How to Derive Principles of Interpretability Logic
A Toolkit

Joost J. Joosten
and
Albert Visser

September 4, 2004

Abstract

In this paper, we develop a toolkit to derive principles for the interpetability logic of all reasonable theories by fine-tuning **ILM** and **ILP** proofs.

1 Introduction

1.1 Preludium

When, around 1987, the project of Interpetability Logic started, Dick de Jongh, in collaboration with Frank Veltman, took responsibility for the modal part of the project. The first fruit of this research was the fundamental paper [dJV90]. From this time on, Dick actively worked on the subject. His work can be found in such papers as [dJV91], [AdJH98], [dJV99], [dJJ98]. We always enjoyed his remarkable style of doing modal logic, where visualization of models plays such an important role.

Dick was not so active on the arithmetical side of the project. For this reason, it seemed a good idea to show that arithmetical soundness proofs can be given using modal methods. Perhaps, this will draw him into arithmetical matters again.

In this paper, we use enriched versions of interpretability logic to verify arithmetically sound principles of ordinary interpretability logic. The general methodology of using modal logic to study modal logic has always been part of the strongly self-reflexive provability logic tradition. We just remind the reader of the seminal paper [JJM91] in which the Solovay result itself is proved in a modal logic.

The logics we study arise from well known logics by 'diversifying' the modal operators. In another contribution to this Liber, Johan van Benthem discusses the idea of scattering. It seems to us that the auxiliary logics presented in this paper are typical examples of such scattered logics. This either proves the coherence of the Dutch tradition or the naturalness of the idea.

1.2 Contents of the paper

In this paper we present and explore a remarkable methodology. We want to derive principles for he interpetability logic of all reasonable theories. This logic is a sublogic both of **IL**M and **IL**P, but it is not the intersection of **IL**M and **IL**P. (See [Vis97].) It turns out that one may derive principles for the logic of all theories by fine-tuning proofs in **IL**M or **IL**P. In fact, *every principle we know* can be derived *both* by fine-tuning an **IL**M and an **IL**P proof.

The fine-tuning procedure is best mediated by auxiliary modal logics that we present in Section 2 (for the **IL**M-case) and in Section 3 (for the **IL**P-case). In Section 4, we put the methods developed in action, and show how to derive a number of principles.

The phenomenon of having two different proofs for the same theorem is always strange. In some cases different proofs reflect different underlying concepts (like the two proofs of the commutativity of addition). In our case the strangeness is increased: we do not just have two proofs, but two natural classes of proofs for the same theorems. No different underlying concepts are in sight. The only clear difference is a difference in *scope*: the M-style proofs use sequentiality and the P-style proofs do not.

The explicit development of modal logics to analyze the methods known so far to derive interpretability principles, has the following aim. We hope that it will enable us to pin down precisely what principles can be derived using M-style methods and which ones in the P-style. The question which interpretability principles can be derived in P-style and which in M-style becomes a modal question open to study with Kripke models. Apart from this aim, we submitt that we found exciting and intriguing modal systems.

At the present stage, it is not yet fully clear that our extended systems are definitive. Do we have the right notions? Did we articulate all possible principles for the chosen notions? Do we have to extend or restrict the expressive power? More experimentation is necessary.

1.3 Convention

Interpretability will in this paper be *theorems interpretability*, i.o.w.

- $k : U \rhd V :\iff \forall \phi \, (\Box_V \phi \to \Box_U \phi^k)$.

2 A logic for relativization to cuts

In this section, we present a logic that incorporates a number of principles concerning provability predicates relativized to definable cuts. This logic will enable us to fine-tune **IL**M-proofs.

In the present section, theories will be Δ^b_1-axiomatized *sequential* theories. We assume that every theory comes equiped with a designated interpretation of $I\Delta_0 + \Omega_1$. Quantifiers will range over the numbers given by this interpretation. Arithmetized concepts will be implicitly relativized to this interpretation. For example, suppose our theory is ZF and neumann

is the von Neumann interpretation of number theory in ZF, then

$$\mathsf{ZF} \vdash \forall x, y, z\ (x + y = z \to \Box_{\mathsf{ZF}} x + y = z)$$

means:

$$\mathsf{ZF} \vdash (\forall x, y, z\ (x + y = z \to \Box_{\mathsf{ZF}} (x + y = z)^{\mathsf{neumann}}))^{\mathsf{neumann}}.$$

We specify the promised logic, which we will call CuL, and its arithmetical semantics. The logic will have two levels, the inner and the outer. We specify the language. We have a set of propositonal variables $p_0, p_1, \ldots$. The meta-variables $p, q, r, \ldots$ will range over the propositional variables. We have a set of cut-variables, $I_0, I_1, \ldots$. We have one cut-constant id. The meta-variables $I, J, J', \ldots$ will range over the cut-variables and id. *Outer formulas* are the smallest class containing the propositional variables, closed under the formation rules corresponding to the propositional connectives (incluiding $\bot$ and $\top$) and closed under the rule:

- if A and B are outer formulas, then so are $\Box^I A$ and $A \rhd B$.

We will write $\Box A$ for $\Box^{\mathsf{id}} A$. *Inner formulas* are the smallest set containing the propositional variables, closed under the formation rules corresponding to the propositional connectives, such that

- if if A and B are outer formulas, then $\Box A$ and $A \rhd B$ are inner formulas.

Our logic is specified as an extension of a suitable sequent system for propositional logic for the modal language. The following rules are shared by the inner and the outer system, but for the fact that, in the inner system, formulas are constrained to be inner fomulas. Concretely, this means that, for the inner system, we must delete the letter I in the statement of the principles, as formulated below.

$$\begin{array}{ll}
(\to)^{\mathsf{J}} & \vdash \Box^I A \to \Box A \\
\mathsf{L}_1^{\mathsf{J}} & \vdash \Box^I (A \to B) \to (\Box^I A \to \Box^I B) \\
\mathsf{L}_2^{\mathsf{J}} & \vdash \Box^I A \to \Box^I \Box^J A \\
\mathsf{L}_3^{\mathsf{J}} & \vdash \Box^I (\Box^J A \to A) \to \Box^I A \\
\mathsf{J}_1^{\mathsf{J}} & \vdash \Box (A \to B) \to A \rhd B \\
\mathsf{J}_2^{\mathsf{J}} & \vdash (A \rhd B) \wedge (B \rhd C) \to A \rhd C \\
\mathsf{J}_3^{\mathsf{J}} & \vdash (A \rhd C) \wedge (B \rhd C) \to A \vee B \rhd C \\
\mathsf{J}_4^{\mathsf{J}} & \vdash A \rhd B \to (\Diamond A \to \Diamond B) \\
\mathsf{J}_5^{\mathsf{J}} & \vdash \Diamond^J A \rhd A \\
\mathsf{Nec}^{\mathsf{J}} & \vdash A \Rightarrow \vdash \Box^I A
\end{array}$$

