JUMPING IN ARITHMETIC

ALBERT VISSER

Abstract. In this paper we study a new relation between sentences: *the jump relation*. The idea of the jump relation is based on an analysis of Feferman's Theorem that the inconsistency of a theory U is interpretable over U. The jump relation is based on a converse of Feferman's Theorem: if a sentence is interpretable over a theory U, it is, in a sense, an inconsistency statement over U. We introduce an *antipode* of the inconsistency statement *the presistency statement*. The jump relation allows one to 'jump' from presistencies to inconsistencies.

We show that for a wide classes of theories U the jump relation coincides with interpretability over U and for an even wider class it coincides with Π_1-conservativity over U. Thus, the jump relation provides a new way of looking at interpretability and Π_1-conservativity. On the other hand, we will show that the jump relation admits variations that are distinct from interpretability and Π_1-conservativity.

We show that the jump relation satisfies the interpretability logic ILM.

Date: January 3, 2014.
2000 *Mathematics Subject Classification.* 03A05, 03B25, 03F25, 03F30, 03F45.
Key words and phrases. Interpretability, Provability Logic, Second Incompleteness Theorem.
I am grateful to Rosalie Iemhoff and Thomas Müller for their comments and advice.

Contents

1. Introduction

The story I want to tell you in this paper has three beginnings. Rather than choose one starting point as the 'leading' one, let me give you each of the beginnings separately.

1.1. **Role Provability Predicates.** Syntactical approaches to modality come in two flavors. A first idea is to add a predicate or predicates to a language that has sufficient coding possibilities. Then, we stipulate that the predicate, considered as a predicate of sentences, satisfies a number of desired modal properties. An important question is which properties we can consistently (or also conservatively) demand of such predicates and whether we can define a Kripke style semantics for them. For examples of this approach, see e.g. [KM60], [Mon63], [Tho80], [RL86], [HL01], [HLW03] and [SF13].

A second approach is the modal study of predicates that are definable in theories with sufficient coding possibilities. This line of research usually zooms in on specific predicates like *provability* and *interpretability*. Provability Logic is a perfect example of this study. The classical papers in this field are [Göd33], [Kre53], [Löb55], [Sol76]. For expository texts, see: [BS91], [Boo93], [Lin96], [JdJ98], [Šve00], [AB04]. There are many variations.

1. Over EA, also known as $\mathrm{I}\Delta_0 + \mathsf{exp}$, cutfree provability and ordinary provability are not equivalent. On the other hand they both validate Löb's Logic. See [Vis90] and [Kal91].
2. Over EA, provability with an oracle for Σ_1-truth and ordinary provability are not equivalent. On the other hand they both validate Löb's Logic.
3. Over PA we can consider the predicates 'provable in PA with a an oracle for Π_{n+1}-truth'. The logic of the hierarchy of such predicates is Japaridze's Logic GLP. See [Jap85]. See also [Boo93]. This logic was used by Lev Beklemishev to extract proof theoretic ordinals from its closed fragment. See e.g. [Bek04, Bek05, BJV05, Bek06].
4. We consider over the theory ZF, the predicate *truth in all transitive models of* ZF. This example was studied by Solovay in [Sol76]. See also [Boo93]. A closely related example is to consider truth in all V_κ where κ is inaccessible. See [Boo93].
5. Let PA^2 be the first-order version of second order arithmetic. We may consider the arithmetization of provability in PA^2 with the ω-rule. This predicate was studied e.g. in [Boo93].
6. Recently Graham Leach-Krouse studied an internal version of Ω-validity over ZFC with the von Neumann interpretation.

All predicates in the above list validate Löb's logic. There are however other modally interesting predicates. The principal of the alternative unary predicates is the Feferman Predicate that was introduced in [Fef60]. It was studied in [Mon78], [Vis89] and [Sha94]. Of a quite different kind is the binary predicate for interpretability over a given theory. This can be viewed as a generalization of ordinary provability. We refer the reader to e.g. [JdJ98], [Vis98], [JV00], [AB04], [GJ08].

An alternative arithmetical interpretation of interpretability logics is provided by various notions of conservativity. In the present paper we will provide yet another interpretation: the jump relation.

This paper is in the second tradition: we treat modal predicates as *objets trouvés*. They are already present in a given theory. On the other hand, we will not zoom in on specific predicates in the given theory, but we will be interested in the totality of predicates over the given theory satisfying such-and-such properties. The appropriate analogy is as follows. A predicate of a theory satisfying a given modal theory is like a model of a theory, for example *group theory*. We will be interested in the relationships between these 'models'. In the analogy: groups are models of group theory. One studies the relations between different groups and constructions on groups. In the same spirit we want to study certain transformations of predicates satisfying modal principles.

1.2. **Generalizing Feferman's Theorem and Giving It a Converse.** We have the following theorem:

Theorem 1.1 (Feferman)**.** *Consider any theory U with a p-time decidable axiom set. Suppose N is an interpretation of Buss' theory S^1_2 in U, then there is an interpretation K of $U + \mathsf{incon}^N(U)$ in U.*

In the statement of the theorem we assume that the theory U is given with a Δ^{b}_1-formula representing its axiom set. We note that Feferman's Theorem is a strengthening of the Second Incompleteness Theorem. The theory U not only fails to prove its own consistency as coded in any choice N for the natural numbers, no, it is positively able to produce uniformly internal models of itself in which we have its inconsistency coded in N. Feferman's Theorem enables us to view the Second Incompleteness Theorem as a strength rather than as a weakness. Feferman proved this theorem in [Fef60]. In [Fef97], Feferman provides a historical discussion of a.o. the genesis of his theorem which is warmly recommended. We give simple proof of Feferman's Theorem in Section 3.

In the paper we show that Feferman's result lifts to a class of predicates that we call HBL-predicates and also to an even wider class the regular HBL-predicates. In this generalized form, the theorem admits a converse. Let us restrict ourselves for definiteness to PA. We find: $\mathsf{PA} + B$ is interpretable in PA iff B is of the form $\triangle\bot$ for some HBL-predicate $\triangle$ over PA. So the inconsistency of PA is interpretbale over PA. Conversely a sentence is only interpretable over PA because it can be viewed as an inconsistency.

The generalization of Feferman's Theorem can be extended to a version involving regular HBL-predicates. This will allow us, e.g. for extensions U of PA, to find a new characterization of the relation $U+B$ *is interpretable in* $U+A$. This possibility leads to our third beginning.

1.3. **Semantics of Interpretability Logic.** The idea of interpretability logic is very simple, given that we already know provability logic. It is the modal study of the predicate A *interprets* B *over* U or $A \rhd_U B$, which means $U + A$ *interprets* $U + B$. We refer the reader to the papers [JdJ98], [Vis98], [JV00], [AB04], [GJ08]

for more information. The basic system of interpretability logic is IL which is Löb's Logic plus the following principles.[1]

IL1. $\vdash \Box(\phi \to \psi) \to \phi \rhd \psi$

IL2. $\vdash (\phi \rhd \psi \wedge \psi \rhd \chi) \to \phi \rhd \chi$

IL3. $\vdash (\phi \rhd \chi \wedge \psi \rhd \chi) \to (\phi \vee \psi) \rhd \chi$

IL4. $\vdash \phi \rhd \psi \to (\Diamond\phi \to \Diamond\psi)$

IL5. $\vdash \Diamond\phi \rhd \phi$

If we take an essentially reflexive theory like Peano Arithmetic (PA) as our basic theory, then we have the extra principle M.

