omega_1$ with a propositional constant for EXP. *Journal of Symbolic Logic*, 57:131–165, 1992.

[Vis93] A. Visser. The unprovability of small inconsistency. *Archive for Mathematical Logic*, 32:275–298, 1993.

[Vis95] A. Visser. A course on bimodal provability logic. *Annals of Pure and Applied Logic*, pages 109–142, 1995.

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

[Vis02] A. Visser. Faith & Falsity: a study of faithful interpretations and false Σ^0_1-sentences. Logic Group Preprint Series 216, Department of Philosophy, Utrecht University, Heidelberglaan 8, 3584 CS Utrecht, October 2002.

[Wan51] H Wang. Arithmetical models of formal systems. *Methodos 3*, pages 217–232, 1951.

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

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

[Zam96] D Zambella. Notes on polynomial bounded arithmetic. *Journal of Symbolic Logic*, 61:942–966, 1996.

Index

List of Symbols

Curriculum Vitae

10-10-1972	Geboren, Joost Johannes Joosten, te Diemen
1984-1990	VWO, Montessori Lyceum te Amsterdam
1990-1994	Student natuurkunde
1990-1998	Student wiskunde
1992	Propedeutisch examen natuurkunde
1992	Propedeutisch examen wiskunde
1995-1996	Erasmusstudent, te Siena, Italië
1998	Doctoraalexamen wiskunde (Cum Laude)
1999	Docent wiskunde aan het *Colegio Internacional Costa Blanca*, Spanje
1999-2004	Promovendus van de vakgroep Theoretische Filosofie van de Faculteit der Wijsbegeerte van de Universiteit Utrecht

Samenvatting

Deze dissertatie is in de eerste plaats een verhandeling over wiskundige interpretaties. Hierbij worden interpretaties zelf onderzocht, maar ook worden zij gebruikt als hulpmiddel bij de bestudering van formele theorieën.

Vergelijken met behulp van interpretaties zegt op natuurlijke wijze iets over de (bewijs) sterkte van formele theorieën. Immers, wat zou het kunnen betekenen dat een theorie S minstens zo sterk is als een theorie T?

Als eerste en meest eenvoudige uitleg kan worden gegeven dat S alles bewijst wat ook door T bewezen wordt. In symbolen schrijft men

$$\forall \varphi \ (T \vdash \varphi \Rightarrow S \vdash \varphi).$$

Een directe complicatie doet zich hier voor als T en S 'verschillende talen spreken'. Onmiddellijk dient zich nu het idee van een vertaling aan om tot de volgende uitleg van "S is minstens zo sterk als T" te komen.

Voor een zekere vertaling j moet S alle vertaalde stellingen van T bewijzen. Dit zou in symbolen als volgt kunnen worden weergegeven.

$$\exists j \, \forall \varphi \ (T \vdash \varphi \Rightarrow S \vdash \varphi^j) \tag{12.9}$$

In grote lijnen is een dergelijke vertaling j hetzelfde als een interpretatie van T in S. Dat wil zeggen, via (12.9) is de notie van interpretatie gedefinieerd. Opdat de notie van interpreteerbaarheid daadwerkelijk informatief is, zijn er nog enkele restricties aan j opgelegd. Zo zal j bijvoorbeeld zekere structuur moeten behouden waardoor alles naar een trivialiteit vertalen, zeg $\forall x \ (x = x)$, uitgesloten wordt.

Het wordt nu ook al snel duidelijk wat de verdiensten van interpretaties voor de grondslagen van de wiskunde zijn. Sinds de onvolledigheidsstellingen van Gödel is het bekend dat er niet zo iets kan zijn als een volledige axiomatisering van de wiskunde. Formele wiskundige systemen kunnen dus niet op natuurlijke wijze worden ingebed in één universeel formeel systeem. De vraag is dan, hoe verschillende formele systemen zich tot elkaar verhouden, hoe zij vergeleken kunnen worden. Interpretaties bieden hier een mogelijke uitkomst.

In dit proefschrift worden interpretaties tussen eerste orde theorieën met een zekere minimale rekenkracht bestudeerd. Zoals gezegd, worden in het proefschrift interpretaties gebruikt om theorieën te vergelijken, maar ook zijn zij zelf het onderwerp van studie. In dit laatste geval ligt de nadruk op het structurele gedrag van interpreteerbaarheid, hetgeen zich onder andere in zogeheten *interpreteerbaarheidslogicas* manifesteert. Het proefschrift valt op natuurlijke wijze uiteen in drie delen.