By a well-known trick, we can derive $\mathsf{L}_2^{\mathsf{J}}$ from $\mathsf{L}_3^{\mathsf{J}}$. Moreover, we can derive $\mathsf{L}_{2,\mathsf{in}}^{\mathsf{J}}$ also from $\mathsf{J}_5^{\mathsf{J}}$ and $\mathsf{J}_4^{\mathsf{J}}$. The further rules are as follows.

$$\begin{array}{ll}
\mathsf{M}^{\mathsf{J}} & \Gamma, (A \wedge \Box^J C \rhd B \wedge \Box^{J'} C) \vdash_{\mathsf{in}} D \Rightarrow \Gamma, A \rhd B \vdash_{\mathsf{in}} D \\
 & \text{Here } J \text{ must be a variable and } J \neq J' \text{ and} \\
 & J \text{ does not occur in } \Gamma, A, B, D \\
\mathsf{IO}^{\mathsf{J}} & \Gamma \vdash_{\mathsf{in}} A \Rightarrow \Gamma \vdash_{\mathsf{out}} A \\
\mathsf{OI}^{\mathsf{J}} & \vdash_{\mathsf{out}} A \Rightarrow \vdash_{\mathsf{in}} \Box A
\end{array}$$

Note that $\mathsf{Nec}^{\mathsf{J}}_{\mathsf{in}}$ follows from IO^{J} and OI^{J}.[1]

Inspection of the verifications of interpretability principles in Section 4, reveals that we only use one cut-variable. So the restriction of our system to one variable, say I_0, is of some interest. Let us write ΔA for $\Box^{I_0} A$ and ∇A for $\neg\Delta\neg A$. We get the following bimodal system.[2] (We omitt some superfluous principles.)

$(\rightarrow)^{\mathsf{J},1}$	$\vdash_{\mathsf{out}} \Delta A \rightarrow \Box A$
$\mathsf{L}_1{}^{\mathsf{J},1}$	$\vdash \Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B)$
	$\vdash_{\mathsf{out}} \Delta(A \rightarrow B) \rightarrow (\Delta A \rightarrow \Delta B)$
$\mathsf{L}_2{}^{\mathsf{J},1}$	$\vdash \Box A \rightarrow \Box\Delta A$
	$\vdash_{\mathsf{out}} \Delta A \rightarrow \Delta\Delta A$
$\mathsf{L}_3{}^{\mathsf{J},1}$	$\vdash \Box(\Delta A \rightarrow A) \rightarrow \Box A$
	$\vdash_{\mathsf{out}} \Delta(\Delta A \rightarrow A) \rightarrow \Delta A$
$\mathsf{J}_1{}^{\mathsf{J},1}$	$\vdash \Box(A \rightarrow B) \rightarrow A \rhd B$
$\mathsf{J}_2{}^{\mathsf{J},1}$	$\vdash (A \rhd B) \wedge (B \rhd C) \rightarrow A \rhd C$
$\mathsf{J}_3{}^{\mathsf{J},1}$	$\vdash (A \rhd C) \wedge (B \rhd C) \rightarrow A \vee B \rhd C$
$\mathsf{J}_4{}^{\mathsf{J},1}$	$\vdash A \rhd B \rightarrow (\Diamond A \rightarrow \Diamond B)$
$\mathsf{J}_5{}^{\mathsf{J},1}$	$\vdash \nabla A \rhd A$
$\mathsf{M}^{\mathsf{J},1}$	$\Gamma, (A \wedge \Delta C \rhd B \wedge \Box C) \vdash_{\mathsf{in}} D \Rightarrow \Gamma, A \rhd B \vdash_{\mathsf{in}} D$
	Γ, A, B, D must be Δ-free
$\mathsf{Nec}^{\mathsf{J},1}$	$\vdash_{\mathsf{out}} A \Rightarrow \vdash_{\mathsf{out}} \Delta A$
$\mathsf{IO}^{\mathsf{J},1}$	$\Gamma \vdash_{\mathsf{in}} A \Rightarrow \Gamma \vdash_{\mathsf{out}} A$
$\mathsf{OI}^{\mathsf{J},1}$	$\vdash_{\mathsf{out}} A \Rightarrow \vdash_{\mathsf{in}} \Box A$

We turn to the 'arithmetical semantics'. Consider a sequential theory T and let N be the designated interpetation of $I\Delta_0 + \Omega_1$. For any formula α of the language of T having at most x free, we define:

- $\mathsf{cut}(\alpha)$ is the formula that expresses that α is an N-cut, i.e. that $\{x \mid \alpha\}$ is a subset of $\{x \mid \delta_N\}$ and that $\{x \mid \alpha\}$ is closed under the operations $0, \mathsf{S}, +, \times, \omega_1$ and that $\{x \mid \alpha\}$ is downwards closed under $<$.
- $\mathsf{c}(\alpha) := (\delta_N \wedge (\mathsf{cut}(\alpha) \rightarrow \alpha))$.

It is easy to see that $T \vdash \mathsf{cut}(\mathsf{c}(\alpha))$ and $T \vdash \mathsf{cut}(\alpha) \rightarrow (\alpha \leftrightarrow \mathsf{c}(\alpha))$. so if α ranges over all formulas having at most x free, then $\mathsf{c}(\alpha)$ ranges modulo T-provable equivalence over all T-cuts.

Arithmetical realizations in a theory T are given by pairs σ, τ, where σ sends the propositional variables to sentences of the language of T and τ sends the cut-variables to formulas α of the language of T having only x free. We demand that σ and τ take for all but finitely many arguments

[1] A possible strengthening of the principle M^{J} that we will not further explore in this paper is:

$\mathsf{M}^{\mathsf{J}+}$ $\Gamma, \{(A_i \wedge \Box^J C_i \rhd B_i \wedge \Box^{J'_i} C_i) \mid i < n\} \vdash_{\mathsf{in}} D \Rightarrow \Gamma, \{(A_i \rhd B_i) \mid i < n\} \vdash_{\mathsf{in}} D$

Here J must be a variable and $J \notin \{J'_0, \ldots, J'_{n-1}\}$ and J may not occur in $\Gamma, A_0, \ldots, A_{n-1}, B_0, \ldots, B_{n-1}, D$

Alternatively, we could add an operation $\cap$ on cuts and formulate the appropriate axiom.

