Characterizations of Interpretability

Joost J. Joosten
and
Albert Visser

July 18, 2004

Abstract

This paper provides an overview of results related to the Orey-Hájek characterization and the Friedman characterization. We sketch the groundwork on which the results rest and we elaborate the mathematical structure behind the results.

A second goal of this paper is to give arithmetical soundness proofs of interpretability principles. The proofs should be of such generality that they make very little reference to the specific base theory.

We provide two modal systems within which such soundness proofs can be given. One system is based on definable cuts. The other is based on finite approximations of theories.

Contents

1 Introduction and preliminaries

Interpretations as we shall study them have proven to be a useful tool in comparing theories. If a theory U interprets a theory V, we write $U \rhd V$, then in a sense, U is as least as strong as V. Thus, interpretations gives us a means of comparing proof strength.

In words, an interpretation j of V in U, we write $j : U \rhd V$, is a structure preserving map sending axioms of V to theorems of U. Structure preserving means that the map should commute with proof constructions and with boolean connectives. Quantifiers get relativized to domain specifiers.

We demand the map to act on axioms of V rather than on theorems, so that indeed interpretations preserve proof structure. Our notion of interpretation is in rough lines the interpretations as studied by Tarski et al in [TMR53]. Let us put it into a formal definition. The theories that we study in this paper 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.

Definition 1.1. 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 free variables of R and $F(R)$ are the same. 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 $\perp^k = \perp$)
- $(\forall x\ \varphi(x))^k = \forall x\ (\delta(x) \to \varphi^k)$ and analogously for the existential quantifier

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

We can distinguish four different approaches to the study of interpretability.

1. Use interpretations in a series of case studies to relate the proof theoretic strength of various theories to each other. This study can be compared with other comparative methods like, for example, different proof theoretic ordinals.
2. 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. By identifying interpretations, we can also consider the category of theories where the interpretations are morphisms.
3. Study the general behavioral properties of interpretability and try to find logics describing this. This leads to the study of interpretability logics.

4. Study the nature of interpretations, for example by relating it to other meta-mathematical notions.

In this paper, the emphasis will be on the last three approaches. As the approaches are rather closely related, also in this paper they will come more or less intertwined. However, we have tried to set out some traces in our road map to this paper for the reader with a special focus.

1.1 A road map to the paper

For the reader primarily interested in provability logics we recommend the following track: Sections 1, 2 (without 2.4), 3, and 5.

The reader that is primarily interested in theorems on the nature of interpretability can follow the following track (in this order): Sections 1, 3 (without 3.4), 2.4, and 4.

For the reader with an interest in interpretability and categories there is the following track: Sections 1, 3 (without 3.4), and 4.2.

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

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 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, as in the way 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. 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.

1.3 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.

1.3.1 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 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 chose[1] to work with S^1_2.

Definition 1.2. 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.

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 $\exists\Delta^b_1$-axiomatization. That is, there is some $\exists\Delta^b_1$-formula $\mathsf{axioms}_U(x)$ with

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

Thus, the axiom set of U should be decidable in polynomial time. The choice of $\exists\Delta^b_1$-axiomatizations is also motivated by Lemma 1.3.

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 1.3. *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.

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

1.3.2 Reflexive theories

Many meta-mathematical statements involve the notion of reflexivity. The idea of a theory being reflexive is essentially that it proves the consistency of all of its finite subtheories. Throughout literature we can find many different variants of this notion. For stronger theories, all these notions coincide. But for weaker theories, the differences are essential. We give some notions of reflexivity.

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 number $\leq n$.

[1]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.

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 number $\leq n$. All formulas φ in 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 1.8) preserved under translations. An example of such a ρ is given in [Vis93].

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

We shall write $\Box_{U,n}$ for $\neg\mathsf{con}_n(U + \neg\varphi)$ or $\exists p\ \mathsf{proof}_{U,n}(p, \varphi)$. 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 1.4. An inspection of the proof of provable Σ_1-completeness (Lemma 1.3) gives us some more information. The proof p that witnesses the provability in U of some $\exists\Sigma^b_1$-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 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 4.2.

If U is a reflexive theory, we do not necessarily have any reflection principles. That is, we do not have $U \vdash \Box_V \varphi \rightarrow \varphi$ 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 1.5. *Let U be a reflexive theory. Then*

$$T \vdash \forall^{\forall\Pi^b_1}\pi\, \forall n\ \Box_U \forall x\ (\Box_{U,n}\pi(x) \rightarrow \pi(x)).$$

Proof. Reason in T and fix π and n. Let m be such that we have (see Lemma 1.3 and Remark 1.4)

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

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 $\forall x\ (\Box_{U,n}\pi(x) \rightarrow \pi(x))$. $\dashv$

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

1.4 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 1.6. 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[2] $\bigwedge_{x_i \in \mathsf{FV}(\varphi)} \delta(x_i) \to \varphi^k$. 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 1.7 below. For example, the translation of

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

will be

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

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[3] in S^1_2, as the number of free variables in $\varphi \wedge \psi$ is bounded by $|\varphi \wedge \psi|$.

Lemma 1.7. *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

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

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

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

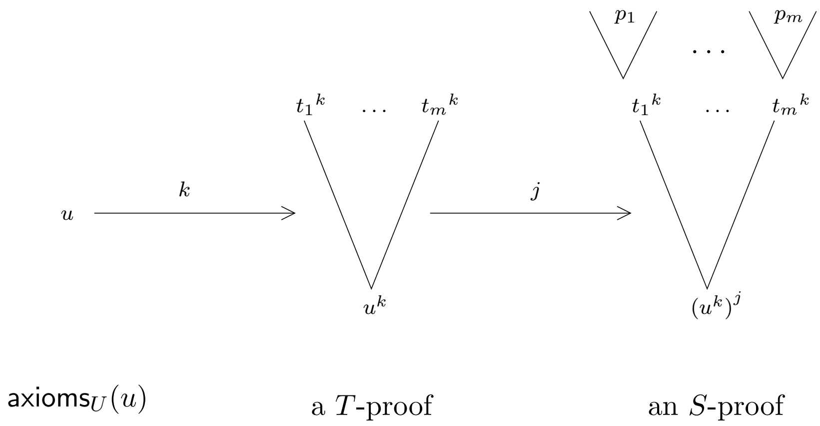

Figure 1: Transitivity of interpretability

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 1.6, which is tailored precisely to let the induction go through. $\dashv$

Remark 1.8. 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 interpretations. Thus, if we have $j : S \rhd T$ and $k : T \rhd U$ we would like to have[4] $j \circ k : S \rhd U$.

If we try to perform this proof as depicted in Figure 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 interpretability as in Definition 1.9. In the presence of $\mathrm{B}\Sigma_1$

[4] A formal definition of $j \circ k$ is given in Section 2.1.

In S^1_2:

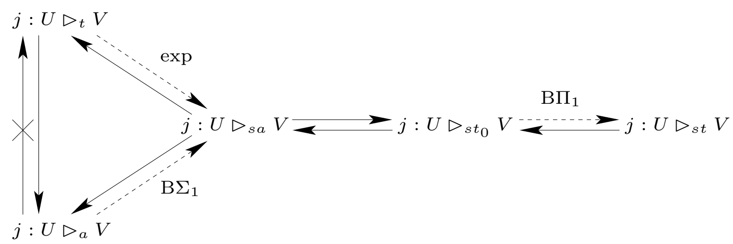

Figure 2: Versions of relative interpretability

this notion will coincide with the earlier defined notion of interpretability, as Theorem 1.10 tells us.

Definition 1.9. We define the notions of axioms interpretability $\rhd_a$, theorems interpretability $\rhd_t$ and smooth axioms interpretability $\rhd_{sa}$. For completeness and academic perversion we also add the notions of weak smooth theorems interpretability $\rhd_{st_0}$ and smooth theorems interpretability $\rhd_{st}$.

$$\begin{array}{lcl}
j : U \rhd_a V & := & \forall v\, \exists p\ (\mathsf{axioms}_V(v) \to \mathsf{proof}_U(p, v^j)) \\
j : U \rhd_t V & := & \forall \varphi\, \forall p\, \exists p'\ (\mathsf{proof}_V(p, \varphi) \to \mathsf{proof}_U(p', \varphi^j)) \\
j : U \rhd_{sa} V & := & \forall x\, \exists y\, \forall\, v {\leq} x\, \exists\, p {\leq} y\ (\mathsf{axioms}_V(v) \to \mathsf{proof}_U(p, v^j)) \\
j : U \rhd_{st_0} V & := & \forall x\, \exists y\, \forall\, \varphi {\leq} x\, \forall\, p {\leq} x\, \exists\, p' {\leq} y\ (\mathsf{proof}_V(p, \varphi) \to \mathsf{proof}_U(p', \varphi^j)) \\
j : U \rhd_{st} V & := & \forall x\, \exists y\, \forall\, \varphi {\leq} x\, \exists\, p' {\leq} y\, \forall p\ (\mathsf{proof}_V(p, \varphi) \to \mathsf{proof}_U(p', \varphi^j))
\end{array}$$

Theorem 1.10. *In S^1_2 we have all the arrows as depicted in Figure 2: Versions of relative interpretability. The dotted arrows indicate that an additional condition is needed; the condition written next to it.*

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 $\forall v\, \exists p\ (\mathsf{axioms}_V(v) \to \mathsf{proof}_U(p, v^j))$. If we fix some x, we get $\forall\, v {\leq} x\, \exists p\ (\mathsf{axioms}_V(v) \to \mathsf{proof}_U(p, v^j))$. By $\mathrm{B}\Sigma_1$ we get the required $\exists y\, \forall\, v {\leq} x\, \exists\, p {\leq} y\ (\mathsf{axioms}_V(v) \to \mathsf{proof}_U(v^j))$. It is not clear if $T \vdash \mathrm{B}\Sigma_1$ 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 $T \vdash \exp$. If T is reflexive, we get by Corollary 4.9 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 $T \vdash \exp$, 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

$$\forall x\, \exists y\, \forall\, v {\leq} x\, \exists\, p {\leq} y\ (\mathsf{axioms}_V(v) \to \mathsf{proof}_U(v^j)). \tag{1}$$

So, we pick x arbitrarily and consider[5] $\nu := \bigwedge_{\mathsf{axioms}_V(v_i) \wedge v_i \leq x} v_i$. Notice that in the worst case, for all $y \leq x$, we have $\mathsf{axioms}_V(y)$, whence the length of ν can be bounded by $x \cdot |x|$. Thus, ν itself can be bounded by x^x, which exists whenever $T \vdash \exp$. 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). It is not clear if $T \vdash \exp$ is a necessary condition.