Deel één In het eerste deel wordt de notie van interpretatie geïntroduceerd en vergeleken met andere noties met zekere meta-mathematische importantie. Dit resulteert in de zogeheten karakteriseringen van interpreteerbaarheid, waarbij consistentie uitspraken, definieerbare snedes en Π_1-conservativiteit centrale begrippen zijn.

Interpreteerbaarheid, als zijnde een zuiver syntactische notie, wordt geformaliseerd en ook de karakteriseringen vinden in een volledig geformaliseerde omgeving plaats. Hierbij wordt voor elke implicatie in de karakterisering nauwkeurig bijgehouden welke principes uit de meta-theorie worden gebruikt. De karakteriseringen krijgen een bijzonder elegante vorm als zij in termen van categorieëntheorie worden uitgedrukt.

Aan het eind van het eerste deel wordt de focus verlegd naar interpreteerbaarheidslogicas en in het bijzonder wordt gesproken over een modale karakterisering van $\mathbf{IL}(\mathrm{All})$, de interpreteerbaarheidslogca van alle redelijke rekenkundige theorieën . Er wordt een nieuw geldig principe R voor deze logica gepresenteerd en dit principe wordt aritmetisch correct bewezen. De correctheid wordt op twee manieren bewezen. Er worden modale systemen gepresenteerd die met deze twee bewijsmethoden corresponderen. Alle tot dusver bekende andere principes in $\mathbf{IL}(\mathrm{All})$ worden in beide modale systemen aritmetisch correct bewezen.

Deel twee Het tweede deel van het proefschrift is volledig gewijd aan modale semantiek van interpreteerbaarheidslogicas. Een centrale vraag is hier of een logica volledig is ten opzichte van haar modale semantiek. Modale volledigheidsbewijzen worden gegeven voor de logica's $\mathbf{IL}$, $\mathbf{IL}\mathsf{M}$, $\mathbf{IL}\mathsf{M}_0$, $\mathbf{IL}\mathsf{W}$ en $\mathbf{IL}\mathsf{W}^*$. De volledigheidsbewijzen voor $\mathbf{IL}\mathsf{M}_0$ en $\mathbf{IL}\mathsf{W}^*$ kunnen als eerste bewijzen worden aangemerkt. Ook zijn er enige toepassingen van de volledigheidsbewijzen.

Er wordt een poging gedaan een soort uniformiteit in volledigheidsbewijzen aan te brengen. Echter, hier valt zeker nog werk te verrichten. Een stap in de goede richting is de invoering van de *full labels* en de ontwikkeling van de theorie hiervan.

In het laatste hoofdtuk wordt een modaal onvolledigheidsbewijs gegeven van $\mathbf{IL}\mathsf{P}_0\mathsf{W}^*$. In dit bewijs speelt het principe R een centrale rol.

Deel drie In het derde en laatste deel van het proefschrift wordt een studie naar primitief recursieve rekenkunde (PRA) uitgevoerd. Met name de verhouding tot $\mathrm{I}\Sigma_1$ wordt bekeken. De stelling van Parsons wordt op twee verschillende manieren bewezen. Eén bewijs is model-theoretisch van aard en geeft inzicht in bewijsbare geslotenheidseigenschappen van de bewijsbaar totaal recursieve functies.

Hoewel PRA en IΣ_1 equi-consistent over PRA zijn, is het wel mogelijk om in IΣ_1 de consistentie van PRA op een definieerbare snede te bewijzen.

Een belangrijk probleem is de modale karakterisering van **IL**(PRA), de interpreteerbaarheidslogica van PRA. Er wordt een frame conditie voor Beklemishev's principe berekend. Ook wordt laten zien dat Beklemishev's principe minstens zo sterk is als het principe van Zambella.

Er wordt een karakterisering gegeven van het gesloten fragment van de interpreteerbaarheidslogica van PRA met een constante voor IΣ_1. Ook wordt een grove bovengrens voor **IL**(PRA) gegeven. Deze bovengrens wordt berekend door de mogelijke substituties in Solovay's stelling te beperken.

De toegift van het proefschrift behandelt modale semantiek voor het gesloten fragment van **GLP**. Er wordt een kleine variatie op het model van Ignatiev gegeven, een model met 'diepte ϵ_0'. Dit model wordt aan een modale analyse onderworpen zonder dat er verdere aritmetische ingrediënten aan te pas komen.