[2] Remember that $\Box$ is definable from $\rhd$.

the value $\top$ resp. $x = x$, thus making σ and τ into finite objects. The assignments σ and τ are extended to the full language in the usual way. The main novelty is that we take, for a cut-variable I,

- $(\Box^I A)^{\sigma,\tau} := \Box_T^{\mathsf{c}(\tau(I))} A^{\sigma,\tau}$.

The mapping $A \mapsto A^{\sigma,\tau}$ can be arithmetized. A consequence of this is that we can treat τ in the interpretation of *inner formulas* as a *numerical variable*. This observation makes the definition of *inner validity* below sensible. We define:

- $\Gamma \models_{T,\mathsf{out}} A :\Leftrightarrow \forall \sigma, \tau \;\; T, \Gamma^{\sigma,\tau} \vdash A^{\sigma,\tau}$,
- $\Gamma \models_{T,\mathsf{in}} A :\Leftrightarrow \forall \sigma \;\; T \vdash \forall \tau \; (\bigwedge \Gamma^{\sigma,\tau} \to A^{\sigma,\tau})$.

It is not hard to see that our principles do indeed hold w.r.t. the intended notions of validity and that our rules are admissible. We will briefly consider each prinple. The principle $(\to)^{\mathsf{J}}$ is a triviality. The principle $\mathsf{L}_1^{\mathsf{J}}$ reflects that concatenation of proofs in a cut, remains within this cut as concatenation is approximately multiplication. The principle $\mathsf{L}_2^{\mathsf{J}}$ is Lemma 2.1. The principle $\mathsf{L}_3^{\mathsf{J}}$ is Löb's theorem with cuts, as proved in Lemma 2.2. The principle $\mathsf{J}_5^{\mathsf{J}}$ follows from the formalized Henkin theorem. The rule $\mathsf{Nec}_{\mathsf{out}}^{\mathsf{J}}$ follows because, if we have a proof of $A^{\sigma,\tau}$, then this proof is standard and, hence, T-provably in every T-cut. The rule M^{J} follows from Lemma 2.3. The rule IO^{J} is trivial. Since the verification of the validity of $\vdash_{\mathsf{in}}$ can be formalized in $I\Delta_0 + \Omega_1$, we have OI^{J}.

We write Form_U^1 for the set of U-formulas having at most x free.

Lemma 2.1. *Consider any theories U, V with designated natural numbers. (U and V need not be sequential.) We have, for any V-sentence α,*

$$I\Delta_0 + \Omega_1 \vdash \forall \beta \in \mathsf{Form}_V^1 \; (\Box_U \alpha \to \Box_V \Box_U^{\mathsf{c}(\beta)} \alpha).$$

Proof. Reason in $I\Delta_0 + \Omega_1$. We assume that $\mathsf{proof}_U(p, \alpha)$ for some p. Since the sentence $\mathsf{proof}_U(p, \alpha)$ is in $\exists\Sigma_1^b$ we get, by verifiable $\exists\Sigma_1^b$-completeness, that, for some p', $\mathsf{proof}_V(p', \mathsf{proof}_U(p, \alpha))$. By the OBIS-principle[3], we have $\Box_V(p \in \mathsf{c}(\beta))$. Ergo $\Box_V \Box_U^{\mathsf{c}(\beta)} \alpha$. $\dashv$

We note that $\mathsf{L}_{2,\mathsf{out}}^{\mathsf{J}}$, follows because we have $I\Delta_0 + \Omega_1$ on T-cut I. We take $U := T$ and $V := T$ and we specialize β to the standard formula for J, noting that its standard code will be T-provably in I. We get $\mathsf{L}_{2,\mathsf{in}}^{\mathsf{J}}$ by noting that we have $I\Delta_0 + \Omega_1$ in N, the designated numbers of T.

Lemma 2.2. *Consider any theory U with designated natural numbers. (U need not be sequential.) We have, for any U-sentence α,*

$$I\Delta_0 + \Omega_1 \vdash \forall \beta \in \mathsf{Form}_U^1 \; (\Box_U(\Box_U^{\mathsf{c}(\beta)} \alpha \to \alpha) \to \Box_U \alpha.$$

Proof. We can do this in two ways. We can simply adapt the proof of Löb's Theorem, or, alternatively, we can use Löb's Theorem. We follow

[3]OBIS stands for *Outside Big Inside Small*. The proof of the OBIS-principle consists in a p-time transformation of a proof q of $\mathsf{cut}(\mathsf{c}(\beta))$ and of p into a proof of $(p \in \mathsf{c}(\beta))$. We note that the possible 'non-standardness' of β does not matter here.

the second rooute here. Reason in $I\Delta_0 + \Omega_1$. Fix $\beta \in \mathsf{Form}^1_U$. We have:

$$\begin{array}{rcl} \Box_U(\Box_U^{\mathsf{c}(\beta)}\alpha \to \alpha) & \to & \Box_U\Box_U^{\mathsf{c}(\beta)}(\Box_U^{\mathsf{c}(\beta)}\alpha \to \alpha) \\ & \to & \Box_U\Box_U^{\mathsf{c}(\beta)}\alpha \\ & \to & \Box_U\alpha. \end{array}$$

$\dashv$

The derivation of $\mathsf{L}^{\mathsf{J}}_{3,\mathsf{out}}$ and $\mathsf{L}^{\mathsf{J}}_{3,\mathsf{in}}$ is immediate. Inspecting the proof of Lemma 2.2, we see that $\mathsf{L}^{\mathsf{J}}_{3,\mathsf{in}}$ follows from $\mathsf{L}^{\mathsf{J}}_{3,\mathsf{out}}$ plus the other principles and rules.

Lemma 2.3. *Suppose U and V are theories with designated numbers. Suppose further that U is sequential. For any Σ^1_0-sentence σ, we have:*

$$I\Delta_0 + \Omega_1 \vdash \forall\beta \in \mathsf{Form}^1_V\ (U \rhd V \to \exists\gamma \in \mathsf{Form}^1_U\ (U + \sigma^{\mathsf{c}(\gamma)}) \rhd (V + \sigma^{\mathsf{c}(\beta)})).$$

(Note our slightly sloppy notation: the function c depends on the designated natural numbers. So, it is not the same in both occurrences.)

Proof. This is a direct consequence of Pudlák's Lemma. We proceed to reason in $I\Delta_0 + \Omega_1$ and suppose $j : U \rhd V$ and $\beta \in \mathsf{Form}^1_V$. We change the designated numbers of V to those given by $\mathsf{c}(\beta)$. Pudlák's Lemma provides us with a U-cut $\mathsf{c}(\gamma)$ and an isomorphism h between $\mathsf{c}(\gamma)$ and an initial segment of $\mathsf{c}(\beta)$. Now, let σ be $\exists y\, \sigma_0(y)$ with $\sigma_0 \in \Delta_0$. We get that $\Box_U \forall\, y{\in}\mathsf{c}(\gamma)\ (\sigma_0(y) \leftrightarrow \sigma_0^{j\circ\mathsf{c}(\beta)}(h(y)))$ and thus certainly

$$\Box_U(\sigma^{\mathsf{c}(\gamma)} \to \sigma^{j\circ\mathsf{c}(\beta)}). \tag{1}$$

The desired fact is now immediate. $\dashv$

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

Lemma 2.4. *We have:* $\mathsf{CuL} \vdash_{\mathsf{in}} A \equiv A \wedge \Box^I\neg A \equiv A \vee \Diamond^J A$.