- $\mathsf{S}^1_2 \vdash j : U \rhd_{sa} V \to j : U \rhd_{st_0} V$. 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.

- $T \vdash j : U \rhd_{st_0} V \to j : U \rhd_{st} V$ if $T \vdash \mathrm{B}\Pi_1$. Suppose

$$\forall x \, \exists y \, \forall \varphi{\leq}x \, \forall p{\leq}x \, \exists p'{\leq}y \ (\mathsf{proof}_V(p, \varphi) \to \mathsf{proof}_U(p', \varphi^j)). \qquad (2)$$

We wish to see

$$\forall x \, \exists y \, \forall \varphi{\leq}x \, \exists p'{\leq}y \, \forall p \ (\mathsf{proof}_V(p, \varphi) \to \mathsf{proof}_U(p', \varphi^j)). \qquad (3)$$

Consider x. Now,

$$\forall \varphi{\leq}x \, \exists p (\Box_V \varphi \to \mathsf{proof}_V(p, \varphi)).$$

As $\Box_V \varphi \to \mathsf{proof}_V(p, \varphi)$ is Π_1, we get by $\mathrm{B}\Pi_1$ that

$$\exists y_0 \, \forall \varphi{\leq}x \, \exists p{\leq}y_0 \ (\Box_V \varphi \to \mathsf{proof}_V(p, \varphi)).$$

Now we can take this $y_0 = x$ and apply 2 to obtain some y such that

$$\forall \varphi{\leq}y_0 \, \forall p{\leq}y_0 \, \exists p'{\leq}y \ (\mathsf{proof}_V(p, \varphi) \to \mathsf{proof}_U(p', \varphi^j)).$$

Clearly, this y is also sufficient for 3. It is not clear if $T \vdash \mathrm{B}\Pi_1$ is a necessary condition. $\dashv$

We note that the notion of smooth theorems interpretability is not really a good one. For it seems to say more on the nature of a theory than on the nature of interpretability. For example, at first sight we seem to need $U \vdash \mathrm{B}\Pi_1$ in order to obtain the desirable $U \vdash \mathsf{id} : U \rhd_{st} U$ as the most straightforward proof of it goes as follows.

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

We pick some x and consider all the $\varphi \leq x$ for which $\exists p\ (\mathsf{proof}_U(p, \varphi) \wedge \forall\, p' {\leq} p\ \neg \mathsf{proof}_U(p', \varphi))$. By $\mathrm{B}\Pi_1$ we can pick the largest such p to serve as a bound. But at $\mathrm{B}\Pi_1$, all the notions of interpretability coincide ($\mathrm{B}\Pi_1 \vdash \mathrm{B}\Sigma_1$).

Finally, 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 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 the paper, 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.

1.5 Interpretations and models

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

Definition 1.11. Let $j : U \rhd V$ with $j = \langle \delta, F \rangle$. If $M \models U$, we denote by M^j the following model.

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

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

2 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. 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 propositions 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. As we shall see, in doing so 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 paper 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 2.1. 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.

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 2.2. 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 **IL**(T), is the set of all the interpretability principles of T or a logic that generates it.

2.1 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 **IL** $\subset$ **IL**(T) for any reasonable T.

Definition 2.3. The logic **IL** is the smallest set of formulas being closed under the rules of Necessitation and of Modus Ponens, that contains all tautological formulas and all instantiations of the following axiom schemata.

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

L2 $\Box A \to \Box\Box A$

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

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

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

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

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

J5 $\Diamond A \rhd A$

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

Lemma 2.4.

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 $\mathsf{L3}$ 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 $\mathsf{J5}$ 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 see 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 $\Box_T(\alpha \to \beta^j)$. If now $\Box_T\neg\beta$ then $\Box_T(\alpha \to \neg\beta^j)$ and we get $\Box_T\neg\alpha$.

The only principle of **IL** that needs some serious argument is J_5. We shall discuss this in the next subsection.

2.2 The formalized henkin construction

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. In a more general setting, we have the following theorem.

Theorem 2.5. *If $U \vdash \mathsf{con}(V)$, then $U \rhd V$.*

A proof 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 $\exists x\varphi(x) \rightarrow \varphi(c_{\exists x\varphi(x)})$. 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 3. We get the following theorem.

Theorem 2.6. *For any numberizable theories T and U, we have that for any V,*

$$T \vdash \Box_U \mathsf{con}(V) \rightarrow \exists k\ (k : U \rhd V\ \&\ \forall\varphi\ \Box_U(\Box_V\varphi \rightarrow \varphi^k)).$$

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

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

$\dashv$

2.3 More logics

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

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

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** ($\mathsf{L1}$-$\mathsf{J5}$) and of the axiom schemata of X.

A central theme in the study of formalized interpretability is to determine **IL**(T) for a specific T. For two classes of theories, **IL**(T) is known.

Definition 2.7. A theory T is essentially reflexive if all of its finite sentential extensions are reflexive.

Theorem 2.8 (Berarducci [Ber90], Shavrukov [Sha88]). *If T is an essentially reflexive theory, then* $\mathbf{IL}(\mathrm{T}) = \mathbf{IL}\mathsf{M}$.

Theorem 2.9 (Visser [Vis90a]). *If T is finitely axiomatizable, then,* $\mathbf{IL}(\mathrm{T}) = \mathbf{IL}\mathsf{P}$.

2.4 Essentially reflexive theories

In Definition 2.7 we stressed that we only considered sentential extensions of T. We can also consider extensions with formulas. This gives rise to the notion of *essentially globally reflexive theories*. We can restate the definition as follows.

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

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

Lemma 2.10. *If T is an essentially globally reflexive theory extending*[6] EA*, then T satisfies full induction.*

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

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

Then, for some m,

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

Thus, also

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

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 $T \vdash \forall x \; \varphi(x)$. ⊣

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 2.10 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

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

conjunction of the first n axioms. By definition $\psi \wedge U[n] \rightarrow \mathsf{con}(\psi \wedge U[n])$ is an axiom of U, whence $U \vdash \psi \rightarrow \mathsf{con}(\psi \wedge U[n])$.

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

Thus, for example $\mathrm{EA} + \{\varphi \rightarrow \mathsf{con}(\varphi) \mid \varphi \text{ a sentence }\} \subseteq \mathrm{EA} + \mathsf{True}_{\Pi_1}$. It is well known that no Σ_3-axiomatized theory can prove $\mathrm{I}\Sigma_1$ (see for example Fact 2.3 from [Joo03]). But $\mathrm{EA}+\mathsf{True}_{\Pi_1}$ has a Π_2-axiomatization, thus $\mathrm{EA} + \{\varphi \rightarrow \mathsf{con}(\varphi) \mid \varphi \text{ a sentence }\} \nvdash \mathrm{I}\Sigma_1$.

Admittedly, theories like the U above are a bit artificial. All natural theories that are essentially reflexive are globally so and hence by Lemma 2.10 satisfy full induction.

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 posses 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 $\varphi(0) \wedge \forall x\ (\varphi(x) \rightarrow \varphi(x+1))$ 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.

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

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

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 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. (sketch) Let us start by making the quantifier $\exists^{U\text{-}\mathsf{Cut}} J$ a bit more precise. By $\exists^{U\text{-}\mathsf{Cut}} J$ we shall mean $\exists J \ \Box_U \mathsf{Cut}(J)$. 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 that $J(0)$. 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 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 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 $\Delta_0(\omega_1)$. $\dashv$

As all U-cuts are closed under $\omega_1(x)$, 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.

3.2 Cuts and the henkin construction

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.

Theorem 3.5. *We have Theorem 2.6 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 2.6. 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. $\dashv$

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

Lemma 3.6. *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. $\dashv$

Remark 3.7. 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 $\delta'(x) := \delta(x) \wedge \forall\, y{<}x\ (\delta(y) \rightarrow y{\neq^j}x)$. F' agrees with F on all symbols except that it maps identity to identity. By the minimal number principle we can prove $\forall x\ (\delta(x) \rightarrow \exists x'\ (x'{=^j}x) \wedge \delta'(x))$, and thus $\forall \vec{x}\ (\delta'(\vec{x}) \rightarrow (\varphi^j(\vec{x}) \leftrightarrow \varphi^{j'}(\vec{x})))$ for all formulae φ. $\dashv$

It is not the case that the implication in Remark 3.7 can be reversed. For, if U is reflexive, contains $\mathrm{I}\Delta_2$ and $U \rhd V$, the following reasoning can be performed. By reflexivity of U (and the totality of exp), we get by Lemma 4.2 of the Orey-Hájek characterization that $\forall x\ \Box_U \mathsf{con}(V, x)$. We can now perform the henkin construction (Lemma 4.1). This yields an interpretation where all symbols of V get a Δ_2-translation. Thus, by $\mathrm{I}\Delta_2$ we can prove $\forall x\ (\delta(x) \rightarrow \exists x'\ (x'{=^j}x) \wedge \delta'(x))$ 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 Σ_3-IR.

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 M of U and its internally defined model M^j of V if $j : U \rhd V$. If U has full induction, this comparison is fairly easy.

Theorem 3.8. *Suppose $j : U \rhd V$ and U has full induction. Let M be a model of U. We have that $M \subseteq_e M^j$ via a definable embedding.*

Proof. (sketch) If U has full induction and $j : U \rhd V$, we may by Remark 3.7 actually assume that j maps identity in V to identity in U. Thus, we can define the following function.

$$f := \begin{cases} 0 \mapsto 0^j \\ x+1 \mapsto f(x) {+^j} 1^j \end{cases}$$

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 codomain and the codomain is an initial segment of M^j. $\dashv$

If U does not have full induction, a comparison between M and 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 M and a definable embedding $f : I \longrightarrow M^j$ such that $f[I] \subseteq_e M^j$.

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 M and M^j respectively, make true the same Δ_0-formulas.

Lemma 3.9 (Pudlák's Lemma). *In the proof of the lemma we shall make the quantifier $\exists^{j,J\text{-function}} h$ 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. The lemma can now be stated as follows.*

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

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

A detailed proof of Pudlák's lemma is given in Appendix B. If j is an interpretation with $j : \alpha \rhd \beta$, we shall sometimes call the corresponding isomorphic cut that is given by Lemma 3.9, the *Pudlák* cut of j and denote it by the corresponding upper case letter J.