M. $\vdash \phi \rhd \psi \to (\phi \wedge \Box\chi) \rhd (\psi \wedge \Box\chi)$

The arithmetical completeness of ILM for theories like PA was proven by Volodya Shavrukov in [Sha88] and Alessandro Berarducci in [Ber90]. Apart from relative interpretability the binary connective has a natural interpretation as conservativity. See e.g. [HM90], [Ign91], [HM92], [DJ94] Regular L-predicates provide a new interpretation of our modal interpretability logic.

1.4. **Basic Notions and Facts.** In Appendix A we introduce the basic notions and facts needed to read the paper. The reader is also referred to the textbook [HP93]. At this point we just fix a number of conventions and notations.

Theories are, in this paper, theories of first-order predicate logic, that have a finite relational signature and that are axiomatized by an axiom set that is represented by a Δ_1^{b}-formula. We will pretend that a theory also has function symbols. These terms can be eliminated by the well-known term-unwinding algorithm.

We use modal notations as much as possible. For example $\Box_U A$ is $\mathsf{prov}_U(\underline{\ulcorner A \urcorner})$. We use $\Box_{U,x} A$ for restricted provability where the Gödel numbers of the axioms used in the proof are all below x and the complexity (= depth of quantifier alternations) is amaller than x. See the appendix for more information. We use $A \rhd_U B$ for interpretability of U.

2. Jumping

In this section we define regular L-predicates and present their connection to interpretability logic.

2.1. **Regular L-predicates.** Let U be any theory and let N be an interpretation of the the Tarski-Mostowski-Robinson Arithmetic R in U. Let P be a predicate of the N-numbers. We write $\triangle A$ for $P(\underline{\ulcorner A \urcorner})$, where $\underline{\ulcorner \cdot \urcorner}$ is some standard efficient Gödelnumbering used w.r.t. N. We use $\vdash$ for $\vdash_U$.

A predicate $\triangle$ is *a regular L-predicate* w.r.t. U, N if it satisfies the following principles.

rL1. $A \vdash B \quad \Rightarrow \quad \triangle A \vdash \triangle B$

[1] Usually, we have two operators $\Box$ and $\rhd$. However, $\Box$ can be defined by: $\Box A :\leftrightarrow \neg A \rhd \bot$.

rL2. $\triangle A, \triangle(A \to B) \vdash \triangle B$

rL3. $\triangle A \vdash \triangle\triangle A$

The name *regular* is taken from the regular modal logics described in [HC96].

Remark 2.1. The original Hilbert-Bernays conditions ([HB39], in the second edition: p294, 295) were approximately, in modern notation:

1. $A \vdash B \quad \Rightarrow \quad \triangle A \vdash \triangle B$

2. $\vdash \triangle \neg\, Ax \to \triangle \neg\, A\underline{k}$

3. $S \vdash \triangle S$, where S is a Σ_1-sentence

So it is a curious fact is that rL1 was the true first Hilbert-Bernays condition. It should be noted that Hilbert and Bernays assumed that $\triangle$ was Σ_1, so their conditions were not really intended as fully abstract conditions. ❑

We note one alternative formulation for our axioms. Let Γ and Θ range over finite sets of U-sentences. We write $\triangle\Theta := \{\triangle C \mid C \in \Theta\}$. A predicate $\triangle$ is a regular L-predicate w.r.t. U, N iff it satisfies.

- $\Gamma, \triangle\Theta \vdash A \quad \Rightarrow \quad \triangle\Gamma, \triangle\Theta \vdash \triangle A$, where we demand that $\Gamma \cup \Theta \neq \emptyset$.

A predicate $\triangle$ is a *normal* L-predicate or simply an L-predicate if it is a regular L-predicate and if we have in addition that $\vdash \triangle\top$.

Example 2.2. Even over R, we have non-trivial examples of L-predicates. E.g., let A be the conjunction of the axioms of a finite axiomatization of S^1_2, or, rather, a version of S^1_2 in the arithmetical language. Then, $\triangle B :\leftrightarrow (A \to \Box_{\mathsf{R}} B)$ is an L-predicate. Here $\Box_{\mathsf{R}}$ is a standard formalization of provability. ❑

Since we have the necessary fixed points in R, we can prove Löb's Theorem.

Theorem 2.3 (Löb's Theorem)**.** *Suppose U is a theory with natural numbers N satisfying R. Suppose that $\triangle$ is a regular L-predicate. Then $\triangle(\triangle A \to A) \vdash \triangle A$.*

Proof. By the Gödel fixed Point Lemma, we find a B such that

$$(\dagger) \quad \vdash B \leftrightarrow (\triangle B \to A).$$

Since $B, \triangle B \vdash A$, we find $\triangle B \vdash \triangle A$. So, we have $\triangle A \to A \vdash \triangle B \to A$ and, hence, $\triangle A \to A \vdash B$. Ergo, $\triangle(\triangle A \to A) \vdash \triangle(\triangle B \to A)$ and $\triangle(\triangle A \to A) \vdash \triangle B$. By the usual reasoning it follows that $\triangle(\triangle A \to A) \vdash \triangle A$. ❑

We could also have given a purely modal formulation of the insight contained in Löb's theorem. The only role of R is the fact that it supports the Fixed Point Lemma. If we have a purely modal language and a modal operator satisfying the regular L-property extended with constants and axioms for fixed points for guarded (aka modalized) propositional variables, then the above reasoning works.

2.2. **Regular Löb's Logic.** Corresponding to the idea of a *regular L-predicate* we have the purely modal theory Regular Löb's Logic or rGL. It has the following axioms.

rGL1. $\phi \vdash \psi \quad \Rightarrow \quad \triangle\phi \vdash \triangle\psi$

rGL2. $\triangle\phi, \triangle(\phi \to \psi) \vdash \triangle\psi$

rGL3. $\triangle\phi \vdash \triangle\triangle\phi$

rGL4. $\triangle(\triangle\phi \to \phi) \vdash \triangle\phi$

The following result is rather useful.

Theorem 2.4. $\mathsf{GL} \vdash \phi$ *iff* $\mathsf{rGL} \vdash \triangle\top \to \phi$.

Proof. From left to right is an induction of proof length. From right to left is trivial, since GL extends rGL. ❑

2.3. **Jump Devices.** We define $A \blacktriangleright B$ as: for some regular L-predicate $\triangle$, we have $A \vdash \triangle\top$ and $\triangle\bot \vdash B$.

We will call a triple $\langle A, \triangle, B\rangle$ a *device* if $\triangle$ is a regular L-predicate and $A \vdash \triangle\top$ and $\triangle\bot \vdash B$. We write $\triangle : B \blacktriangleleft A$ or $\triangle : A \blacktriangleright B$ for: $\langle A, \triangle, B\rangle$ is a device.

Two devices $\triangle : A \blacktriangleright B$ and $\triangle' : A' \blacktriangleright B'$ are *equivalent* iff $\vdash A \leftrightarrow A'$, $\vdash B \leftrightarrow B'$ and, for all C, we have $\vdash \triangle C \leftrightarrow \triangle' C$.