Proof. Just copy the proofs from **IL**, replacing some regular principles with the new principles relativized to a cut. Note that for the first mutual interpretability, we do need to detour via the outer system. $\dashv$

Lemma 2.5. $\mathsf{CuL} \vdash_{\mathsf{in}} \neg(A \rhd \neg C) \to \Diamond(A \wedge \Box^J C)$.

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

3 The approximating theory

In this section we develop a logic for fine-tuning **ILP**-proofs. Theories in this section will be Δ^b_1-axiomatized theories with a designated interpretation of S^1_2. (Since S^1_2 is finitely axiomatizable, it is in this context more convenient to use than $I\Delta_0 + \Omega_1$. Of course, nothing substantial is entailed by this choice.)

3.1 The approximating theory defined

For finitely axiomatized theories V, we have: $\mathsf{S}^1_2 \vdash U \rhd V \to \Box_{\mathsf{S}^1_2}(U \rhd V)$, because $U \rhd V$ is a $\exists\Sigma^b_1$-sentence. To mimic the P-style behavior for an arbitrary theory V, we will modify V to a new theory V' approximating V to obtain $\vdash U \rhd V \to \Box_{\mathsf{S}^1_2}(U \rhd V')$. Of course, V' should be sufficiently like V to be useful. Thus, we define a theory V' that is extensionally the same as V, but for which $U \rhd V'$ is so simple that we can easily infer $\Box_{\mathsf{S}^1_2}(U \rhd V')$.

The idea is as follows. Let us *define* the set of axioms V' as consisting of just those axioms ϕ of V such that $U \vdash \phi^k$. Note that, if $k : U \rhd V$, then V and V' have the same axioms. However, we cannot take this insight with us when we proceed to reason inside a box. This idea works modulo some trifling details. Firstly, the definition of the new axiom set does not have the right complexity. Secondly, if the argument is not set up in a careful way, we may seem to need both Σ_1-collection and exp. We shall use a variation of Craig's trick so that our axiom set will be Δ^b_1-definable. The same trick makes the use of strong pinciples, like Σ_1-collection and exp, superfluous.

Definition 3.1. Let k be a translation of the right type. We define $V^{[k]}$ as follows.

$$\begin{aligned}\mathsf{axioms}_{V^{[k]}}(x) \quad := \quad & \exists p\ (x = \ulcorner \varphi \wedge (\underline{p} = \underline{p}) \urcorner \wedge \\ & \qquad \mathsf{axioms}_V(\varphi) \wedge \mathsf{proof}_U(p, \varphi^k)).\end{aligned}$$

It is clear that $\mathsf{axioms}_{V^{[k]}}(x)$ is in poly-time decidable if $\mathsf{axioms}_V(x)$ and $\mathsf{axioms}_U(x)$ are. Note that it is essential here that we work with efficient numerals $\underline{p}$.

Lemma 3.2.

1. $\mathsf{S}^1_2 \vdash \forall k\ (\mathsf{id} : V \rhd V^{[k]})$.
2. $\mathsf{S}^1_2 \vdash \forall k\ (k : U \rhd V \to \mathsf{id} : V^{[k]} \rhd V)$.

We see that S^1_2 verifies that $k : U \rhd V$ implies that V and $V^{[k]}$ are extensionally equal.

Proof. Ad (1). Reason in S^1_2. We have to show: $\Box_{V^{[k]}}\varphi \to \Box_V \varphi$. This is easily seen to be true, since we can replace every axiom $\varphi \wedge (\underline{p} = \underline{p})$ of $V^{[k]}$ by a proof of $\varphi \wedge (\underline{p} = \underline{p})$ from the V-axiom φ. The resulting transformation is clearly p-time.

Ad (2). Reason in S^1_2. Suppose $k : U \rhd V$ and $\Box_V \varphi$. We have a proof p of φ from axioms, $\tau_0, \ldots, \tau_n$. Let τ be the conjunction of these axioms. Note that τ is bounded by p. Since, clearly, $\Box_V \tau$, we may find, using $k : U \rhd V$, a U-proof q of τ^k. We may use q to obtain U-proofs of q_i of ${\tau_i}^k$. Clearly, $|q_i|$ is bounded by a term of order $|q|^2$. We can now replace every axiom occurrence of τ_i in p by

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

and obtain a $V^{[k]}$-proof r of φ. We find that $|r|$ is bounded by a term of order $|p| \cdot |q|^2$. So r can indeed be found in p-time from the given p and q. $\dashv$

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

Lemma 3.3. $\mathsf{S}^1_2 \vdash \forall k\ (k : U \rhd V^{[k]})$.

Proof. Reason in S^1_2. Suppose p is a $V^{[k]}$-proof of ϕ. We want to construct a U-proof of ϕ^k. As a first step we transform p into a V-proof p' as we did in the proof of Lemma 3.2,(1). Next we transform p', using k, into a predicate logical proof q of ϕ^k from assumptions τ^k, where τ is a V-axiom. It is well known that this transformation is p-time. Finally each axiom τ extracted from p, comes from a $V^{[k]}$-axiom $\tau \wedge (\underline{r} = \underline{r})$, where r is a U-proof of τ^k. So our final step is to extend q to a U-proof q' by prepending the U-proofs r above the corresponding τ^k. This extension will at most double the number of symbols of q, so $q' \approx q^2$. $\dashv$

3.2 A modal logic for approximation

We proceed to articulate modal principles reflecting facts about approximations. We will call our modal system AtL. We first specify the language. We have propositional variables $p_0, p_1, p_2 \ldots$, we will use $p, q, r, \ldots$, and we have interpretation variables $k_0, k_1, k_2, \ldots$. We have one interpretation constant id. the meta-variables $k, \ell, m, \ldots$ will range over the interpretation variables and id. The modal language is the smallest language containing the propositional variables, closed under the propositional connectives, including $\top$ and $\bot$, and closed under the following rule.

- If A, B are in the language and k is an interpretation term, then $\Box^{[k]}A$ and $A \rhd^{[k]} B$ are in the language.