3.4 Cuts and interpretability logics

In a modal setting, facts about cuts are reflected in the following way.

$(\rightarrow)^{\mathsf{J}}$	$\Box^J A \rightarrow \Box A$	$(\forall)$
$\mathsf{L}_1^{\mathsf{J}}$	$\Box^J(A \rightarrow B) \rightarrow (\Box^J A \rightarrow \Box^J B)$	$(\forall)$
$\mathsf{L}_2^{\mathsf{J}}$	$\Box^I A \rightarrow \Box^I \Box^J A$	$(\forall\forall)$
$\mathsf{L}_3^{\mathsf{J}}$	$\Box^J(\Box^I A \rightarrow A) \rightarrow \Box^J A$	$(\forall\forall)$
$\mathsf{J}_5^{\mathsf{J}}$	$\Diamond^I A \rhd A$	$(\forall)$
M^{J}	$A \rhd B \rightarrow A \wedge \Box^J C \rhd B \wedge \Box C$	$(\exists)$
$\mathrm{Nec}^{\mathsf{J}}$	$\vdash A / \vdash \Box^J A$	$(\forall)$

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.

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

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; $\mathsf{L}_1^{\mathsf{J}}$ reflects that concatenation of proofs in a cut, remains within this cut as concatenation is approximately multiplication; $\mathsf{L}_2^{\mathsf{J}}$ is Lemma 3.10; $\mathsf{L}_3^{\mathsf{J}}$ is Löb's theorem with cuts, as proved in Lemma 3.11; $\mathsf{J}_5^{\mathsf{J}}$ follows from Theorem 3.5; M^{J} is Lemma 3.12. Also $\mathsf{Nec}^{\mathsf{J}}$ is easy to see. If α is provable, it has a standard proof, and standard proofs are provably in any cut.

Lemma 3.10. *For any U-cuts I and J we have that $T \vdash \Box_U^I \alpha \to \Box_U^I \Box_U^J \alpha$.*

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 1.3 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 $\Box_U^I \mathsf{proof}_U(\alpha)$. Lemma 3.2 now gives us the desired $\Box_U^I \Box_U^J \alpha$. ⊣

Lemma 3.11. *For any U-cuts I and J we have that $T \vdash \Box_U^J(\Box_U^I \alpha \to \alpha) \to \Box_U^J \alpha$.*

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 $\mathsf{Nec^J}$ 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.

$$
\begin{array}{lllr}
\Box^J \Box^I (F \to (\Box^I F \to \alpha)) & \to & \Box^J(\Box^I F \to (\Box^I \Box^I F \to \Box^I \alpha)) & \text{by } \mathsf{L_2^J} \\
 & \to & \Box^J(\Box^I F \to \Box^I \alpha) & \text{by } \mathbb{A} \\
 & \to & \Box^J(\Box^I F \to \alpha) & (*) \\
 & \to & \Box^J F & \text{by } \mathsf{L_2^J} \\
 & \to & \Box^J \Box^I F & \text{by } (*) \\
 & \to & \Box^J \alpha &
\end{array}
$$

⊣

Lemma 3.12. *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 3.9. 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

$$\Box_{T+\alpha}(\Box^J \gamma \to (\Box \gamma)^j). \tag{4}$$

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) we now get the required $\Box_{T+\alpha+\Box^J\gamma} \varphi^j$. ⊣

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 3.13. *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. ⊣

Lemma 3.14. *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) $\Box(A \to \Diamond^J \neg C) \to A \rhd \Diamond^J \neg C \rhd \neg C$. ⊣

4 Characterizations of interpretability

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

In Subsection 4.1 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 Subsection 4.2 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.

Finally, in Subsection 4.3 we give a model theoretic characterization of interpretability.

4.1 The Orey-Hájek characterizations

We consider the diagram from Figure 3. We shall comment on all the arrows in the diagram.

Lemma 4.1. *In S^1_2 we can prove $\forall n\ U \vdash \mathsf{con}_n(V) \Rightarrow U \rhd V$.*

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 3.6 this is automatically satisfied.

Let us first give the informal proof. Thus, let $\mathsf{axioms}_V(x)$ 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, $\mathsf{axioms}_{V'}(x) := \mathsf{axioms}_V(x) \wedge \mathsf{con}_x(V)$.

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

$$U \vdash \mathsf{con}(V'). \tag{5}$$

Thus, by Theorem 2.5 we get that $U \rhd V'$. Second, we shall see that

$$V = V'. \tag{6}$$

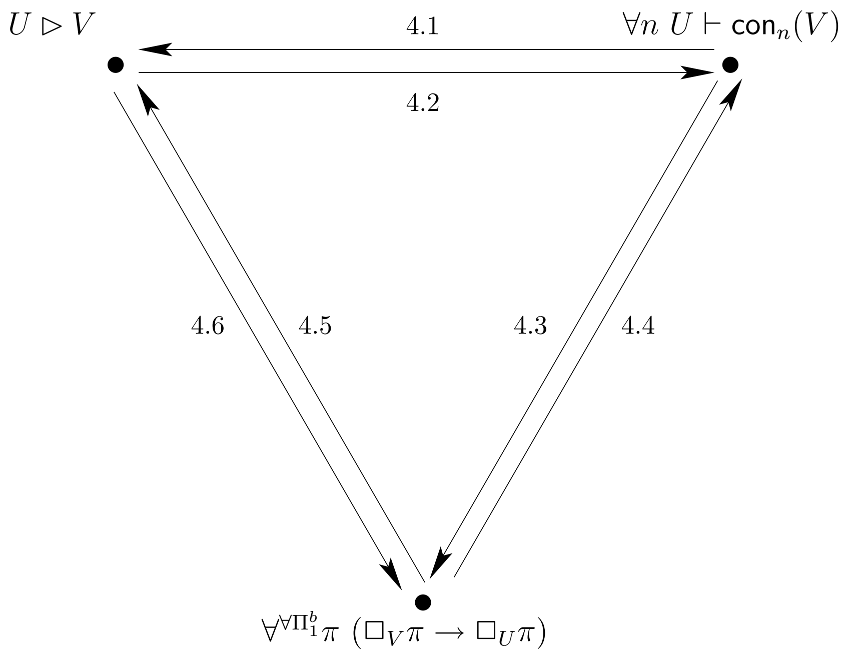

Figure 3: Characterizations of interpretability

To see (5), 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 $\mathsf{axioms}_{V'}(v)$ 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 can not 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 (6). Clearly $\mathbb{N} \models \mathsf{axioms}_{V'}(v) \rightarrow \mathsf{axioms}_V(v)$ for any $v \in \mathbb{N}$. To see that the converse also holds, we reason as follows.

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

$$\mathbb{N} \subseteq_e M. \tag{7}$$

As $M \models \mathsf{con}_v(V)$ 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 $\mathbb{N} \models \mathsf{axioms}_V(v)$, thus we get that $\mathbb{N} \models \mathsf{axioms}_{V'}(v)$. We conclude that

$$\mathbb{N} \models \mathsf{axioms}_V(x) \leftrightarrow \mathsf{axioms}_{V'}(x) \tag{8}$$

whence, that $V = V'$. As $U \vdash \mathsf{con}(V')$, we get by Theorem 2.5 that $U \rhd V'$. We may thus infer the required $U \rhd V$.

It is not possible to directly formalize the informal proof. At (8) we concluded that $V = V'$. This actually uses some form of Π_1-reflection which is manifested in (7). The lack of reflection in the formal environment will be compensated by another sort of reflection, as formulated in Theorem 2.6.

Moreover, to see (5), 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).

$$V = V_0 \subseteq V_1 \subseteq V_2 \subseteq \cdots \text{with } \mathsf{con}^*(V_i) \tag{9}$$

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 (9) 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.

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

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

$\dashv$

In U:

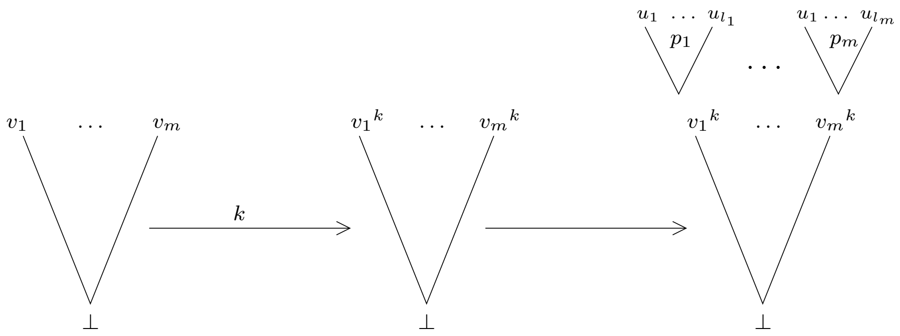

Figure 4: Transformations on proofs

Lemma 4.2. *In the presence of* exp*, we can prove that for reflexive* U*,* $U \rhd V \to \forall x\ \Box_U \mathsf{con}_x(V)$*.*

Proof. The informal argument is conceptually very clear and we have depicted it in Figure 4. 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 $\forall u\, \exists p\ (\mathsf{axioms}_V(u) \to \mathsf{proof}_U(p, u^k))$. 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

$$\forall\, u{\leq}x\, \exists p\ (\mathsf{axioms}_V(u) \to \mathsf{proof}_{U,l}(p, u^k)). \tag{10}$$

This formula is actually equivalent to the Σ_1-formula

$$\exists n\, \forall\, u{\leq}x\, \exists\, p{\leq}n\ (\mathsf{axioms}_V(u) \to \mathsf{proof}_{U,l}(p, u^k)) \tag{11}$$

from which we may conclude by provable Σ_1-completeness,

$$U \vdash \exists n\, \forall\, u{\leq}x\, \exists\, p{\leq}n\ (\mathsf{axioms}_V(u) \to \mathsf{proof}_{U,l}(p, u^k)). \tag{12}$$

We now reason in U and suppose that there is some V, x-proof p of $\perp$. The assumptions in p are axioms $v_1 \ldots v_m$ of V, with each $v_i \leq x$. Moreover, all the formulas ψ in p have $\rho(\psi) \leq x$. By Lemma 1.7, p transforms to a proof p^k of $\perp^k$ which is again $\perp$.

The assumptions in p^k are now among the ${v_1}^k \ldots {v_m}^k$. By Remark 1.8 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 (12), 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 $\perp$ by substituting all the p_i for the $(v_i)^k$. Thus we have shown $\mathsf{proof}_{U,l+n'}(p', \perp)$. But this clearly contradicts the reflexivity of U.