Given any device from A to B, we can find a device from A to B that satisfies some further desirable properties. We will call such devices *basic devices*. A device $\triangle : A \blacktriangleright B$ is *basic* iff it satisfies:

bD1. $C \vdash D \quad \Rightarrow \quad \triangle C \vdash \triangle D$

bD2. $\triangle C, \triangle(C \to D) \vdash \triangle D$

bD3. $\vdash \triangle\top \leftrightarrow (A \vee B)$.

bD4. $\vdash \triangle\bot \leftrightarrow B$.

bD5. $\vdash \triangle\triangle\bot \leftrightarrow (A \vee B)$.

It is easy to see that a basic device is a device. Suppose $\triangle : A \blacktriangleright B$ is an arbitrary device. Consider $\Psi\langle A, \triangle, B\rangle := \langle A, \triangle^*, B\rangle$, where $\triangle^*$ is defined as

$$\triangle^* C :\leftrightarrow B \vee (A \wedge \triangle(B \to C)).$$

We claim that $\triangle^* : A \blacktriangleright B$ is a basic device. The verification is an easy exercise in modal logic. We treat as an example bD5. We have:

$$\begin{array}{rcl} B & \vdash & B \vee (A \wedge \triangle(B \to \bot)) \\ & \vdash & B \vee (A \wedge \triangle(\triangle\bot \to \bot)) \\ & \vdash & B \vee (A \wedge \triangle\bot) \\ & \vdash & B \vee (A \wedge B) \\ & \vdash & B \end{array}$$

We note that Ψ preserves equivalence of devices. The operation Ψ is idempotent, since: $\Psi\langle A, \Psi\langle A, \triangle, B\rangle, B\rangle = \langle A, \triangle^\circ, B\rangle$, where:

$$\triangle^\circ C := B \vee (A \wedge (B \vee (A \wedge \triangle(B \to C))))$$

We verify the validity of IL for $\blacktriangleright$. The verification is for the moment in the meta-language. We want it to be verifiable in the U itself. We postpone discussion of the demands on U until after the proof. We remind the reader that the logic IL is defined as follows.

$\mathsf{IL}1$. $\vdash \Box(\phi \to \psi) \to \phi \rhd \psi$

$\mathsf{IL}2$. $\vdash (\phi \rhd \psi \wedge \psi \rhd \chi) \to \phi \rhd \chi$

$\mathsf{IL}3$. $\vdash (\phi \rhd \chi \wedge \psi \rhd \chi) \to (\phi \vee \psi) \rhd \chi$

$\mathsf{IL}4$. $\vdash \phi \rhd \psi \to (\Diamond\phi \to \Diamond\psi)$

$\mathsf{IL}5$. $\vdash \Diamond\phi \rhd \phi$

Each principle except $\mathsf{IL}4$ corresponds to an operation on devices.[2] The operations as chosen by us all yield a basic device as output independent of whether the input devices are basic. In our verifications we will use the fact that $\Gamma, \triangle\Theta \vdash A \Rightarrow \triangle\Gamma, \triangle\Theta \vdash \triangle A$, provided that $\Gamma \cup \Theta$ is non-empty.

$\mathsf{IL}1$: Suppose $\vdash A \to B$. We define $\Phi_1(A, B) := \langle A, \triangle^*, B\rangle$, where $\triangle^* C :\leftrightarrow B$.

It is easy to verify that $\Phi_1(A, B)$ is a basic device.

$\mathsf{IL}2$: Suppose $\triangle_0 : A \blacktriangleright B$ and $\triangle_1 : B \blacktriangleright C$. We define $\Phi_2(A, \triangle_0, B, \triangle_1, C) := \langle A, \triangle^*, C\rangle$, where:

$$\begin{aligned}\triangle^* D \quad :\leftrightarrow \quad & C \vee (A \wedge ((B \wedge \triangle_1(C \to D)) \ \vee \\ & (\neg B \wedge \triangle_0(B \to ((C \wedge D) \vee (\neg C \wedge \triangle_1(C \to D))))))\end{aligned}$$

We verify $\mathsf{bD}1$. Suppose $D \vdash E$. It follows that $(C \to D) \vdash (C \to E)$. Ergo:

$$\text{(a)} \quad \triangle_1(C \to D) \vdash \triangle_1(C \to E).$$

Hence, also:

$$B \to ((C \wedge D) \vee (\neg C \wedge \triangle_1(C \to D))) \vdash B \to ((C \wedge E) \vee (\neg C \wedge \triangle_1(C \to E))).$$

It follows that:

$$\text{(b)} \quad \triangle_0(B \to ((C \wedge D) \vee (\neg C \wedge \triangle_1(C \to D)))) \vdash \triangle_0(B \to ((C \wedge E) \vee (\neg C \wedge \triangle_1(C \to E)))).$$

It is immediate from (a) and (b) that $\triangle^* D \vdash \triangle^* E$.

We verify $\mathsf{bD}2$. We can easily derive: (a) $\triangle_1(C \to D), \triangle_1(C \to (D \to E)) \vdash \triangle_1(E)$ and from this:

$$\begin{aligned}B \to ((C \wedge D) \vee (\neg C \wedge \triangle_1(C \to D))), & \\ B \to ((C \wedge (D \to E)) \vee (\neg C \wedge \triangle_1(C \to (D \to E)))) & \vdash \\ B \to ((C \wedge E) \vee (\neg C \wedge \triangle_1(C \to E))).&\end{aligned}$$

[2] In hindsight is would have been more natural to have $\mathsf{IL}4$ as the last principle of the list. However, we do not want to diverge from the traditional order.

It follows that:

$$\begin{aligned}(b)\ \triangle_0(B \to ((C \wedge D) \vee (\neg C \wedge \triangle_1(C \to D)))),\\ \triangle_0(B \to ((C \wedge (D \to E)) \vee (\neg C \wedge \triangle_1(C \to (D \to E))))) \vdash\\ \triangle_0(B \to ((C \wedge E) \vee (\neg C \wedge \triangle_1(C \to E)))).\end{aligned}$$

We reason in U. Suppose (c) $\triangle^* D$ and (d) $\triangle^*(D \to E)$. It follows that we have one of the exclusive cases C or $\neg C \wedge A \wedge B$ or $\neg C \wedge A \wedge \neg B$. In case we have C we are immediately done. Suppose we have $\neg C$ and A and B. In this case (c) gives us $\triangle_1(C \to D)$ and (d) gives us $\triangle_1(C \to (D \to E))$. By (a), we find $\triangle_1(C \to E)$. Ergo $\triangle^* E$. Suppose we have $\neg C$ and A and $\neg B$. In this case (c) and (d) give us $\triangle_0(B \to ((C \wedge D) \vee (\neg C \wedge \triangle_1(C \to D))))$ and $\triangle_0(B \to ((C \wedge (D \to E)) \vee (\neg C \wedge \triangle_1(C \to (D \to E)))))$. By (b) we find the desired conclusion $\triangle_0(B \to ((C \wedge E) \vee (\neg C \wedge \triangle_1(C \to E))))$, and hence $\triangle^* E$.