We will write $\Box$ for $\Box^{[\mathsf{id}]}$ and $\rhd$ for $\rhd^{[\mathsf{id}]}$.

$(\to \Box)^{\mathsf{k}}$	$\vdash \Box^{[k]}A \to \Box A$
$(\to \rhd)^{\mathsf{k}}$	$\vdash A \rhd B \to A \rhd^{[k]} B$
$\mathsf{L}^{\mathsf{k}}_1$	$\vdash \Box^{[k]}(A \to B) \to (\Box^{[k]}A \to \Box^{[k]}B)$
$\mathsf{L}^{\mathsf{k}}_2$	$\vdash \Box^{[\ell]}A \to \Box^{[k]}\Box^{[\ell]}A$
$\mathsf{L}^{\mathsf{k}}_3$	$\vdash \Box^{[k]}(\Box^{[k]}A \to A) \to \Box^{[k]}A$
$\mathsf{J}^{\mathsf{k}}_1$	$\vdash \Box^{[k]}(A \to B) \to A \rhd^{[k]} B$
$\mathsf{J}^{\mathsf{k}}_2\mathsf{a}$	$\vdash (A \rhd B) \wedge (B \rhd^{[k]} C) \to A \rhd^{[k]} C$
$\mathsf{J}^{\mathsf{k}}_2\mathsf{b}$	$\vdash (A \rhd^{[k]} B) \wedge \Box^{[k]}(B \to C) \to A \rhd^{[k]} C$
$\mathsf{J}^{\mathsf{k}}_3$	$\vdash (A \rhd^{[k]} C) \wedge (B \rhd^{[k]} C) \to A \vee B \rhd^{[k]} C$
$\mathsf{J}^{\mathsf{k}}_4$	$\vdash A \rhd^{[k]} B \to (\Diamond A \to \Diamond^{[k]} B)$
$\mathsf{J}^{\mathsf{k}}_5$	$\vdash A \rhd^{[k]} \Diamond^{[n]} B \to A \rhd^{[n]} B$
P^{k}	$\Gamma, \Delta, \Box(A \rhd^{[k]} B) \vdash C \Rightarrow \Gamma, A \rhd B \vdash C$
$\mathsf{Nec}^{\mathsf{k}}$	$\vdash A \Rightarrow \vdash \Box^{[k]}A$

Here P^{k} is subject to the following conditions[4] :

[4] We realize that this formulation of P^{k} is not going to win a beauty contest. However, the primary focus of this paper is to formulate systems that (i) are arithmetically correct

1. k is an interpretation variable;
2. k does not occur in Γ, A, B, C;
3. Δ consists of formulas of the form $(E \rhd^{[k]} F \to E \rhd F)$ and $(\Box E \to \Box^{[k]} E)$.

We will call the licence to use $(\Box E \to \Box^{[k]} E)$ provided by P^{k}: $(\mathsf{E}\Box)^{\mathsf{k}}$, and we will call the licence to use $(E \rhd^{[k]} F \to E \rhd F)$: $(\mathsf{E}\rhd)^{\mathsf{k}}$.[5] In our verification of interpretability principles in Section 4, we use only one interpretation variable, say k_0. Thus, it could be a good idea to study the bimodal system obtained by restricting AtL to one variable. We leave the obvious formulation of the bimodal case to the reader.

We proceed to formulate the desired notion of arithmetical validity. Let T be any theory with a designated interpretation, say N, of S^1_2. Let $\alpha^\star$ be a conjunction of T-axioms that implies $(\mathsf{S}^1_2)^N$.

Define, for any translation k of the language of T to the language of T:

- $\mathsf{s}(k) := k\langle(\alpha^\star)^k\rangle\mathsf{id}$.

Thus $\mathsf{s}(k)$ is the interpretation that is is equal to k in case $\alpha^\star$ and that is the identity interpretation otherwise. We have, verifiably in S^1_2, for any k,

- $T \vdash (\alpha^\star)^{\mathsf{s}(k)}$,
- $T \vdash (\alpha^\star)^k \to (\phi^{\mathsf{s}(k)} \leftrightarrow \phi^k)$,

Thus, modulo T-provable equivalence, $\mathsf{s}(k)$ ranges precisely over all interpretations of $\alpha^\star$. We have:

Lemma 3.4. $\mathsf{S}^1_2 \vdash \forall k\, \Box_{T^{\mathsf{s}(k)}} (\mathsf{S}^1_2)^N$.

Proof. Reason in S^1_2. Consider any k. We have a proof in T of $(\alpha^\star)^{\mathsf{s}(k)}$. Hence, we have proofs of $\alpha^{\mathsf{s}(k)}$, for standardly finitely many T-axioms α together implying $(\mathsf{S}^1_2)^N$. Clearly, $T^{[\mathsf{s}(k)]}$ will imply each of these α, and, hence, $(\mathsf{S}^1_2)^N$. $\dashv$

We will take N as the designated interpretation of S^1_2 in the $T^{[k]}$.

Our assignments for the arithmetical interpretation are pairs σ, τ, where σ maps the propositional variables to T-sentences and τ maps the interpretation variables to translations from the language of T to the language of T. We stipulate that the σ are $\top$ for all but finitely many arguments and that the τ are id for all but finitely many arguments. The σ, τ are lifted to the arithmetical language in the obvious way, taking:

- $(\Box^{[k]} A)^{\sigma,\tau} := \Box_{T^{[\mathsf{s}(k)]}} A^{\sigma,\tau}$,

for the given interpretation, (ii) enable us to formalize the desired reasoning and (iii) are as parsimonious as possible in expressive power. There is definitely work to do to obtain formalizations of the systems fitting a good proof-theoretical format.

[5] We might wish to consider the following possible strengthening of P^{k}.

$$\mathsf{P}^{\mathsf{k}+} \quad \Gamma, \Delta, \{\Box(A_i \rhd^{[k]} B_i \mid i < n+1)\} \vdash C \Rightarrow \Gamma, \{A_i \rhd B_i \mid i < n+1\} \vdash C$$

We put the obvious conditions on occurrences of k and on Δ.

- $(A \rhd^{[k]} B)^{\sigma,\tau} := (T + A^{\sigma,\tau}) \rhd (T^{[\mathsf{s}(k)]} + B^{\sigma,\tau})$.

Note that the interpretation $\mathsf{s}(k)$ is applied only at locations where it is inside a provability. Thus, we can arithmetize its use in T. For this reason, the following definition makes sense.

- $\Gamma \models_T A :\Leftrightarrow \forall \sigma \;\; \mathsf{S}^1_2 \vdash \forall \tau \; (\bigwedge \Gamma^{\sigma,\tau} \to A^{\sigma,\tau})$.