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

First of all, to conclude that (10) is equivalent to (11), 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 10.

In addition we need that T proves the totality of exponentiation. For weaker theories, we only have provable $\exists\Sigma_1^b$-completeness. But if $\mathsf{axioms}_V(u)$ is Δ_1^b, we can only guarantee that $\forall\, u{\leq}m\, \exists\, p{\leq}n\ (\mathsf{axioms}_V(u) \to \mathsf{proof}_U(p, u^k))$ is Π_2^b. As far as we know, exponentiation is needed to prove $\exists\Pi_2^b$-completeness.

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

⊣

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

$$\forall V\ (U \rhd V \to \forall x\ \Box_U \mathsf{con}_x(V)), \tag{13}$$

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

$$\forall U\ (U \rhd V \to \forall x\ \Box_U \mathsf{con}_x(V)), \tag{14}$$

that then V is reflexive. However, V being reflexive could never be a sufficient condition for (14) to hold, as we know from [Sha97] that interpreting reflexive theories in finitely many axioms is complete Σ_3.

Lemma 4.3. *In S_2^1 we can prove $\forall x\ \Box_U \mathsf{con}_x(V) \to \forall^{\forall\Pi_1^b}\pi\ (\Box_V \pi \to \Box_U \pi)$.*

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_2^1 and assume $\forall x\ \Box_U \mathsf{con}_x(V)$. Now we consider any $\pi \in \forall\Pi_1^b$ 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 1.4)

$$\Box_U(\neg\pi \to \Box_{V,x}\neg\pi). \tag{15}$$

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

$$\Box_U \Box_{V,x}\pi. \tag{16}$$

Combining (15), (16) and the assumption that $\forall x\ \Box_U \mathsf{con}_x(V)$, we see that indeed $\Box_U \pi$. ⊣

Lemma 4.4. *In S_2^1 we can prove that for reflexive V we have*

$$\forall^{\forall\Pi_1^b}\pi\ (\Box_V \pi \to \Box_U \pi) \to \forall x\ \Box_U \mathsf{con}_x(V).$$

Proof. If V is reflexive and $\forall^{\forall\Pi_1^b}\pi\ (\Box_V \pi \to \Box_U \pi)$ then, as for every x, $\mathsf{con}_{\overline{x}}(V)$ is a $\forall\Pi_1^b$-formula, also $\forall x\ \Box_U \mathsf{con}_x(V)$. ⊣

It is obvious that

$$\forall U\ [\forall^{\forall\Pi_1^b}\pi\ (\Box_V\pi \to \Box_U\pi) \to \forall x\ \Box_U \mathsf{con}_x(V)] \tag{17}$$

implies that V is reflexive. Likewise,

$$\forall V\ [\forall^{\forall\Pi_1^b}\pi\ (\Box_V\pi \to \Box_U\pi) \to \forall x\ \Box_U \mathsf{con}_x(V)] \tag{18}$$

implies that U is reflexive. However, U being reflexive can never be a sufficient condition for (18) to hold. An easy counterexample is obtained by taking U to be PRA and V to be $\mathrm{I}\Sigma_1$. (See for example [Joo03].)

Lemma 4.5. *For reflexive V we have $\forall^{\forall\Pi_1^b}\pi\ (\Box_V\pi \to \Box_U\pi) \to U \rhd V$.*

Proof. We know of no direct proof of this implication. Also, all proofs in the literature go via Lemmata 4.4 and 4.1, and hence use reflexivity of V. ⊣

Again, by [Sha97] and Lemma 4.6 we see that U being reflexive can not be a sufficient condition for $\forall^{\forall\Pi_1^b}\pi\ (\Box_V\pi \to \Box_U\pi) \to U \rhd V$ 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 4.6. *Let U be a reflexive and sequential theory. We have that $U \rhd V \to \forall^{\forall\Pi_1^b}\pi\ (\Box_V\pi \to \Box_U\pi)$.*

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

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 3.7 assume that j maps identity to identity. By Theorem 3.8 we now see that $M \subseteq_e M^j$. If for some $\pi \in \Pi_1$, $\Box_V\pi$ then by soundness $M^j \models \pi$, whence $M \models \pi$. As 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 mayor 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 3.9, we now find[7] a definable U-cut J and a j, J-function h such that

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

[7]Remark B.3 ensures us that we can find them also in the case of theorems interpretability.

We shall see that for this cut J we have that

$$\forall^{\Pi_1}\pi\ (\Box_V\pi \to \Box_U\pi^J). \tag{19}$$

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

$$\Box_U\forall x\ (\delta(x) \to \varphi^j(x)). \tag{20}$$

We are to see $\Box_U\forall x\ (J(x) \to \varphi(x))$. 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 $\delta(h(x))$, whence by (20), $\varphi^j(h(x))$. Pudlák's lemma now yields the desired $\varphi(x)$. As x was arbitrary, we have proved (20).

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 3.2 and Remark 3.3, for large enough n, this implies

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

and by Lemma 1.5 (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 4.2) theories, reflexivity is also a necessary condition. That is to say, that for such theories,

$$\forall V\ [U \rhd V \to \forall^{\forall\Pi_1^b}\pi\ (\Box_V\pi \to \Box_U\pi)], \tag{21}$$

implies that U is reflexive.[8] For, if U is sequential, we get by Lemma 4.12 that for every n, $U \rhd \mathsf{S}^1_2 + \mathsf{con}_n(U)$. Thus, by (21) we get that $\forall n\ U \vdash \mathsf{con}_n(U)$.

The sequentiallity is essentially used here: we can expose a non-sequential non-reflexive U which satisfies (21).

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

First, U satisfies (21). Suppose that $\mathsf{PA} \boxplus T \rhd V$ and $\Box_V\pi$. Then $V \rhd \mathsf{S}^1_2 + \pi$ whence $\mathsf{PA} \boxplus T \rhd \mathsf{S}^1_2 + \pi$. As we have pairing in $\mathsf{S}^1_2 + \pi$, we can

[8] 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 $\mathsf{S}^1_2 + \forall^{\Pi_1}\pi\ \varphi(\pi)$.

apply Theorem A.5 and obtain that $\mathsf{PA} \rhd \mathsf{S}^1_2 + \pi$ 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 $\mathsf{PA} \rhd \mathsf{S}^1_2 + \pi$. By Lemma 4.6 we conclude that $\mathsf{PA} \vdash \pi$, whence $\mathsf{PA} \boxplus T \vdash \pi$.

Second, we see that $\mathsf{PA} \boxplus T$ can not be reflexive. Suppose for a contradiction that $\forall n\ \mathsf{PA} \boxplus T \vdash \mathsf{con}_n((\mathsf{PA} \boxplus T))$. Then, for all n, $\mathsf{PA} \vdash \mathsf{con}_n((\mathsf{PA} \boxplus T))$ and thus, for sufficiently large n, $\mathsf{PA} \vdash \mathsf{con}(T)$. But this would imply that $\mathsf{PA} \vdash \mathsf{con}(\mathsf{PA})$ which is a contradiction.

Again, by [Sha97] we note that V being reflexive can never be a sufficient condition for $\forall U\ [U \rhd V \rightarrow \forall^{\forall\Pi^b_1}\pi\ (\Box_V \pi \rightarrow \Box_U \pi)]$.

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

Corollary 4.7. *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 4.8. *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) $\Rightarrow$ (2) is already proved in Lemma 4.6, (2) $\Rightarrow$ (3) follows from the transitivity of V and (3) $\Rightarrow$ (1) is a sharpening of Lemma 4.1. which closely follows Theorem 3.5. Note that $\rhd$ may denote denote smooth or theorems interpretability. ⊣

Corollary 4.9. *If V is reflexive, then*

$$\vdash U \rhd_t V \leftrightarrow U \rhd_s V.$$

Proof. By Remark B.3 and Corollary 4.8. ⊣

Corollary 4.10. *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. $\forall^{\forall\Pi^b_1}\pi\ (\Box_V \pi \rightarrow \Box_U \pi)$
3. $\forall x\ \Box_U \mathsf{con}_x(V)$

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

As an application we can for example see[9] that $\mathsf{PA} \rhd \mathsf{PA} + \mathsf{InCon}(\mathsf{PA})$. It is well known that PA is essentially reflexive, so we use Corollary 4.10.

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

Thus, it is sufficient to show that $\mathsf{PA}+\mathsf{InCon}(\mathsf{PA})$ is Π_1-conservative over PA.

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

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 $\Box\bot \to \pi$ and we conclude π. A contradiction.

4.2 Characterizations and functors

In this subsection, 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 Orey-Hájek and the Friedman becomes fully perspicuous.

We shall work extensively with the notion of local interpretability. Basically, 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.[10] In this subsection we will not worry about these subtle distinctions, assuming that our metalanguage 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 4.16, 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 4.1 expressed the following basic insight.

Theorem 4.11. $h_V : \mho_V \rhd V$.

Here, h_V denotes the 'henkin interpretation' which is the syntactic variant of the henkin model.

In this subsection we distinguish three kinds of reflexivity.

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

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

- 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 $V \rhd_{\mathsf{loc}} \mho_V$.

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

Proof. (sketch) Given some fixed $n \in \omega$, we are to interpret $\mathsf{S}^1_2 + \mathsf{con}_n(V)$. Going from an 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[11], 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. $\dashv$

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

4.2.1 Reflexivization as a Functor

In this subsubsection, we treat the basic insight (Theorem 4.13), from which, in combination with Theorem 4.11, many others will follow by simple semi-modal arguments. Theorem 4.13 also tells us that $\mho_{(\cdot)}$ can act as a functor between various categories.

Theorem 4.13. *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 $j : U \rhd V_n$. 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:[12] $\forall\, \phi{\in}\alpha_{V,n}\, \exists r{<}k\;\; r{:}\Box_{U,m} \phi^j$. 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,

[11] It would be good to explicitly define truth-predicates for the formula complexities as used in [Ger03]. Alternatively, the formula complexities from [Vis93] for which we know to have truth predicates could be linked to the ones used in [Ger03].

[12] We see that a 'carefree metatheory' for this argument should be something like EA plus Σ_1-collection.

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, $\mathsf{con}_n(V)$. $\dashv$

Theorem 4.13 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 4.13 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[13] $U \mapsto \mathsf{S}^1_2 + \mathsf{con}(U)$. We collect some immediate consequences of Theorem 4.13.