We treat $\mathsf{bD3,4,5}$. First we have:

$$\begin{aligned}\vdash \triangle^*\top \quad &\leftrightarrow \quad C \vee (A \wedge ((B \wedge \triangle_1(C \to \top)) \vee\\ &\qquad (\neg B \wedge \triangle_0(B \to ((C \wedge \top) \vee (\neg C \wedge \triangle_1(C \to \top)))))))\\ &\leftrightarrow \quad C \vee (A \wedge ((B \wedge \top) \vee\\ &\qquad (\neg B \wedge \triangle_0(B \to (C \vee (\neg C \wedge \top))))))\\ &\leftrightarrow \quad C \vee (A \wedge (B \vee (\neg B \wedge \top)))\\ &\leftrightarrow \quad C \vee A\end{aligned}$$

We note that since $\triangle_0\bot \vdash B$, we have $\neg B \vdash \neg\, \triangle_0\bot$, and, hence, $\triangle_0 \neg B \vdash \triangle_0(\neg\, \triangle_0\bot)$ and, so, $\triangle_0 \neg B \vdash \triangle_0\bot$. Similarly, $\triangle_1 \neg C \vdash \triangle_1\bot$. We have:

$$\begin{aligned}\vdash \triangle^*\bot \quad &\leftrightarrow \quad C \vee (A \wedge ((B \wedge \triangle_1 \neg C) \vee\\ &\qquad (\neg B \wedge \triangle_0(B \to ((C \wedge \bot) \vee (\neg C \wedge \triangle_1 \neg C))))))\\ &\leftrightarrow \quad C \vee (A \wedge ((B \wedge \triangle_1\bot) \vee (\neg B \wedge \triangle_0(B \to (\neg C \wedge \triangle_1\bot)))))\\ &\leftrightarrow \quad C \vee (A \wedge ((B \wedge C) \vee (\neg B \wedge \triangle_0 \neg B)))\\ &\leftrightarrow \quad C \vee (A \wedge ((B \wedge C) \vee (\neg B \wedge \triangle_0\bot)))\\ &\leftrightarrow \quad C \vee (A \wedge B \wedge C)\\ &\leftrightarrow \quad C\end{aligned}$$

Finally:

$$\begin{aligned}\vdash \triangle^*\triangle^*\bot \quad &\leftrightarrow \quad \triangle^* C\\ &\leftrightarrow \quad C \vee (A \wedge ((B \wedge \triangle_1(C \to C)) \vee\\ &\qquad (\neg B \wedge \triangle_0(B \to ((C \wedge C) \vee (\neg C \wedge \triangle_1(C \to C)))))))\\ &\leftrightarrow \quad C \vee (A \wedge ((B \wedge \top) \vee (\neg B \wedge \triangle_0(B \to (\top \vee (\neg C \wedge \triangle_1\top))))))\\ &\leftrightarrow \quad C \vee (A \wedge (B \vee (\neg B \wedge \top)))\\ &\leftrightarrow \quad C \vee A\end{aligned}$$

$\mathsf{IL3}$: Suppose $\triangle_0 : A \blacktriangleright C$ and $\triangle_1 : B \blacktriangleright C$, We define $\Phi_3(A, \triangle_0, B, \triangle_1, C) := \langle A \vee B, \triangle^*, C\rangle$, where:

$$\triangle^* D :\leftrightarrow C \vee (A \wedge \triangle_0(C \to D)) \vee (\neg\, A \wedge B \wedge \triangle_1(C \to D)).$$

All cases except bD4 are like the cases of IL2 but simpler. We treat bD4.

$$\begin{array}{rcl} \vdash C & \to & C \vee (A \wedge \triangle_0 \neg C) \vee (\neg A \wedge B \wedge \triangle_1 \neg C) \\ & \to & C \vee (A \wedge C) \vee (\neg A \wedge B \wedge C) \\ & \to & C \end{array}$$

IL4: Suppose $\triangle : A \blacktriangleright B$ and $\vdash \neg B$. It follows that $\triangle\top \vdash \triangle\neg B$, and, hence, $\triangle\top \vdash \triangle\neg\triangle\bot$ and, thus, that $\triangle\top \vdash \triangle\bot$. We may conclude that $A \vdash B$ and, so, $A \vdash \bot$, i.e., $\vdash \neg A$.

IL5: We prove a stronger fact, say IL5$^+$. Suppose $\triangle$ is a regular L-predicate and $A \vdash \triangle\top$. We show that $(A \wedge \triangledown B) \rhd B$. We define $\Phi_4(A, B, \triangle) := \langle (A \wedge \triangledown B), \triangle^*, B\rangle$, where :

$$\triangle^* C :\leftrightarrow B \vee (A \wedge \triangledown B \wedge \triangle(B \to C)).$$

We leave the simple verifications to the reader.

We will call Φ_1, Φ_2, Φ_3, Φ_4: *the* Φ*-operations*. We will call a class of devices closed under the Φ-operations: Φ*-closed*.

To truly obtain IL, we need verifiability of the above proofs in U itself w.r.t. some chosen $N : \mathsf{S}^1_2 \lhd U$. Fortunately all the transformations in our verification are p-time, so we do not encounter a problem in internalizing the argument.

Clearly every Φ-closed class of devices $\mathcal{D}$ will satisfy IL. The basic devices are an example of such a class. We will write $\blacktriangleright_{\mathcal{D}}$ for the jump relation obtained by only considering devices from $\mathcal{D}$.

Open Question 2.5. One would hope that the devices (or a closed subclass of the devices) form a category, but its seems that our definitions do not yield the associativity of composition. Since neither the class of devices nor the chosen operations on devices are uniquely determined, there is still some hope that we can find the desired category. So we formulate the open question: *can we find a category of devices?* ❑

2.4. **Regular HBL-predicates.** In this subsection we introduce the class of regular HBL-predicates. We will show that devices associated with these predicates are Φ-closed.

We formulate our relevant notion of $\exists\Sigma^{\mathsf{b}}_1$-completeness. Consider a theory U and an interpretation $N : \mathsf{S}^1_2 \lhd U$. We define:

- r-C: $\triangle\top, S^N \vdash \triangle S^M$, where S is a $\exists\Sigma^{\mathsf{b}}_1$-sentence and M is any interpretation of S^1_2 in U.
- r-C$_0$: $\triangle\top, S^N \vdash \triangle S^N$, where S is a $\exists\Sigma^{\mathsf{b}}_1$-sentence.

Note that the definition assumes that we have U and N fixed in the background. We call a regular L-predicate that satisfies r-C w.r.t. U, N: *a regular HBL-predicate*. We call a regular L-predicate that satisfies r-C$_0$ w.r.t. U, N: *a regular HBL$_0$-predicate*. The name "HBL" stands for: Hilbert-Bernays-Löb. The reason for this choice is the fact that the third Hilbert-Bernays was verifiable Σ_1-completeness.

Here is a basic theorem about regular $\exists\Sigma^{\mathsf{b}}_1$-completeness, connecting it with restricted provability.

Theorem 2.6. *Suppose that U is sequential. Let N interpret S^1_2 in U. Suppose that $\triangle$ is a regular L-predicate for U, N. Then, the following are equivalent:*

i. $\triangle$ is a regular HBL-predicate.

ii. For all U-sentences A, and for all n, we have $\triangle\top \vdash \Box^N_n A \to \triangle A$.

Proof. Suppose that U is sequential and $N : \mathsf{S}^1_2 \lhd U$. Suppose that $\triangle$ is a regular L-predicate for U, N.

(i) $\Rightarrow$ (ii). Suppose $\triangle$ is a regular HBL-predicate. Consider any sentence A and any number n. Since U is sequential, there is an interpretation $M : \mathsf{S}^1_2 \lhd U$, such that $U \vdash \Box^M_n A \to A$. By $\mathsf{r\text{-}C}$, we have $\triangle\top \vdash \Box^N_n A \to \triangle\Box^M_n A$. Since $\vdash \Box^M_n A \to A$, we have, by $\mathsf{rL1}$, that $\vdash \triangle\Box^M_n A \to \triangle A$. It follows that $\triangle\top \vdash \Box^N_n A \to \triangle A$.