The iterated modalities make sense because of Lemma 3.4. Note that, if we drop the superscripts in $\mathsf{J}^{\mathsf{k}}_5$, we get a formula that is equivalent over $\mathsf{J1}$, $\mathsf{J2}$ to the ordinary version of $\mathsf{J5}$. Moreover, in our system, we can derive: $\vdash \Diamond^{[k]} A \rhd^{[k]} A$.

We turn to checking the validity of our principles and rules. The principles $(\to \Box)^{\mathsf{k}}$ and $(\to \rhd)^{\mathsf{k}}$ are immediate from Lemma 3.2. The principles $\mathsf{L}^{\mathsf{k}}_1$ to $\mathsf{J}^{\mathsf{k}}_4$ are simple. The validity of $\mathsf{J}^{\mathsf{k}}_5$ follows from the observation that:

$$\mathsf{S}^1_2 \vdash \forall k, n \; (T^{[\mathsf{s}(k)]} + \mathsf{con}(T^{[\mathsf{s}(n)]} + B)) \rhd (T^{[\mathsf{s}(n)]} + B).$$

This follows by the usual formalization of Henkin's Theorem.

We now consider P^k. Reason in S^1_2. Suppose $k^\star : (T + \alpha) \rhd (T + \beta)$. Let $k' := k^\star\langle\alpha\rangle\mathsf{id}$ be the translation that acts like $k^\star$ if α and like id if $\neg\alpha$. Let $k := \mathsf{s}(k')$. (This last move is only of an administrative nature, since, in the present context, k' and k will be the same in their behaviour as interpretations.) Then, we have both $k : T \rhd T$ and $k : (T + \alpha) \rhd (T + \beta)$. By Lemma 3.3, we have $\Box_T(k : T \rhd T^{[k]})$. Also we have $k : (T+\alpha) \rhd \beta$, and, hence, $\Box_T(k : (T+\alpha) \rhd \beta)$. Combining, we find: $\Box_T(k : (T+\alpha) \rhd (T^{[k]}+\beta))$.

By Lemma 3.2, we find that $\mathsf{id} : T \equiv T^{[k]}$. So, from $\Box_T\delta$ we will get $\Box_{T^{[k]}}\delta$. Moreover, Suppose $m : (T + \delta) \rhd (T^{[k]} + \varepsilon)$. It follows that:

$$T + \varepsilon \xrightarrow{\mathsf{id}} T^{[k]} + \varepsilon \xrightarrow{m} T + \delta.$$

So, $m : (T + \delta) \rhd (T + \varepsilon)$.

Finally, $\mathsf{Nec}^{\mathsf{k}}$ is evident.

4 Arithmetical soundness results

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 **ILP** and **ILM**. We shall see that these two modal proofs give rise to two different arithmetical soundness proofs. The M-style proofs use definable cuts and find place is the modal system CuL. The P-style proofs are based on the use of approximating theories. This behavior is captured in te system AtL.

We now come to the soundness proofs of the following principles.

W	$\vdash A \rhd B \to A \rhd (B \wedge \Box\neg A)$
M_0	$\vdash A \rhd B \to (\Diamond A \wedge \Box C) \rhd (B \wedge \Box C)$
W^*	$\vdash A \rhd B \to (B \wedge \Box C) \rhd (B \wedge \Box C \wedge \Box\neg A)$
P_0	$\vdash A \rhd \Diamond B \to \Box(A \rhd B)$
R	$\vdash A \rhd B \to \neg(A \rhd \neg C) \rhd B \wedge \Box C$

The principles M_0 and P_0 both follow from R and W^* follows from M_0 and W. So it would be sufficient[6] to just prove the soundness of R and W. However, we have decided to give short proofs for all principles. Like this, the close match between the modal systems comes better to the fore. For every principle we shall give a proof in **ILP** and in **ILM**. These proofs can then be copied almost literally to yield arithmetical soundness proofs.

In CuL and AtL we will reason in an informal way, as if M^{J} and P^{k} were fomulated with an existential quantifier. It is easy to see how to convert this reasoning to the official format.

4.1 The principle W

We start with the **ILP**-proof of W.

Fact 4.1. $\mathbf{ILP} \vdash \mathsf{W}$.

Proof. We reason in **ILP**. Suppose $A \rhd B$. Then, $\Box(A \rhd B)$. Hence, $(*)$ $\Box(\Diamond A \to \Diamond B)$, and, thus, $(**)$ $\Box(\Box\neg B \to \Box\neg A)$.

Moreover, from $A \rhd B$, we have $A \rhd (B \wedge \Box\neg A) \vee (B \wedge \Diamond A)$. So it is sufficient to show: $B \wedge \Diamond A \rhd B \wedge \Box\neg A$. We have, by $(*)$,

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

⊣

Note that the proof of Fact 4.1, already works in $\mathbf{ILP}_{\mathsf{R}}$, where:

$\mathsf{P}_{\mathsf{R}} \vdash A \rhd B \to \Box(\Diamond A \to \Diamond B)$.

We turn to the **ILM**-proof of W.

Fact 4.2. $\mathbf{ILM} \vdash \mathsf{W}$.

Proof. We reason in **ILM**. Suppose $A \rhd B$ We find $A \wedge \Box\neg A \rhd B \wedge \Box\neg A$. But $A \rhd A \wedge \Box\neg A$, whence $A \rhd B \wedge \Box\neg A$. ⊣

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

Reason in AtL. Suppose that $A \rhd B$. By P^{k} we have that, for some k, $\Box(A \rhd^{[k]} B)$. Hence, by $\mathsf{J}^{\mathsf{k}}_4$, we have $(*)$ $\Box(\Diamond A \to \Diamond^{[k]} B)$ and, so, $(**)$ $\Box(\Box^{[k]}\neg B \to \Box\neg A)$.

Moreover, from $A \rhd B$, we have $A \rhd (B \wedge \Box\neg A) \vee (B \wedge \Diamond A)$. So it is sufficient to show $B \wedge \Diamond A \rhd B \wedge \Box\neg A$. We have, by $(*)$,

$$\begin{array}{lll} B \wedge \Diamond A \rhd \Diamond^{[k]} B & & \text{by } \mathsf{L}^{\mathsf{k}}_3 \\ \rhd \Diamond^{[k]}(B \wedge \Box^{[k]}\neg B) & & \text{by } \mathsf{J}^{\mathsf{k}}_5 \text{ and } (\mathsf{E}\rhd)^{\mathsf{k}} \\ \rhd B \wedge \Box^{[k]}\neg B & & \text{by } (**) \\ \rhd B \wedge \Box\neg A. & & \end{array}$$

[6] Originally in [JG04], a slightly different, but equivalent version of R was given. In the contribution of Joosten and Goris to this volume, a new principle is given that is precisely W and R together.