Lemma 4.14.
(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 4.13, 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 4.11 and 4.13, we find that: $V \equiv_{\mathsf{loc}} U \equiv_{\mathsf{loc}} \mho_U \equiv \mho_V$. Hence V is ℓ-reflexive. $\dashv$

Lemma 4.15. *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. $\dashv$

Lemma 4.16. *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 4.11 indeed $\mho_U \rhd U$. By the ℓ-reflexivity of U we get that $U \equiv_{\mathsf{loc}} \mho_U$, whence, by Theorem 4.13 $\mho_U = \mho_{\mho_U}$ 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 4.15 that $T \rhd U$. By Theorem 4.13 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. $\dashv$

Note that the reflexivization of the theory of pure identity is S^1_2, which is finitely axiomatized and, hence, not reflexive. Since $\mho_V$ is itself ℓ-reflexive, we always have that $\langle \mho^2_V, \mathsf{id}\rangle$ is reflexive.

Open Question 4.17. Give an example of a U that is not ℓ-reflexive, but where $\langle \mho_U, \mathsf{id}\rangle$ is reflexive. ❐

Lemma 4.18. *Suppose U is ℓ-reflexive. Then, $U \rhd_{\mathsf{loc}} V$ iff $\mho_U \supseteq \mho_V$.*

[13] 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, $U \rhd_{\mathsf{loc}} \mho_U \supseteq \mho_V \rhd V$. $\dashv$

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

Lemma 4.18 tells us that reflexivization is an embedding of $\mathsf{DEG}^{\mathsf{loc}}_{\mathsf{lr}}$ in THRY. The ℓ-reflexivity is really necessary, as $\mho$ is in general not an embedding.

For example, let SEIN be the theory of pure identity and let ONE be $\mathsf{SEIN} + \forall x, y\ x = y$. Then, $\mho_{\mathsf{SEIN}} = \mathsf{S}^1_2 = \mho_{\mathsf{ONE}}$. But not $\mathsf{SEIN} \equiv_{\mathsf{loc}} \mathsf{ONE}$. It would be interesting to see a less trifling counterexample.

4.2.2 The Orey-Hájek Characterization

We can now reformulate the Orey-Hájek characterizations using local interpretability. All the characterizations are direct consequences of Theorem 4.13. Here is the first Orey-Hájek characterization.

Theorem 4.19 (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 4.20 (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 4.21. *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 4.21, 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[14] of Theorem 4.21.

$$U \rhd_{\mathsf{loc}} V \Leftrightarrow \mho_U \rhd V \tag{22}$$

We reformulate this equivalence a bit, to make the adjunction fully explicit. We now treat, par abus de langage, $\mho$ as a functor from $\mathsf{DEG}^{\mathsf{loc}}_{\mathsf{lr}}$ to $\mathsf{DEG}_{\mathsf{lr}}$. Let emb be the embedding functor of $\mathsf{DEG}_{\mathsf{lr}}$ in $\mathsf{DEG}^{\mathsf{loc}}_{\mathsf{lr}}$. Now (22)

[14] 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.

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 4.16 is more informative.

4.2.3 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 4.22. $\digamma_U \equiv \mho_U$.

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 4.23. *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 sequentiallity 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 5.1.)

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

Lemma 4.24 tells us that $\mathsf{Cnt}_\Gamma(\cdot)$ is a functor from $\mathsf{DEG}^{\mathsf{loc}}$ to THRY.

Lemma 4.25.

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,

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

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 $\mathsf{S}^1_2 \vdash \forall z{\in}J\ 2^z{\downarrow}$. ⊣

4.2.4 Conservativity

We define two notions of conservativity as follows.

- V is ℓ-Γ-conservative over U, or: $U \rhd_{\mathsf{loc},\Gamma} V$, iff $U \rhd_{\mathsf{loc}} \mathsf{Cnt}_\Gamma(V)$.
- Suppose $\langle U, j\rangle$ and $\langle V, k\rangle$ are numberized. Then, $\langle V, k\rangle$ is Γ-conservative over $\langle U, j\rangle$, or: $\langle U, j\rangle \rhd_\Gamma \langle V, k\rangle$, iff $j : U \rhd \mathsf{Cnt}_\Gamma(\langle V, k\rangle)$.

Note that the notion of ℓ-Γ-conservativity is, in general, not a very interesting notion. It only makes sense for certain special classes Γ.

Lemma 4.26. *Suppose U and V are ℓ-reflexive. Then,*

$$U \rhd_{\mathsf{loc},\forall\Pi_1^b} V \Leftrightarrow U \rhd_{\mathsf{loc}} V.$$

Lemma 4.27. *Suppose $\langle U, j\rangle$ and $\langle V, k\rangle$ are numberized theories. Then,*

$$\langle U, j\rangle \rhd_\Gamma \langle V, k\rangle \Leftrightarrow \mathsf{Cnt}_\Gamma(\langle U, j\rangle) \supseteq \mathsf{Cnt}_\Gamma(\langle V, k\rangle).$$

4.2.5 The Friedman Functor

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

Lemma 4.28. *Suppose U is ℓ-reflexive. Then, $\mho_U^+$ and $\mho_U$ prove the same $\forall\Pi_1^b$-sentences.*

Proof. Consider $\pi \in \forall\Pi_1^b$. 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 [Vis90b] or [Vis92]), we have, for some cut J, $\mathsf{S}_2^1 + \mathsf{con}_n(U) \vdash \pi^J$. Let $k : U \rhd (\mathsf{S}_2^1 + \mathsf{con}_n(U))$. We have, for sufficiently large m,

$$\begin{array}{rcl} \mho_U \vdash \neg\pi & \rightarrow & \Box_{U,m}(\neg\,\pi^J)^k \\ & \rightarrow & \Box_{U,m}(\Box_{U,n}\bot)^k \\ & \rightarrow & \Box_{U,m}\bot \\ & \rightarrow & \bot \end{array}$$

Hence $\mho_U \vdash \pi$. ⊣

Lemma 4.29. *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 4.28. ⊣

By cut elimination (see [Ger03] and [Ger0X]), we can show that over EA we may replace the $\mathsf{con}_n(U)$ by $\mathsf{cutfree\text{-}con}(U[n])$.[15]

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

- U is ℓ-reflexive iff $U \rhd (\mathsf{S}_2^1 + \mathsf{cutfree\text{-}con}(U))$.
- $\mho_U^+ = \mathsf{EA} + \mathsf{cutfree\text{-}con}(U)$.

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

Theorem 4.30. *Suppose U and V are finitely axiomatized and ℓ-reflexive, then:*

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

[15] Alternatively, we can show that, for finitely axiomatized V, for some V-cut J, we have $\mathsf{S}_2^1 \vdash \mathsf{cutfree\text{-}con}(V) \rightarrow \mathsf{con}_n^J(V)$. This uses the construction of an interpretation k via a variant of the henkin construction. Subsequently, we produce a proof of n-reflection using an appropriate truth-predicate true_m: $\mathsf{S}_2^1 \vdash \forall\phi\,(\Box_{V,n}^J\phi \rightarrow \mathsf{true}_m(\phi^k))$. Finally, we apply the results of Wilkie and Paris to obtain: $\mathsf{EA} \vdash \mathsf{cutfree\text{-}con}(V) \rightarrow \mathsf{con}_n(V)$.

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.

We can view $\mho^+$ as a functor from $\mathsf{DEG}_{\mathsf{loc}}$ to AR, the ordering of extensions of Q in the language of arithmetic with the subset relation. We can show that this functor preserves sums.

Lemma 4.31. *Let $W = U \oplus V$. Then, for every k, there is an n and a S^1_2-cut J, such that $\mathsf{S}^1_2 + \mathsf{con}_n(U) + \mathsf{con}_n(V) \vdash \mathsf{con}^J_k(W)$. By the results of Wilkie and Paris, it follows that for every k, there is an n, such that $\mathsf{EA} + \mathsf{con}_n(U) + \mathsf{con}_n(V) \vdash \mathsf{con}_k(W)$.*

Proof. Choosing n sufficiently large, we can construct interpretations

$$i : (\mathsf{S}^1_2 + \mathsf{con}_n(U)) \rhd U_k \text{ and } i' : (\mathsf{S}^1_2 + \mathsf{con}_n(V)) \rhd V_k.$$

Using these interpretations, we can construct an interpretation $j : (\mathsf{S}^1_2 + \mathsf{con}_n(U)) \rhd W_k$. Now we can construct a satisfaction-predicate for W-formulas of complexity k in $\mathsf{S}^1_2 + \mathsf{con}_n(U)$, adapted to j. This predicate gives us the usual proof of $\mathsf{S}^1_2 + \mathsf{con}_n(U) + \mathsf{con}_n(V) \vdash \mathsf{con}^J_k(W)$, for a suitable cut J. ⊣

It follows that:

Lemma 4.32. $\mho^+_{U\oplus V} = \mho^+_U \cup \mho^+_V$.

We end with an insight in the relation between $\mho$ and $\mho^+$.

Lemma 4.33. $\mho^+ \circ \mho = \mho^+$.

Proof. Since $\mho_U \rhd U$, we immediately have $\mho^+_{\mho_U} \supseteq \mho^+_U$. In the converse direction, it is sufficient to prove that, for any k (that is sufficiently large to make the statement meaningful):

(†) $\mathsf{EA} + \mathsf{con}_k(U) \vdash \mathsf{con}_k(\mathsf{S}^1_2 + \mathsf{con}_k(U))$.

To prove (†) it is, by the results of Wilkie and Paris, sufficient to show:

(‡) For some definable cut J, $\mathsf{S}^1_2 + \mathsf{con}_k(U) \vdash \mathsf{con}^J_k(\mathsf{S}^1_2 + \mathsf{con}_k(U))$.

But this last fact is immediate from the ℓ-reflexivity of $\mho_U$. (Since $\mho_U$ is sequential, we have the consistency statements always on definable cuts, by Pudlák's Result. See Lemma 3.9.) ⊣

4.3 End-extensions

We have a model-theoretic characterization of interpretability between extensions of PA in the language of PA(see Theorem 3.8). It is simply that $U \rhd V$ iff every model $\mathcal{M}$ of U has an endextension $\mathcal{N}$ satisfying V. In this subsection 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 4.34. *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 $l : \mathcal{M} \preceq_{\mathsf{end}} \mathcal{N}$.*

We will present two proofs of the theorem as both proofs have some interest on their own. Here is the first proof.