(ii) $\Rightarrow$ (i), Suppose that, for all U-sentences A, and for all n, we have $\triangle\top \vdash \Box^N_n A \to \triangle A$. Consider any $\exists\Sigma^{\mathsf{b}}_1$-sentence S and any $M : \mathsf{S}^1_2 \lhd U$. We have, for sufficiently large n, $\vdash S^N \to \Box^N_n S^M$ and $\triangle\top \vdash \Box^N_n S^M \to \triangle S^M$. Hence, $\triangle\top \vdash S^N \to \triangle S^M$. ❑

We remind the reader of our operations:

- $\Psi\langle A, \triangle, B\rangle := \langle A, \triangle^*, B\rangle$, where $\triangle^*$ is defined as:
  $\triangle^* C :\leftrightarrow B \vee (A \wedge \triangle(B \to C))$.
- $\Phi_1(A, B) := \langle A, \triangle^*, B\rangle$, where $\triangle^* C :\leftrightarrow B$.
- Suppose $\triangle_0 : A \blacktriangleright B$ and $\triangle_1 : B \blacktriangleright C$. We define $\Phi_2(A, \triangle_0, B, \triangle_1, C) := \langle A, \triangle^*, C\rangle$, where:
  $\triangle^* D :\leftrightarrow C \vee (A \wedge ((B \wedge \triangle_1(C \to D)) \vee$
  $(\neg B \wedge \triangle_0(B \to ((C \wedge D) \vee (\neg C \wedge \triangle_1(C \to D))))))$.
- Suppose $\triangle_0 : A \blacktriangleright C$ and $\triangle_1 : B \blacktriangleright C$, We define $\Phi_3(A, \triangle_0, B, \triangle_1, C) := \langle A \vee B, \triangle^*, C\rangle$, where:
  $\triangle^* D :\leftrightarrow C \vee (A \wedge \triangle_0(C \to D)) \vee (\neg A \wedge B \wedge \triangle_1(C \to D))$.
- Suppose $\triangle : A \blacktriangleright D$. Note that D is not necessarily B. We define $\Phi_4(A, B, \triangle) := \langle (A \wedge \nabla B), \triangle^*, B\rangle$, where:
  $\triangle^* C :\leftrightarrow B \vee (A \wedge \nabla B \wedge \triangle(B \to C))$.

If the predicates in the input of the operations are HBL (HBL_0) for U, N, then so are the predicates in the output.

We treat the case of Φ_2 for HBL. Suppose $\triangle_0 : A \blacktriangleright B$ and $\triangle_1 : B \blacktriangleright C$, where $\triangle_0$ and $\triangle_1$ are HBL. We have, for any $M : \mathsf{S}^1_2 \lhd U$, that: $A \vdash S^N \to \triangle_0 S^M$, so *a fortiori* $A \vdash S^N \to \triangle_0(B \to S^M)$. Similarly, $B \vdash S^N \to \triangle_1(C \to S^M)$ and, hence, $A \vdash \triangle_0(B \to (S^N \to \triangle_1(C \to S^M)))$. We also have $A \vdash S^N \to (\triangle_0 S^N \wedge \triangle_0 S^M)$. So:

$$A \vdash S^N \to \triangle_0(B \to ((C \wedge S^M) \vee (\neg C \wedge \triangle_1(C \to S^M)))).$$

From these facts the desired result is immediate.

We show that, if we restrict ourselves to devices based on HBL (HBL_0) predicates for U, N, we have:

M: Suppose S is $\exists\Sigma^{\mathsf{b}}_1$, then: $A \blacktriangleright B \Rightarrow (A \wedge S^N) \blacktriangleright (B \wedge S^N)$.

Suppose $\triangle : A \blacktriangleright B$. Let $\Phi_5(A, \triangle, B) := \langle A \wedge S^N, \triangle^*, B \wedge S^N \rangle$, where $\triangle^* C :\leftrightarrow (S^N \wedge \triangle C)$.

We leave the easy verification that $\triangle^*$ is indeed a HBL (HBL_0) predicate for U, N and that $\triangle^* : (A \wedge S^N) \blacktriangleright (B \wedge S^N)$ to the reader. We note that, since $\Box_U C$ is $\exists \Sigma^{\mathsf{b}}_1$, the usual form of M follows:

$$A \blacktriangleright B \quad \Rightarrow \quad (A \wedge \Box^N C) \blacktriangleright (B \wedge \Box^N C).$$

3. Feferman's Theorem

In this section we present a simple proof of Feferman's Theorem. We remind the reader of the Theorem.

Feferman's Theorem: *Consider any theory U with a p-time decidable axiom set. Suppose N is an interpretation of Buss' theory S^1_2 in U, then there is an interpretation K of $U + \mathsf{incon}^N(U)$ in U.*

Proof. Consider any theory U with p-time decidable axiom set and an interpretation $N : \mathsf{S}^1_2 \lhd U$. Clearly, we have $\Diamond^N \top \vdash_U \Diamond^N \Box^N \bot$ and $\Diamond^N \Box^N \bot \rhd_U \Box^N \bot$, by, respectively, the Second Incompleteness Theorem and the Gödel-Hilbert-Bernays-Wang-Henkin-Feferman Theorem (Theorem A.1 of the Appendix). By composition, $\Diamond^N \top \rhd_U \Box^N \bot$. Suppose K witnesses that $\Diamond^N \top \rhd_U \Box^N \bot$. We also have $\mathsf{ID} : \Box^N \bot \rhd_U \Box^N \bot$. Hence $K\langle \Diamond^N \top \rangle \mathsf{ID} : \top \rhd_U \Box^N \bot$. Here $K\langle \Diamond^N \top \rangle \mathsf{ID}$ is the disjunctive interpretation that 'is' K if $\Diamond^N \top$ and ID if not $\Diamond^N \top$. (See Appendix A.2 for the definition of disjunctive interpretations.) ❑

The proof of Feferman's Theorem presented here was given in [Vis90]. The same proof is reported in [Fef97]. Feferman learned the proof in conversation from Per Lindström. It seems likely that Per discovered the proof independently.

4. Jumping, Interpretatibility, Conservativity

In this section, we prove that, for essentially reflexive theories, jumping and interpretability coincide and we prove that for theories interpreting EA jumping and Π_1-conservativity coincide. (The preceding formulation is still not fully precise. It will be refined below.)

4.1. Interpretability over Essentially Reflexive Theories.

In this subsection, we show that HBL-jumping and interpretability coincide for essentially reflexive theories. We first prove Interpretation Existence for HBL predicates.

A theory U is *locally sententially essentially reflexive* if, for every U-sentence A and for every n, there is an $M : \mathsf{S}^1_2 \lhd U$ such that $U \vdash \Box^M_{U,n} A \to A$. Here $\Box_{U,n}$ is restricted provability as described in Appendix A.4. As is well known, sequential theories are locally sententially essentially reflexive.