M-style soundness proof of W We reason in inner CuL. We assume $A \rhd B$. By M^{J}, we may find a J such that $A \wedge \Box^J \neg A \rhd B \wedge \Box \neg A$. By Lemma 2.4, $A \rhd A \wedge \Box^J \neg A$, whence $A \rhd B \wedge \Box \neg A$. We may conclude $A \rhd B \wedge \Box \neg A$.

4.2 The principle M_0

We start with the **IL**P-proof of M_0.

Fact 4.3. **IL**P $\vdash \mathsf{M}_0$.

Proof. Reason in **IL**P. Suppose $A \rhd B$. By P, we have:

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

$\dashv$

Note that the proof of Fact 4.3, already works in **IL**P$_{\mathsf{R}}$, We proceed with the **IL**M-proof of M_0.

Fact 4.4. **IL**M $\vdash \mathsf{M}_0$.

Proof. Reason in **IL**M. Suppose $A \rhd B$. We find $A \wedge \Box C \rhd B \wedge \Box C$. But, $\Diamond A \wedge \Box C \rhd \Diamond(A \wedge \Box C) \rhd A \wedge \Box C$, whence $\Diamond A \wedge \Box C \rhd B \wedge \Box C$. $\dashv$

P-style soundness proof of M_0 Reason in AtL. Suppose $A \rhd B$. By P^{k}, we have, for some k,

$$\begin{array}{lll} \Box(A \rhd^{[k]} B) & \to & \mathsf{J}_4^{\mathsf{k}} \\ \Box(\Diamond A \to \Diamond^{[k]} B) & \to & \\ \Box(\Diamond A \wedge \Box C \to \Diamond^{[k]} B \wedge \Box C) & \to & \\ \Diamond A \wedge \Box C \rhd \Diamond^{[k]} B \wedge \Box C & \to & \text{a.o. by } \mathsf{L}_2^{\mathsf{k}} \\ \Diamond A \wedge \Box C \rhd \Diamond^{[k]}(B \wedge \Box C) & \to & \text{by } \mathsf{J}_5^{\mathsf{k}} \text{ and } (\mathsf{E}\rhd)^{\mathsf{k}} \\ \Diamond A \wedge \Box C \rhd B \wedge \Box C & & \end{array}$$

M-style soundness proof of M_0 We reason in inner CuL. Suppose $A \rhd B$. We have, for some J, $A \wedge \Box^J C \rhd B \wedge \Box C$. By $\mathsf{L}_2^{\mathsf{J}}$, and necessitation, we also have $\Box(\Diamond A \wedge \Box C \to \Diamond A \wedge \Box\Box^J C)$, whence

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

4.3 The principle W^*

We start with the **IL**P-proof of W^*.

Fact 4.5. **IL**$\mathsf{P} \vdash \mathsf{W}^*$.

Proof. We reason in **IL**P: Suppose $A \rhd B$. Then:

$$\Box(\Box\neg B \to \Box\neg A) \tag{2}$$

and

$$\Box(\Diamond A \wedge \Box C \to \Diamond B \wedge \Box C) \tag{3}$$

Moreover, $B \wedge \Box C \rhd (B \wedge \Box C \wedge \Box\neg A) \vee (B \wedge \Box C \wedge \Diamond A)$. Thus, it is sufficient to show $B \wedge \Box C \wedge \Diamond A \rhd B \wedge \Box C \wedge \Box\neg A$. We have:

$$\begin{array}{lll} B \wedge \Box C \wedge \Diamond A & \rhd & \Diamond A \wedge \Box C \qquad\qquad \text{by (3)} \\ & \rhd & \Diamond B \wedge \Box C \qquad\qquad \text{by } \mathsf{L}_3 \\ & \rhd & \Diamond(B \wedge \Box\neg B) \wedge \Box C \qquad \text{by } \mathsf{L}_2 \\ & \rhd & \Diamond(B \wedge \Box C \wedge \Box\neg B) \qquad \text{by } \mathsf{J}_5 \\ & \rhd & B \wedge \Box C \wedge \Box\neg B \qquad\qquad \text{by (2)} \\ & \rhd & B \wedge \Box C \wedge \Box\neg A \end{array}$$

$\dashv$

Note that the proof of Fact 4.5, already works in **IL**P_R, We proceed with the **IL**M-proof of W^*.

Fact 4.6. **IL**$\mathsf{M} \vdash \mathsf{W}^*$.

Proof. We reason in **IL**M. Suppose $A \rhd B$. We have:

$$B \wedge \Box C \rhd (B \wedge \Box C \wedge \Box\neg A) \vee (B \wedge \Box C \wedge \Diamond A).$$

So it sufficient to show, $B \wedge \Box C \wedge \Diamond A \rhd B \wedge \Box C \wedge \Box\neg A$. We have:

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

$\dashv$

P-style soundness proof of W^* Reason in AtL. Suppose $A \rhd B$. By P^k we obtain a k, such that we have:

$$\begin{array}{lll} \Box(A \rhd^{[k]} B) & \to & \text{by } \mathsf{J}_4^\mathsf{k} \\ \Box(\Diamond A \to \Diamond^{[k]} B) & \to & \\ \Box(\Box^{[k]}\neg B \to \Box\neg A) & \to & (*) \\ \Box(\Diamond A \wedge \Box C \to \Diamond^{[k]} B \wedge \Box C) & & (**) \end{array}$$

By the usual reasoning, we only need to show:

$$B \wedge \Box C \wedge \Diamond A \rhd B \wedge \Box C \wedge \Box\neg A.$$

We have:

$$\begin{array}{lll} B \wedge \Box C \wedge \Diamond A & \rhd & \Diamond A \wedge \Box C \qquad \text{by } (**) \\ & \rhd & \Diamond^{[k]} B \wedge \Box C \qquad \text{by } \mathsf{L}_3^{\mathsf{k}} \\ & \rhd & \Diamond^{[k]}(B \wedge \Box^{[k]} \neg B) \wedge \Box C \qquad \text{by } \mathsf{L}_2^{\mathsf{k}} \\ & \rhd & \Diamond^{[k]}(B \wedge \Box C \wedge \Box^{[k]} \neg B) \qquad \text{by } \mathsf{J}_5^{\mathsf{k}} \text{ and } (\mathsf{E}\rhd)^{\mathsf{k}} \\ & \rhd & B \wedge \Box C \wedge \Box^{[k]} \neg B \qquad \text{by } (*) \\ & \rhd & B \wedge \Box C \wedge \Box \neg A \end{array}$$

M-style soundness proof of W* We reason in inner CuL. Suppose $A \rhd B$. By M^{J}, we have, for some J, $A \wedge \Box^J(C \wedge \neg A) \rhd B \wedge \Box(C \wedge \neg A)$, and, hence, by elementary reasoning (via the outer system),

$$A \wedge \Box^J C \wedge \Box^J \neg A \rhd B \wedge \Box C \wedge \Box \neg A.$$

It is sufficient to show: $B \wedge \Box C \wedge \Diamond A \rhd B \wedge \Box C \wedge \Box \neg A$. We have:

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

4.4 The principle P_0

We start with the **IL**P-proof of P_0.