Proof. Suppose U is sequential and e-reflexive. We first find m. Pick any $l : U \rhd \mathsf{S}^1_2$. By Lemma 4.23, we can find $p : U \rhd \nabla_{U,l}$. Recall that $\nabla_{U,l}$ contains U. We take $m := p \circ \mathcal{I}$.

(1) $\Rightarrow$ (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 5 to support our argument. Let us first resume the list of interpretations that will be used in our proof.

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

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 $\mathcal{M}^\star$:

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

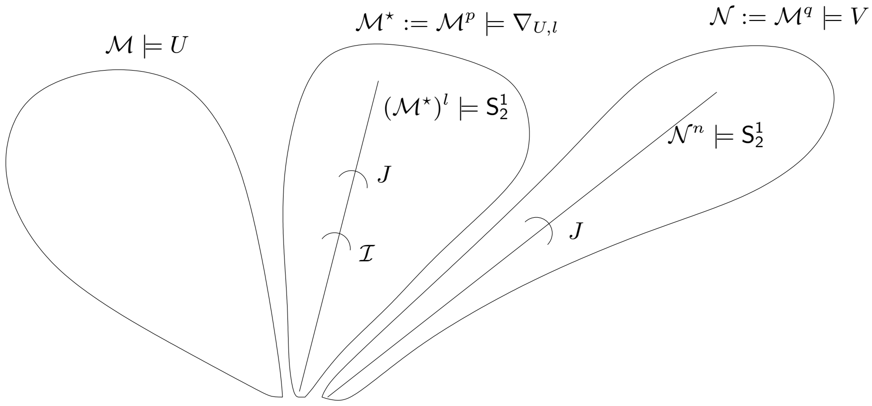

Figure 5: End extension theorem

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 $U \vdash \mathsf{con}^k_n(V)$. Hence $l : U \rhd \mho_V$, and, thus $U \rhd V$. $\dashv$

We give the second proof of Theorem 4.34.

Second Proof of Theorem 4.34. Our alternative only concerns the step (1) $\Rightarrow$ (2). Suppose $m_0 : U \rhd \mho_U$. We can shorten m_0 to m_1 such that (i) there is a Σ_1-truth-predicate for the Σ_1-sentences inside m_1 definable in U and (ii) m_1 satisfies $I\Delta_0 + \Omega_1 + B\Sigma_1$. We will also have: $m_1 : U \rhd \mho_U$.

Let m_2 be the interpretation that extends m_1 by adding the Σ_1-truth-predicate. Now consider any theory V. We assume that the axiom set of V is not too complex. So, in the de-luxe circumstances of $I\Delta_0 + \Omega_1 + B\Sigma_1$ provability in V will be Σ_1.

Now we build in $\mho_V + I\Delta_0 + \Omega_1 + B\Sigma_1$, using the Feferman predicate, an interpretation q of V. Note that this interpretation is restricted: we have a satisfaction-predicate for the interpretation that works for standard formulas. We can construct the isomorphic cut J between m_1 and q. Say the isomorphism is Φ. For any $m' : V \rhd \mathsf{S}^1_2$, we can define:

- $J'(x) :\Longleftrightarrow J(x) \wedge \exists y\ (x(\Phi)y \wedge \delta_{q\circ m'}(y))$.

Note that Φ will give an initial embedding from J' to $q \circ m'$. The construction of J' is uniform in the provability predicate of V. So in the theory $\heartsuit_V$ which is $\mho_V + I\Delta_0 + \Omega_1 + B\Sigma_1$ extended with a Σ_1-truth-predicate and the appropriate axioms, we can give J' a fixed complexity, say n, independent of the specific V we are considering. Let sat_n be the satisfaction predicate for this complexity. We define:

- $J^\star(x) :\Longleftrightarrow \forall y{\in}\Gamma_n\ (\mathsf{cut}(\{z \mid \mathsf{sat}_n(y,z)\}) \rightarrow \mathsf{sat}_n(y,x))$.

We take $m := m_2 \circ J^\star$.

Now suppose $U \rhd V$. By Theorem 4.19, it follows that $m_2 : U \rhd \heartsuit_V$. We build, inside m_2, a restricted interpretation q of V, which has an appropriately simple satisfaction predicate for standard formulas. Consider $k := m_2 \circ q : U \rhd V$. Let $\mathcal{M}$ be any model of U. Consider $\mathcal{N} := \mathcal{M}^k$. Suppose $\mathcal{N}^{m'} \models \mathsf{S}^1_2$. We can find a cut J of $\mathcal{M}^{m_2}$ that is isomorphic to a cut of $\mathcal{N}^{m'}$, such that J is given by a formula of complexity n. By construction $\mathcal{M}^m$ is a cut of $(\mathcal{M}^{m_2})^J$ and we are done. $\dashv$

5 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 is some sort of modal system as described in Subsection 3.4. The P-style proofs are based on finite approximations of interpretability. This behavior is captured also in a modal-like system as we shall see below.

5.1 Finite approximations of interpretability

It is a triviality that for finitely axiomatized theories we have $\alpha \rhd \beta \rightarrow \Box(\alpha \rhd \beta)$. 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[16] $T + \alpha \rhd T' + \beta$. 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.

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

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 verifyably 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 little trick a-là Craig 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 5.1. If $k : T+\alpha \rhd T+\beta$, we define T^k as follows.

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

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

Lemma 5.2. *(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 $\Box(\Box_{T^k}\varphi \to \Box_T\varphi)$ we shall in general not have $\Box(\Box_T\varphi \to \Box_{T^k}\varphi)$.

Lemma 5.3. $\mathsf{S}^1_2 \vdash k : T+\alpha \rhd T+\beta \to \Box_T(T+\alpha \rhd T^k+\beta)$

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

$$\Box_T\Box_{T+\alpha}\beta^k. \tag{23}$$

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 $\mathsf{axioms}_{T^k+\beta}(\tau)$, there is a proof $p' \leq y$ with $\mathsf{proof}_{T+\alpha}(p', \tau^k)$.

If $\mathsf{axioms}_{T^k+\beta}(\tau)$, either $\tau = \beta$ and we are done by (23), or we have that $\mathsf{axioms}_{T^k}(\tau)$. Let us consider the latter case. Again, if $\tau = s_2^1$ 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 $x \cdot f(x, k)$. $\dashv$

We note that we may replace $\rhd_s$ in the antecedent of Lemma 5.3 by $\rhd_t$.

5.2 Finite approximations and modal logics

Just as in Subsection 3.4, we can make some sort of modal system in which facts about finite 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 5.2 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 5.4. 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 5.5. $\mathbf{IL}\mathsf{X} = \mathbf{IL}\mathsf{X}^\Box$

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. $\dashv$

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 $k : \top \rhd_T \top$.

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

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 *extensionallity 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^1_2, many arguments like $\mathsf{L}_1^{\mathsf{k}}$-$\mathsf{L}_3^{\mathsf{k}}$ and $\mathsf{J}_5^{\mathsf{k}}$ go as always. $\mathsf{J}_1^{\mathsf{k}}$ is easy. But if it is under a $\Box$, we need $(\rightarrow \Box)^{\mathsf{k}}$. Now, $(\rightarrow \Box)^{\mathsf{k}}$ itself together with $(\mathsf{E}\Box)^{\mathsf{k}}$ is just Lemma 5.2, and $(\mathsf{E}\rhd)^{\mathsf{k}}$ is a direct consequence of it. Finally, P^{k} is just Lemma 5.3.

5.3 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 of R and W.[17] 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. Per principle we shall give a proof in $\mathbf{IL}\mathsf{P}$ and in $\mathbf{IL}\mathsf{M}$. These proofs can then be copied almost literally to yield arithmetical soundness proofs.

5.3.1 The principle W

Lemma 5.6. $\mathbf{IL}\mathsf{P} \vdash \mathsf{W}$ *and* $\mathbf{IL}\mathsf{P}_{\mathsf{R}} \vdash \mathsf{W}$

Proof.

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

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

[17]In Gorris Joosten, [GJ04] a principle is given that is precisely W and R together.

$B \wedge \Diamond A \rhd B \wedge \Box\neg A$. 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 5.7. **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 $(*)$

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

M-style soundness proof of W We assume $j : \alpha \rhd \beta$ and fix the corresponding Pudlák cut J. By Lemma 3.13, $\alpha \rhd \alpha \wedge \Box^J\neg\alpha$, whence by $\mathsf{M^J}$ and J_2, $\alpha \rhd \beta \wedge \Box\neg\alpha$.

5.3.2 The principle M_0

Lemma 5.8. **IL**$\mathsf{P} \vdash \mathsf{M}_0$ *and* **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 5.9. **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.

$$\begin{array}{lcll}
\alpha \rhd \beta & \to & \Box(\alpha \rhd^k \beta) & \mathsf{J_4^k} \\
 & \to & \Box(\Diamond\alpha \to \Diamond^k\beta) & \\
 & \to & \Box(\Diamond\alpha \wedge \Box\gamma \to \Diamond^k\beta \wedge \Box\gamma) & \\
 & \to & \Diamond\alpha \wedge \Box\gamma \rhd \Diamond^k\beta \wedge \Box\gamma & \text{o.a. by } \mathsf{L_2^k} \\
 & \to & \Diamond\alpha \wedge \Box\gamma \rhd \Diamond^k(\beta \wedge \Box\gamma) & \text{by } \mathsf{J_5^k} \\
 & \to & \Diamond\alpha \wedge \Box\gamma \rhd \beta \wedge \Box\gamma &
\end{array}$$

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 $\mathsf{L_2^J}$ for this particular J we get $\Diamond\alpha \wedge \Box\gamma \to \Diamond\alpha \wedge \Box\Box^J\gamma$, whence

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

5.3.3 The principle $\mathsf{W^*}$

Lemma 5.10. $\mathbf{IL}\mathsf{P} \vdash \mathsf{W^*}$ *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

$$\Box(\Box\neg B \to \Box\neg A) \tag{24}$$

and

$$\Box(\Diamond A \wedge \Box C \to \Diamond B \wedge \Box C) \tag{25}$$

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.

$$\begin{array}{lcll}
B \wedge \Box C \wedge \Diamond A & \rhd & \Diamond A \wedge \Box C & \text{by (25)} \\
 & \rhd & \Diamond B \wedge \Box C & \text{by } \mathsf{L_3} \\
 & \rhd & \Diamond(B \wedge \Box\neg B) \wedge \Box C & \text{by } \mathsf{L_2} \\
 & \rhd & \Diamond(B \wedge \Box C \wedge \Box\neg B) & \text{by } \mathsf{J_5} \\
 & \rhd & B \wedge \Box C \wedge \Box\neg B & \text{by (24)} \\
 & \rhd & B \wedge \Box C \wedge \Box\neg A &
\end{array}$$

$\dashv$

Lemma 5.11. $\mathbf{IL}\mathsf{M} \vdash \mathsf{W^*}$

Proof. So, in $\mathbf{IL}\mathsf{M}$, assume $A \rhd B$. $B \wedge \Box C \rhd (B \wedge \Box C \wedge \Box\neg A) \vee (B \wedge \Box C \wedge \Diamond A)$. 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.