The theory U is *sententially essentially reflexive* if, there is a fixed $N : \mathsf{S}^1_2 \lhd U$ such that, for every U-sentence A and for every n, we have $U \vdash \Box^N_{U,n} A \to A$. We often make N part of the data and say, e.g., *U is essentially reflexive w.r.t. N.* Sequential theories with full induction, such as PA and ZF, are essentially reflexive

and, hence, *a fortiori* sententially essentially reflexive. For a worked out-example of a theory that is sententially essentially reflexive but not essentially reflexive, see [Vis12].

Theorem 4.1. *Suppose that U is locally sententially essentially reflexive. Let the interpretation N provide natural numbers satisfying S^1_2. Suppose that $\triangle$ is a regular HBL-predicate for U, N. Then, $(\triangle\top \wedge \triangledown A) \rhd A$.*

Proof. By Theorem 2.6, we have, for every n, that $\triangle\top \vdash \triangledown A \to \Diamond^N_{U,n} A$. Hence, by Theorem A.3, we find that $(\triangle\top \wedge \triangledown A) \rhd A$. ❑

Next we prove Feferman's Theorem w.r.t. HBL predicates.

Theorem 4.2. *Suppose that U is locally sententially essentially reflexive. Let our natural numbers be given by $N : \mathsf{S}^1_2 \lhd U$. Suppose that $\triangle$ is a regular HBL-predicate for U, N. Then, $\triangle\top \rhd \triangle\bot$.*

Proof. First, we trivially have $(\triangle\top \wedge \triangle\bot) \rhd \triangle\bot$. Secondly, we have, by Löb's Theorem, $(\triangle\top \wedge \triangledown\top) \vdash (\triangle\top \wedge \triangledown\triangle\bot)$ and, by Theorem 4.1, $(\triangle\top \wedge \triangledown\triangle\bot) \rhd \triangle\bot$. By $\mathsf{IL3}$ we are done. ❑

We now move to a result where we really need global reflexivity.

Theorem 4.3. *We work over a theory U and $N : \mathsf{S}^1_2 \lhd U$. Suppose U is sententially essentially reflexive w.r.t. N. Then, over U, N, we have: $A \rhd B$ iff $A \blacktriangleright_{\mathsf{hbl}} B$.*

Proof. Suppose $A \rhd B$. It follows that, for every n, $A \vdash \Diamond^N_n B$. We define:

$$\triangle C :\leftrightarrow B \vee (A \wedge \exists x\, (\Box^N_x (B \to C) \wedge \Diamond^N_x B)).$$

We show that $\triangle : A \blacktriangleright B$ is a basic device. We have:

$$\begin{array}{rrcl} \vdash & \triangle\bot & \leftrightarrow & B \vee (A \wedge \exists x\, (\Box^N_x \neg B \wedge \Diamond^N_x B)) \\ & & \leftrightarrow & B \\ \vdash & \triangle\top & \leftrightarrow & B \vee (A \wedge \exists x\, (\Box^N_x \top \wedge \Diamond^N_x B)) \\ & & \leftrightarrow & A \vee B \\ \vdash & \triangle\triangle\bot & \leftrightarrow & \triangle B \\ & & \leftrightarrow & B \vee (A \wedge \exists x\, (\Box^N_x \top \wedge \Diamond^N_x B)) \\ & & \leftrightarrow & A \vee B \end{array}$$

We treat $\mathsf{rL1}$. Suppose $C \vdash D$. Then, (a) for some n, $C \vdash_n D$. We reason in U. Suppose (b) $B \vee (A \wedge \exists x\, (\Box^N_x (B \to C) \wedge \Diamond^N_x B))$ We want to prove (c) $B \vee (A \wedge \exists x\, (\Box^N_x (B \to D) \wedge \Diamond^N_x B))$. If B we are easily done. Suppose (d) $A \wedge \exists x\, (\Box^N_x (B \to C) \wedge \Diamond^N_x B)$. It follows that we have A and hence $\Diamond^N_n B$. Thus, we may assume that for some $a \geq n$, (e) $\Box^N_a (B \to C) \wedge \Diamond^N_a B$. Combining this with (a), we find: (f) $\Box^N_a (B \to D) \wedge \Diamond_a B$. From this we easily find the desired (c).

Both $\mathsf{rL2}$ and the $\exists\Sigma^{\mathsf{b}}_1$-condition are both easy.

The other direction is immediate from Theorem 4.2. ❑

We remind the reader that every essentially reflexive theory U has Orey sentences. This means that, there is a sentence O such that $\top \rhd_U O$ and $\top \rhd_U \neg O$. It follows from Theorem 4.3, that there are HBL-predicates $\vartriangle_0$ and $\vartriangle_1$ such that $U \vdash \vartriangle_0 \bot \leftrightarrow \neg \vartriangle_1 \bot$.

Both Per Lindström and Robert Solovay have shown that interpretability over an essentially reflexive theory is complete Π_1. Inspecting the proof of Theorem 4.3 we can see that we can reduce the question whether $A \rhd_U B$ to the question whether the specific predicate $\vartriangle$ as constructed in the proof is a HBL-predicate. Hence the question whether a predicate is HBL is complete Π_2.

4.2. **Π_1-Conservativity.** Suppose Γ is a set of arithmetical sentences. We define Γ-conservativity. Let $N : \mathsf{S}^1_2 \lhd U$ and $M : \mathsf{S}^1_2 \lhd U$. Then:

- $(U, N) \rhd_\Gamma (V, M)$ iff, for all Γ-sentences C, if $V \vdash C^M$, then $U \vdash C^N$.
- $A \rhd_{U,N,\Gamma} B$ iff $(U + A, N) \rhd_\Gamma (U + B, N)$.

The logic of Π_1-conservativety was studied by Petr Hájek and Franco Montagna in two papers [HM90], [HM92]. They proved the arithmetical completeness of ILM for extensions of $\mathrm{I}\Sigma_1$. A careful analysis of precisely what principles are involved in the proof can be found in [BV05]. The basic system for which the proof works is $\mathrm{III}^- + \mathsf{Exp}$. In this section we prove that $\blacktriangleright_{\mathsf{hbl}_0}$ coincides with $\rhd_{\Pi_1}$ for extensions of EA, aka $\mathrm{I}\Delta_0 + \mathsf{Exp}$.

Theorem 4.4. *Consider any theory U and any $N : \mathsf{S}^1_2 \lhd U$. Suppose $A \blacktriangleright_{\mathsf{hbl}_0} B$ w.r.t. U, N. Then $A \rhd_{\forall \Pi^{\mathsf{b}}_1} B$.*

Proof. Let U and any $N : \mathsf{S}^1_2 \lhd U$ be given. Suppose $A \blacktriangleright_{\mathsf{hbl}_0} B$ w.r.t. U, N. Let $\vartriangle$ be a HBL_0 predicate for U, N and let P be a $\forall\Pi^{\mathsf{b}}_1$-sentence and let S be the negation of P. We have:

$$\begin{aligned} B \vdash P^N &\Rightarrow \vartriangle \bot \vdash P^N \\ &\Rightarrow S^N \vdash \neg \vartriangle \bot \\ &\Rightarrow \vartriangle S^N \vdash \vartriangle \neg \vartriangle \bot \\ &\Rightarrow \vartriangle S^N \vdash \vartriangle \bot \\ &\Rightarrow S^N, \vartriangle \top \vdash \vartriangle \bot \\ &\Rightarrow S^N, \vartriangle \top \vdash \bot \\ &\Rightarrow \vartriangle \top \vdash P^N \\ &\Rightarrow A \vdash P^N \end{aligned}$$

Hence B is $\forall\Pi^{\mathsf{b}}_1$-conservative over A w.r.t. U, N. ❑

If we have the totality of exponentiation in N, then we can transform $\forall\Pi^{\mathsf{b}}_1$-conservativity into Π_1-conservativity. For completeness sake we reproduce the simple argument.