Fact 4.7. **IL**P $\vdash \mathsf{P}_0$.

Proof. Reason in **IL**P. Suppose $A \rhd \Diamond B$. Then, $\Box(A \rhd \Diamond B)$ and, so, $\Box(A \rhd B)$. ⊣

Fact 4.8. **IL**$\mathsf{P}_\mathsf{R} \nvdash P_0$.

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

Fact 4.8 nicely reflects that the frame condition for P_0 essentially involves new S-transitions. We proceed with the **IL**M-proof of P_0.

Fact 4.9. **IL**M $\vdash \mathsf{P}_0$.

Proof. Reason in **IL**M.

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

⊣

P-style soundness proof of P_0 Reason in AtL. Suppose $A \rhd \Diamond B$. We have, for some k, $\Box(A \rhd^{[k]} \Diamond B)$.Hence, by $\mathsf{J}_5^{\mathsf{k}}$, $A \rhd \Diamond B \to \Box(A \rhd B)$.

M-style soundness proof of P_0 We reason in the inner system of CuL. Suppose $A \rhd \Diamond B$. By M^{J}, we have, for some J, $A \wedge \Box^J \neg B \rhd \bot$. It follows that $\Box(A \to \Diamond^J B)$ and, hence, $\Box\Box(A \to \Diamond^J B)$. We may conclude $\Box(A \rhd \Diamond^J B)$. So, $\Box(A \rhd B)$.

Note: the principle $A \rhd \Diamond B \to \Box(A \rhd \Diamond B)$ is also provable in both **IL**M and **IL**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.

4.5 The principle R

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

Lemma 4.10. **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 $(A \rhd B) \wedge \Box(B \to \Diamond\neg C) \to A \rhd \Diamond\neg C$ and $\Diamond\neg C \rhd \neg C$. ⊣

Fact 4.11. **IL**$\mathsf{P} \vdash \mathsf{R}$.

Proof. We reason in **IL**P. Suppose $A \rhd B$. It follows that $\Box(A \rhd B)$. Using this together with Lemma 4.10 we get:

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

⊣

Fact 4.12. **IL**$\mathsf{P}_\mathsf{R} \nvdash \mathsf{R}$.

Proof. By providing a countermodel as in the proof of Fact 4.8. ⊣

Fact 4.13. **IL**$\mathsf{M} \vdash \mathsf{R}$.

Proof. In **IL**, it is easy to see that $\neg(A \rhd \neg C) \to \Diamond(A \wedge \Box C)$. Reason in **IL**$\mathsf{M}$. Suppose $A \rhd B$. Then,

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

⊣

P-style soundness proof of R Reason in AtL. We first show that $(A \rhd^{[k]} B) \wedge \neg(A \rhd \neg C) \rightarrow \Diamond^{[k]}(B \wedge \Box C)$. Suppose that $A \rhd^{[k]} B$ and $\Box^{[k]}(B \rightarrow \Diamond\neg C)$, then, by $\mathsf{J}_2^{\mathsf{k}}\mathsf{b}$, $A \rhd^{[k]} \Diamond\neg C$. Thus, by $\mathsf{J}_5^{\mathsf{k}}$, we find $A \rhd \neg C$. By necessitation,

$$\Box((A \rhd^{[k]} B) \wedge \neg(A \rhd \neg C) \rightarrow \Diamond^{[k]}(B \wedge \Box C)). \tag{4}$$

We now turn to the main proof. Suppose $A \rhd B$. Then, for some k, we have $\Box(A \rhd^{[k]} B)$ and, thus,

$$\begin{array}{lcll} \neg(A \rhd \neg C) & \rhd & \neg(A \rhd \neg C) \wedge (A \rhd^{[k]} B) & \text{by (4)} \\ & \rhd & \Diamond^{[k]}(B \wedge \Box C) & \text{by } \mathsf{J}_5^{\mathsf{k}} \text{ and } (\mathsf{E}\rhd)^{\mathsf{k}} \\ & \rhd & B \wedge \Box C. & \end{array}$$

M-style soundness proof of R Reason in te inner system of CuL. Suppose that $A \rhd B$. Then, for some J, we have $A \wedge \Box^J C \rhd B \wedge \Box C$. By Lemma 2.5, we find that, without assumptions: $\neg(A \rhd \neg C) \rightarrow \Diamond(A \wedge \Box^J C)$. Hence, $\Box(\neg(A \rhd \neg C) \rightarrow \Diamond(A \wedge \Box^J C))$ and, so, $\neg(A \rhd \neg C) \rhd \Diamond(A \wedge \Box^J C)$ We have:

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

References

[AdJH98] C. Areces, D. de Jongh, and E. Hoogland. The interpolation theorem for IL and ILP. In *Proceedings of AiML98. Advances in Modal Logic*, Uppsala. Sweden, October 1998. Uppsala University.

[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.

[dJV90] D.H.J. de Jongh and F. Veltman. Provability logics for relative interpretability. In *[Pet90]*, pages 31–42, 1990.

[dJV91] D.H.J. de Jongh and A. Visser. Explicit fixed points in interpretability logic. *Studia Logica*, 50:39–50, 1991.

[dJV99] Dick de Jongh and Frank Veltman. The modal completeness of ILW. In Gerbrandy et al. [GMdRV99], page /contribs/jongh/. ISBN 90 5629 104 1.

[GMdRV99] Jelle Gerbrandy, Maarten Marx, Maarten de Rijke, and Yde Venema, editors. *JFAK. Essays Dedicated to Johan van Benthem on the Occasion of his 50th Birthday*, Vossiuspers, Amsterdam, 1999. Amsterdam University Press. ISBN 90 5629 104 1.

[JG04] J.J. Joosten and E. Goris. Modal matters in interpretability logic. Logic Group Preprint Series 226, Department of Philosophy, Utrecht University, Heidelberglaan 8, 3584 CS Utrecht, March 2004.

[JJM91] de D. Jongh, M. Jumelet, and F. Montagna. On the proof of Solovay's theorem. *Studia Logica*, 50:51–70, 1991.

[Pet90] P.P. Petkov, editor. *Mathematical logic, Proceedings of the Heyting 1988 summer school in Varna, Bulgaria*. Plenum Press, Boston, 1990.

[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.