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

We follow the modal proof.

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

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

$$\begin{array}{llll} \beta \wedge \Box\gamma \wedge \Diamond\alpha & \rhd & \Diamond\alpha \wedge \Box\gamma & \text{by } \mathsf{L}_3^{\mathsf{J}} \\ & \rhd & \Diamond(\alpha \wedge \Box^J\neg\alpha) \wedge \Box\gamma & \text{by } \mathsf{L}_2^{\mathsf{J}} \\ & \rhd & \Diamond(\alpha \wedge \Box^J\gamma \wedge \Box^J\neg\alpha) & \text{by } \mathsf{J}_5 \\ & \rhd & \alpha \wedge \Box^J\gamma \wedge \Box^J\neg\alpha & \text{by } \mathsf{M}^{\mathsf{J}} \\ & \rhd & \beta \wedge \Box\gamma \wedge \Box\neg\alpha & \end{array}$$

5.3.4 The principle P_0

Lemma 5.12. **IL**$\mathsf{P} \vdash \mathsf{P}_0$

Proof. Within **IL**P: $A \rhd \Diamond B \to \Box(A \rhd \Diamond B) \to \Box(A \rhd B)$. ⊣

Lemma 5.13. **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. ⊣

Lemma 5.14. **IL**$\mathsf{M} \vdash \mathsf{P}_0$

Proof.

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

⊣

P-style soundness proof of P_0 The proof goes conform the modal proof. Thus, for some k, $\alpha \rhd \Diamond\beta \to \Box(\alpha \rhd^k \Diamond\beta)$. But as also $\Box(\Diamond\beta \rhd \beta)$, we get by $\mathsf{J}_2^{\mathsf{k}}$ the required $\alpha \rhd \Diamond\beta \to \Box(\alpha \rhd \beta)$.

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

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

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

5.3.5 The principle R

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

Lemma 5.15. **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 5.16. **IL**$\mathsf{P} \vdash \mathsf{P}_0$

Proof. $A \rhd B \to \Box(A \rhd B)$. Using this together with Lemma 5.15 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 5.17. **IL**$\mathsf{P}_{\mathsf{R}} \nvdash R$

Proof. By exposing a countermodel as in the proof of Lemma 5.13. ⊣

Lemma 5.18. **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 J_1^k, $\beta \rhd^k \Diamond\neg\gamma$. Thus, by J_2^k and J_5, we get $\beta \rhd \neg\gamma$. From $\alpha \rhd^k \beta$ we get, again by J_2^k that $\alpha \rhd \neg\gamma$. We have not used any extensionallity axioms, thus also

$$\Box((\alpha \rhd^k \beta) \wedge \neg(\alpha \rhd \neg\gamma) \to \Diamond^k(\beta \wedge \Box\gamma)) \tag{26}$$

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

$$\begin{array}{lcll} \neg(\alpha \rhd \neg\gamma) & \rhd & \neg(\alpha \rhd \neg\gamma) \wedge (\alpha \rhd^k \beta) & \text{by (26)} \\ & \rhd & \Diamond^k(\beta \wedge \Box\gamma) & \text{by } \mathsf{J}_5^\mathsf{k} \\ & \rhd & \beta \wedge \Box\gamma. & \end{array}$$

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 3.14 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

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

5.3.6 Mixing proof styles

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

Lemma 5.19. *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

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

But, by Lemma 3.14 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 $\mathbf{IL}\mathsf{R}$.

Lemma 5.20. $\mathbf{IL}\mathsf{R} \vdash A \rhd B \to \neg(A \rhd \neg C) \wedge \Box D \rhd B \wedge \Box(C \wedge D)$

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

Lemma 5.21. $\mathbf{IL}\mathsf{R} \vdash A \rhd \Diamond B \to \Box(\neg(A \rhd \neg C) \to \Diamond(B \wedge C))$

Proof. In **IL**R we get

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

⊣

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.

References

[Ber90] A. Berarducci. The interpretability logic of Peano arithmetic. *Journal of Symbolic Logic*, 55:1059–1089, 1990.

[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.

[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] Philipp 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.

[Ger0X] Philipp Gerhardy. the role of quantifier eliminations in cut elimination. *Notre Dame Journal of Formal Logic*, 200X.

[GJ04] E. Goris and J.J. Joosten. Modal matters in interpretability logics. Logic Group Preprint Series 226, University of Utrecht, March 2004.

[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.

[HP93] P. Hájek and P. Pudlák. *Metamathematics of First Order Arithmetic*. Springer-Verlag, Berlin, Heidelberg, New York, 1993.

[Joo03] 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.

[JV00] J.J. Joosten and A. Visser. The interpretability logic of *all* reasonable arithmetical theories. *Erkenntnis*, 53(1–2):3–26, 2000.

[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.

[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.

[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.

[TMR53] A. Tarski, A. Mostowski, and R. Robinson. *Undecidable theories*. North–Holland, Amsterdam, 1953.

[Vis90a] A. Visser. Interpretability logic. In P.P. Petkov, editor, *Mathematical Logic*, pages 175–208. Plenum Press, New York, 1990.

[Vis90b] A. Visser. Interpretability logic. In *[Pet90]*, pages 175–209, 1990.

[Vis91] A. Visser. The formalization of interpretability. *Studia Logica*, 50(1):81–106, 1991.

[Vis92] A. Visser. An inside view of 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.

[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.

[WP87] A. Wilkie and J.B. Paris. On the scheme of of induction for bounded arithmetic formulas. *Annals of Pure and Applied Logic*, 35:261–302, 1987.

A The Sum

The sum $U \oplus V$ is given as a theory W of signature Σ_W, where Σ_W is given as the disjoint union of Σ_U and Σ_V plus two additional fresh unary predicate symbols Δ_U and Δ_V and a new binary identity symbol E_W.[18] Let τ_U and τ_V be the obvious translations of the languages of U, respectively V into the language of W, where we relativize to Δ_U in the first case and to Δ_V in the second case. We take W to be axiomatized by the following axioms.

- $\vdash P_{\tau_U}\vec{v} \rightarrow \vec{v} : \Delta_U$,
- $\vdash P_{\tau_V}\vec{v} \rightarrow \vec{v} : \Delta_V$,
- $\vdash A^{\tau_U}$, for A a U-axiom,
- $\vdash A^{\tau_V}$, for A a V-axiom,

[18] Of course, if $U = V$, we take the appropriate measures to make the Δ disjoint.

- $\vdash \forall x\,(x : \Delta_U \vee x : \Delta_V)$,
- $\vdash xE_Wy \iff \forall z\,((xE_Uz \iff yE_Uz) \wedge (xE_Vz \iff yE_Vz))$.

Note that, in the presence of the other axioms, the last axiom says that E_W is the crudest congruence relation w.r.t. all predicates of W.

It is easy to check that $\oplus$ gives us the supremum, i.e. the sum in the sense of category theory, in DEG and $\mathsf{DEG}^{\mathsf{loc}}$.

We define $U \boxplus V$ as the disjoint union of U and V. this is $U \oplus V$ with the further axiom that Δ_U and Δ_V are disjoint. (It is easy to see that we can simplify the definition a bit, having just one Δ and with E_W the union of E_U and E_V.)

Theorem A.1. *$U \oplus V$ is consistent iff both U and V are consistent. It follows that, if $W \oplus Z \vdash \phi^{\mathsf{in}_0} \vee \psi^{\mathsf{in}_1}$, then $W \vdash \phi$ or $Z \vdash \psi$. (These facts can be verified in $I\Delta_0 + \mathsf{supexp}$.) Similarly for $\boxplus$.*

Theorem A.2. *Consider any formula $\phi\vec{x}$. Partition $\vec{x}$ into $\vec{y}$ and $\vec{z}$. We have that $\phi\vec{y}\vec{z}$ is provably equivalent in $(U \boxplus V) + \vec{y} : \Delta_U + \vec{z} : \Delta_V$ to a boolean combination of U-formulas $\psi\vec{y}$ and V-formulas $\chi\vec{z}$.*

Proof. The proof is by induction of ϕ, quantifiying over all possible partitions of he free variables. $\dashv$

We prove, in our context, a theorem well-known from the research on chapters. See [MPS90].

Lemma A.3. *Suppose W is a theory with a pairing function working for all objects. Suppose $K : U \boxplus V \rhd W$. Then, for some k, $U \boxplus V$ proves that there at most k elements of δ_K modulo E_K in either in Δ_U or in Δ_V.*

Proof. Let W and K be as in the statement of the theorem. Reason in $U \boxplus V$. By Theorem A.2, there are U-formulas $\phi_i(x, z)$ and V-formulas $\psi_i(y)$, for $i = 0, \ldots, n-1$, such that, for all $x, z : (\delta_K \cap \Delta_U)$ and, for all $y : (\delta_K \cap \Delta_V)$, we have $\mathsf{pair}(x, y, z) \iff \bigvee_i(\phi_i(x, z) \wedge \psi_i(y))$. Similarly, there are U-formulas $\phi^\star_j(y, z)$ and V-formulas $\psi^\star_j(x)$, for $j = 0, \ldots, n^\star-1)$ such that, for all $y, z : (\delta_K \cap \Delta_V)$ and, for all $x : (\delta_K \cap \Delta_U)$, we have $\mathsf{pair}(x, y, z) \iff \bigvee_j(\phi^\star_j(y, z) \wedge \psi^\star_j(x))$.