Theorem 4.5. *Consider any theory U and any $N : \mathsf{EA} \lhd U$. Suppose $A \blacktriangleright_{\mathsf{hbl}_0} B$ w.r.t. U, N. Then $A \rhd_{\Pi_1} B$.*

Proof. Let U and any $N : \mathsf{EA} \lhd U$ be given. Suppose $A \blacktriangleright_{\mathsf{hbl}_0} B$ w.r.t. U, N. Let $\triangle$ be a HBL_0 predicate for U, N and let P be a Π_1-sentence and let S be the negation of P. We have $\mathsf{EA} \vdash S \leftrightarrow S_0$, for some S_0 in $\exists\Sigma^{\mathsf{b}}_1$. We have:

$$\begin{aligned} B \vdash P^N &\Rightarrow \triangle\bot \vdash P^N \\ &\Rightarrow S^N \vdash \neg\triangle\bot \\ &\Rightarrow S_0^N \vdash \neg\triangle\bot \\ &\Rightarrow \triangle S_0^N \vdash \triangle\neg\triangle\bot \\ &\Rightarrow \triangle S_0^N \vdash \triangle\bot \\ &\Rightarrow S_0^N, \triangle\top \vdash \triangle\bot \\ &\Rightarrow S^N, \triangle\top \vdash \bot \\ &\Rightarrow \triangle\top \vdash P^N \\ &\Rightarrow A \vdash P^N \end{aligned}$$

Hence B is Π_1-conservative over A w.r.t. U, N. ❑

It would be nice to prove a converse of Theorem 4.4. However, we could not do it. In stead we prove a converse of Theorem 4.5. To prove this converse we need to develop some machinery. Our strategy is to develop an analogue of restricted provability and then simply mimic the proofs we gave for the case of interpretability and jumping.

Open Question 4.6. Do we have the converse of Theorem 4.4? I.o.w., consider any theory U and any $N : \mathsf{S}^1_2 \lhd U$. Suppose $A \rhd_{\forall\Pi^{\mathsf{b}}_1} B$ w.r.t. U, N. Do we have: $A \blacktriangleright_{\mathsf{hbl}_0} B$ w.r.t. U, N? ❑

Let ϕ be a formula of propositional logic. We define: $\mathsf{subst}_U(\phi)$ is the set of all $\sigma : \mathsf{FV}(\phi) \to \mathsf{sent}_U$. We write $\mathsf{taut}(\phi)$ for 'ϕ is a tautology' and $\Box_{\mathsf{prop}}$ for provability in propositional logic.

Lemma 4.7. *Suppose U is any theory and $N : \mathsf{S}^1_2 \lhd U$. We have:*

i. $\mathsf{EA} \vdash \forall\phi\,(\mathsf{taut}(\phi) \to \Box_{\mathsf{prop}}\phi)$,

ii. $\mathsf{EA} \vdash \forall\phi\,\forall\sigma \in \mathsf{subst}_U(\phi)\,(\mathsf{taut}(\phi) \to \Box_U\sigma(\phi))$,

iii. $\mathsf{EA} \vdash \forall\phi\,(\neg\,\mathsf{taut}(\phi) \to \Box_U\neg\,\mathsf{taut}^N(\phi))$,

iv. $\mathsf{EA} \vdash \forall\phi\,\forall\sigma \in \mathsf{subst}_U(\phi)\,\Box_U(\mathsf{taut}^N(\phi) \to \sigma(\phi))$.

Proof. The proof of (i) is simply the formalization of the usual completeness proof of propositional logic. Item (ii) is a direct consequence of (i). Item (iii) is an instance of Σ_1-completeness. Item (iv) follows from (ii) and (iii). ❑

We define sub_0 is follows:

- $\mathsf{sub}_0(A) := \{A\}$ if A is of the form $\mathsf{Q}x\,B$, for $\mathsf{Q} \in \{\forall, \exists\}$, or $P\vec{t}$, where $P\vec{t}$ is an atomic sentence.
- $\mathsf{sub}_0(B \wedge C) := \mathsf{sub}_0(B) \cup \mathsf{sub}_0(C) \cup \{(B \wedge C)\}$, and similarly for the other propositional connectives.

The set $\mathsf{at}_0(A)$ is the set of all B in $\mathsf{sub}_0(A)$ of the form $\mathsf{Q}x\,C$ or $P\vec{t}$, where P is atomic. We define the function θ by $\theta(A) := p_{\ulcorner A \urcorner}$, if A is of the form $\mathsf{Q}x\,B$ or $P\vec{t}$, where P is atomic, and θ commutes with the propositional connectives. Suppose $\nu : p_{\ulcorner A \urcorner} \mapsto A$. Then $\nu(\theta(B)) = B$. Let $\mathsf{taut}^*_U(A) := \mathsf{taut}(\theta(A))$. We have:

Lemma 4.8. *Suppose U is any theory and $N : \mathsf{S}^1_2 \lhd U$. We have:*
$\mathsf{EA} \vdash \forall A\, \Box_U(\,\mathsf{taut}^{*N}_U(A) \to A)$.

Proof. The lemma is immediate by Lemma 4.7(iv). ❑

We will employ a Σ_1-truth predicate in the definition of our restricted-provability-analogue. We note that we have $\mathsf{S}^1_2 \vdash \mathsf{true}_{\Sigma_1}(S) \to S$ and $\mathsf{EA} \vdash S \to \mathsf{true}_{\Sigma_1}(S)$. Suppose $\mathsf{true}_{\Sigma_1}(x)$ is of the form $\exists y\, \mathsf{true}_0(y, x)$, where true_0 is Δ_0. We write $\mathsf{true}^z_{\Sigma_1}(x)$ for: $\exists y \leq z\, \mathsf{true}_0(y, x)$.

Let $\mathsf{S}^*(A)$ be the set of S in Σ_1 such that S^N is in $\mathsf{at}_0(A)$. Here we assume that all such formulas start with an existential quantifier. Let $\mathsf{S}^*(X) := \bigcup_{A\in X} \mathsf{S}^*(A)$. We write X^N for the set of B^N such that B is in X. We write $\Box$ for $\Box_U$ and taut^* for taut^*_U. Let $Y_x := \{B \mid \exists p < x\, \mathsf{proof}_U(p, B)\}$. We define:

$$\blacksquare_x A := \exists\, \mathcal{S} \subseteq \mathsf{S}^*(Y_x \cup \{A\})\, \exists z\, (\forall S \in \mathcal{S}\ \mathsf{true}^z_{\Sigma_1}(S) \wedge \mathsf{taut}^*(\bigwedge(Y_x \cup \mathcal{S}^N) \to A)).$$

The business with variable 'z' is just a trick to avoid the use of Σ_1-collection. In case we do have Σ_1-collection in the ambient theory we can omit 'z' from the definition. We collect the basic facts about $\blacksquare_x$ in a lemma.