Let m be the maximum of $n, n^\star$. Consider $x_0, \ldots, x_{2m}, y_0, \ldots, y_{2m}$, where the x_i are in Δ_U and the y_i are in Δ_V. We assume further that the x_i are pairwise E_K-disjoint and, similarly, for the y_i. Suppose $\mathsf{pair}(x_i, y_i, z_i)$, for $i = 0, \ldots, 2m$. Clearly, for one of Δ_U, Δ_V, there are at least $m+1$ of the z_i in that set. Without loss of generality we may assume that $z_0, \ldots, z_m$ are in Δ_U. So we have that, for each $j \in \{0, \ldots, m\}$, there is an $i \in \{0, \ldots, n-1\}$, such that $\phi_i(x_j, z_j)$ and $\psi_i(y_j)$. By the Pidgeon Hole Principle, for some i, there are j, j' with $j \neq j'$, such that $\phi_i(x_j, z_j)$ and $\psi_i(y_j)$ and $\phi_i(x_{j'}, z_{j'})$ and $\psi_i(y_{j'})$. It follows that $\phi_i(x_j, z_j)$ and $\psi_i(y_{j'})$. Hence, $\mathsf{pair}(x_j, y_{j'}, z_j)$. Since we have both $\mathsf{pair}(x_j, y_j, z_j)$. and $\mathsf{pair}(x_j, y_{j'}, z_j)$, we may conclude that $y_jE_Ky_{j'}$. A contradiction.

So it follows that either in Δ_U or in Δ_V there are less than $2m$ elements of δ_K, modulo E_K. $\dashv$

Lemma A.4. *Suppose W is a theory with a pairing function working for all objects. Supose further that W has two provably definable, provably distinct objects, say 0 and 1. Suppose $K : U \boxplus V \rhd W$. Then, we can find a $\tilde{K} : U \boxplus V \rhd W$, such that $U \boxplus V$ proves that $\delta_{\tilde{K}}$ is either entirely in Δ_U or entirely in Δ_V.*

Proof. By Lemma A.3, for some k, $U \boxplus V$ proves that there are at most k elements of δ_K in one of Δ_U, Δ_V. Consider n with $2^n > k$. For any $0,1$-sequence $\sigma = b_0 \cdots b_{n-1}$ of we define a W-formula $z(F_\sigma)w$ as follows:

- $z(F_\varepsilon)w :\iff z = w$,
- $z(F_{b\tau})w :\iff \exists u\ (z(F_\tau)u \wedge \mathsf{pair}(b, u, z))$.

We define the interpretation $M_\sigma : W \to W$ as follows:

- $\delta_{M_\sigma}(x) :\iff \exists y\ y(F_\sigma)x$,
- $P_{M_\sigma}(\vec{x}) :\iff \exists \vec{y}\ (\vec{y}(F_\sigma)\vec{x} \wedge P(\vec{y}))$.

We take $K_\sigma := K \circ M_\sigma$. Suppose $\sigma_0, \ldots, \sigma_{m-1}$ enumerates the $0,1$-sequences of length n. We define:

- $K^{(0)} := K_{\sigma_0}$,
- $K^{(j+1)} := K_{\sigma_{j+1}} \langle \delta_{K_{\sigma_{j+1}}} \subseteq \Delta_U \vee \delta_{K_{\sigma_{j+1}}} \subseteq \Delta_V \rangle K^{(j)}$,
- $\tilde{K} := K^{(m)}$.

It is easily seen that $\tilde{K}$ is a promised. ⊣

Theorem A.5. *Suppose W is a theory with a pairing function working for all objects. Suppose further that W has two provably definable, provably distinct objects, say 0 and 1. Suppose $K : U \boxplus V \rhd W$. Then, we can find a $K^\star$ such that $K^\star : U \rhd W$ or $K^\star : V \rhd W$.*

Proof. Let K and W be as stipulated in the conditions of the theorem. By Lemma A.4, we may replace K be $\tilde{K}$, such that $U \boxplus V \vdash \delta_{\tilde{K}} \subseteq \Delta_U \vee \delta_{\tilde{K}} \subseteq \Delta_V$. By Theorem A.2, there are U-sentences $\phi_0, \ldots, \phi_{n-1}, \phi^\star_0, \ldots, \phi^\star_{n^\star-1}$, and V-sentences $\psi_0, \ldots, \psi_{n-1}, \psi^\star_0, \ldots, \psi^\star_{n^\star-1}$ such that:

- $U \boxplus V \vdash \delta_{\tilde{K}} \subseteq \Delta_U \iff \bigvee_i (\phi_i \wedge \psi_i)$,
- $U \boxplus V \vdash \delta_{\tilde{K}} \subseteq \Delta_V \iff \bigvee_j (\phi^\star_j \wedge \psi^\star_j)$

It follows that $U \boxplus V \vdash \bigvee_i \phi_i \vee \bigvee_j \psi^\star_j$. Hence, by Theorem A.1, we find $U \vdash \bigvee_i \phi_i$ or $V \vdash \bigvee_j \psi^\star_j$.

Without loss of generality we may assume that the first case obtains. Also we may assume that the $\phi_i \wedge \psi_i$ are consistent with $U \boxplus V$ —otherwise we may omit them from our disjunction. It is sufficient to provide, for each i, $K^{[i]} : (U + \phi_i) \rhd W$. Then we can take:

- $K^\star := K^{[0]} \langle \phi_0 \rangle (K^{[1]} \langle \phi_1 \rangle (\ldots \langle \phi_{n-2} \rangle K^{[n-1]}) \cdots)$.

Consider $(U + \phi_i) \boxplus (V + \psi_i)$. This is a consistent theory. We have $(U + \phi_i) \boxplus (V + \psi_i) \vdash \delta_{\tilde{K}} \subseteq \Delta_U$. It follows that we have:

- $(U + \phi_i) \boxplus (V + \psi_i) \vdash \delta_{\tilde{K}}(x) \iff (\Delta_U(x) \wedge \bigvee_p (\chi^\delta_p(x) \wedge \rho^\delta_p))$, where χ^δ_p is a U-formula and ρ^δ_p is a V-sentence;

- $(U + \phi_i) \boxplus (V + \psi_i) \vdash P_{\tilde{K}}(\vec{x}) \iff (\vec{x} : \Delta_U \wedge \bigvee_q (\chi_q^P(\vec{x}) \wedge \rho_q^P))$, where χ_q^P is a U-formula and ρ_q^P is a V-sentence.

Consider the set R all the V-sentences ρ involved in these equivalences. With each subset X of R we associate a formula γ_X which is the conjunction of all the sentences in X and of all the negations of the sentences in $R \setminus X$. Clearly $V + \psi_i + \gamma_X$ must be consistent for some X. We consider some such X. Now note that:

- $(U+\phi_i)\boxplus(V+\psi_i+\gamma_X) \vdash \delta_{\tilde{K}}(x) \iff (\Delta_U(x) \wedge \bigvee\{\chi_p^\delta(x) \mid \rho_p^\delta \in X\})$,
- $(U + \phi_i) \boxplus (V + \psi_i + \gamma_X) \vdash P_{\tilde{K}}(\vec{x}) \iff (\vec{x} : \Delta_U \wedge \bigvee\{\chi_q^P(\vec{x}) \mid \rho_q^P \in X\})$.

Define:

- $\delta_{K[i]}(x) :\iff \bigvee\{\chi_p^\delta(x) \mid \rho_p^\delta \in X\}$,
- $P_{K[i]}(\vec{x}) :\iff \bigvee\{\chi_q^P(\vec{x}) \mid \rho_q^P \in X\}$.

We have, for any axiom α of W, $(U + \phi_i) \boxplus (V + \psi_i + \gamma_X) \vdash (\alpha^{K[i]})^{\mathsf{in}_0}$. We find, by Theorem A.1, $U + \phi_i \vdash \alpha^{K[i]}$. Thus, we are done. $\dashv$

Here is a variant of Theorem A.1.

Theorem A.6. *Let $W = U \oplus V$. Then, for every k, there is an n and a S^1_2-cut J, such that $\mathsf{S}^1_2 + \mathsf{con}_n(U) + \mathsf{con}_n(V) \vdash \mathsf{con}^J_k(W)$. By the results of Wilkie and Paris, it follows that for every k, there is an n, such that $\mathsf{EA} + \mathsf{con}_n(U) + \mathsf{con}_n(V) \vdash \mathsf{con}_k(W)$*

Proof. Choosing n sufficiently large, we can construct interpretations

$$K : (\mathsf{S}^1_2 + \mathsf{con}_n(U)) \rhd U_k \text{ and } M : (\mathsf{S}^1_2 + \mathsf{con}_n(V)) \rhd V_k.$$

Using these interpretations, we can construct an interpretation $N : (\mathsf{S}^1_2 + \mathsf{con}_n(U)) \rhd W_k$. Now we can construct a satisfaction-predicate for W-formulas of complexity k in $\mathsf{S}^1_2 + \mathsf{con}_n(U)$, adapted to N. This predicate gives us the usual proof of $\mathsf{S}^1_2 + \mathsf{con}_n(U) + \mathsf{con}_n(V) \vdash \mathsf{con}^J_k(W)$, for a suitable cut J. $\dashv$

B Pudlák's lemma

Lemma B.1 (Pudlák's lemma).

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

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

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

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 posess 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 mulptiplication. 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 $\mathsf{Goodsequence}$.

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

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

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

Finally, we define J to be the closure of J' under $+$, $\cdot$ and ωx. Now that we have defined all the machinery we can start the real proof. The reader is encouridged 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 4.6.

We first note that $J'(x)$ indeed defines a U-cut. For $\Box_U J'(0)$ you basically need sequentiallity 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

$$\Box_U \forall\, x,y{\in}J\ (h(x){=^j}h(y) \leftrightarrow x=y) \qquad (27)$$

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[19]

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

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 B.2. $\forall^{\mathsf{Term}} t\ \Box_U \forall\, \vec{x}, y \in J\ (t^j(h(\vec{x})){=^j}h(y) \leftrightarrow t(\vec{x}) = y)$

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 (27) 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 (27).

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:

- $\leftarrow$ If $t_1(\vec{x}) + t_2(\vec{x}) = y$, then by Lemma 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 (27), $h(y_1 + y_2){=^j}h(y)$ and by the definition of h we get that

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

- $\rightarrow$ 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 (27), $y_1 + y_2 = y$, that is, $t_1(\vec{x}) + t_2(\vec{x}) = y$.

$\dashv$

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

$$\Box_U \forall\, \vec{x}{\in}J\ (\varphi^j(h(\vec{x})) \leftrightarrow \varphi(\vec{x})). \tag{28}$$

Again we proceed by an induction on φ.

[19] 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 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 B.2, 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 3.9 we have used twice a Σ_1-induction; In Claim B.2 and in proving (28). 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.

$\dashv$

Remark B.3. 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 3.9 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.