Lemma 4.9. *Suppose U is any theory and $N : \mathsf{S}^1_2 \lhd U$. The variable 'S' ranges over Σ_1-sentences, that begin with an existential quantifier. We have:*

i. $\blacksquare_x A$ *is* Σ_1.

ii. $\mathsf{EA} \vdash \forall A\, (\Box A \to \exists x\, \blacksquare_x A)$.

iii. $\mathsf{EA} \vdash \forall A\, (\Box A \to \exists x\, \Box\, \blacksquare^N_x A)$.

iv. $\mathsf{EA} \vdash \forall S, x\, (\mathsf{true}_{\Sigma_1}(S) \to \blacksquare_x S)$.

v. $\mathsf{EA} \vdash \forall x, A, B\, ((\blacksquare_x A \wedge \blacksquare_x(A \to B)) \to \blacksquare_x B)$.

vi. $\mathsf{EA} \vdash \forall x, A\, \Box(\blacksquare^N_x A \to A)$.

vii. $\mathsf{EA} \vdash \forall A\, (\exists x\, \blacksquare_x A \to \Box A)$.

viii. $\mathsf{EA} \vdash \forall A\, (\exists x\, \blacksquare_x A \leftrightarrow \Box A)$.

Proof. Items (i) and (ii) are trivial. Item (iii) follows from (i) and (ii) by Σ_1-completeness. (iv) is again trivial.

We address item (v). Reason in EA. Consider A, B and x. Suppose $\blacksquare_x A$ and $\blacksquare_x(A \to B)$. We have $\mathcal{S}_0 \subseteq \mathsf{S}^*(A)$ and $\mathcal{S}_1 \subseteq \mathsf{S}^*(A \to B)$ and z_0 and z_1 such that all elements of $\mathcal{S}_0$ are true witnessed below z_0 and and all elements of $\mathcal{S}_1$ are true witnessed below z_1 and $\mathsf{taut}^*(\bigwedge(Y_x \cup \mathcal{S}^N_0) \to A)$ and $\mathsf{taut}^*(\bigwedge(Y_x \cup \mathcal{S}^N_1) \to (A \to B))$. Clearly, all elements of $\mathcal{S}_0 \cup \mathcal{S}_1$ are true witnessed below $z := \mathsf{max}(z_0, z_1)$. Moreover, $\mathsf{taut}^*(\bigwedge(Y_x \cup \mathcal{S}^N_0 \cup \mathcal{S}^N_1) \to B)$. Let $\mathcal{S}_2 := (\mathcal{S}_0 \cup \mathcal{S}_1) \cap \mathsf{S}^*(B)$. By elementary propositional logic we find that $\mathsf{taut}^*(\bigwedge(Y_x \cup \mathcal{S}^N_2) \to B)$ (since the

atoms corresponding to elements of $\mathcal{S}_0 \cup \mathcal{S}_1$ that are not in $\theta(B)$ are irrelevant for the truth of $\theta(B)$ for a given assignment). The elements of $\mathcal{S}_2$ are witnessed below z. So $\blacksquare_x B$.

We prove item (vi). We reason in EA. Consider any A in the language of U and any x. We have:

$$\begin{array}{rcl}
\Box(\blacksquare_x^N A & \to & [\,\exists\,\mathcal{S} \subseteq \mathsf{S}^*(Y_x \cup \{A\})\,\exists z \\
& & (\forall S \in \mathcal{S}\, \mathsf{true}^z_{\Sigma_1}(S) \wedge \mathsf{taut}^*(\bigwedge(Y_x \cup \mathcal{S}^N) \to A))\,]^N \\
& \to & \bigvee_{\mathcal{S} \subseteq \mathsf{S}^*(Y_x \cup \{A\})} (\bigwedge_{S \in \mathcal{S}} \mathsf{true}^N_{\Sigma_1}(S) \wedge \mathsf{taut}^{*N}(\bigwedge(Y_x \cup \mathcal{S}^N) \to A)) \\
& \to & \bigvee_{\mathcal{S} \subseteq \mathsf{S}^*(Y_x \cup \{A\})} (\bigwedge \mathcal{S}^N \wedge (\bigwedge(Y_x \cup \mathcal{S}^N) \to A)) \\
& \to & \bigvee_{\mathcal{S} \subseteq \mathsf{S}^*(Y_x \cup \{A\})} (\bigwedge \mathcal{S}^N \wedge (\bigwedge Y_x \to A)) \\
& \to & (\bigwedge Y_x \to A) \\
& \to & A\,)
\end{array}$$

Finally (vii) follows by combining (iii) and (vi) and (viii) is simply the combination of (ii) and (vii). ❑

With our new notion of 'restricted provability' in hand, we can now proceed to give an 'Orey Hájek characterization' for Π_1-conservativity. We have Π_1 here rather than $\forall\Pi^{\mathsf{b}}_1$ because we need the totality of exponentiation to get everything going.

We write $\blacksquare_{V,M,n}$ for $\blacksquare$ define w.r.t. V, M. We have:

Theorem 4.10 (Orey-Hájek for Π_1-conservativity)**.** *Consider U, N and V, M, where N is an interpretation of EA in U and M is an interpretation of EA in V. Then, $(U, N) \rhd_{\Pi_1} (V, M)$ iff, for all n, we have $U \vdash \blacklozenge^N_{V,M,n} \top$.*

Proof. From left to right: Suppose $(U, N) \rhd_{\Pi_1} (V, M)$. By Lemma 4.9(vi) we have, for any n, that $V \vdash \blacklozenge^M_{V,M,n} \top$. Hence we also have $U \vdash \blacklozenge^N_{V,M,n} \top$.

From right to left: Suppose, for all n, $U \vdash \blacklozenge^N_{V,M,n} \top$. Suppose $V \vdash P^M$, for P in Π_1. It follows that $U \vdash \blacksquare^N_{V,M,n^*} P^M$, for sufficiently large n^*. We can write P as $\neg\, S$, where S is in Σ_1. Reason in U. Suppose S^N. Then, we have $\blacksquare^N_{V,M,n^*} S^M$. So, $\blacksquare^N_{V,M,n^*} \bot$. Quod non. Hence, we may conclude P^N. Leaving U, we see that $U \vdash P^N$. ❑

Open Question 4.11. Consider U, N and V, M, where N is an interpretation of S^1_2 in U and M is an interpretation of EA in V. Can we prove the following?

$(U, N) \rhd_{\forall\Pi^{\mathsf{b}}_1} (V, M)$ iff, for all n, we have $U \vdash \blacklozenge^N_{V,M,n} \top$. ❑

Finally we give our main theorem.

Theorem 4.12. *Suppose U is a theory and $N : \mathsf{EA} \lhd U$. Then, $A \blacktriangleright_{\mathsf{hbl}_0} B$ iff $A \rhd_{\Pi_1} B$.*

Proof. From left to right. This is Theorem 4.5.

From right to left. Suppose $A \rhd_{\Pi_1} B$. By the 'unformalized' version of Lemma 4.9(vi), we have: for all n, $\vdash B \to \blacklozenge^N_n B$ and, hence (†) for all n, $\vdash A \to \blacklozenge^N_n B$. We define the following predicate:

$$\blacktriangle C :\leftrightarrow (B \vee (A \wedge \exists x\, (\blacksquare_x (B \to C) \wedge \blacklozenge_x B))).$$

We claim that $\blacktriangle : A \blacktriangleright_{\mathsf{hbl}_0} B$ (w.r.t. for U, N). It is easy to see that $\vdash \blacktriangle \bot \leftrightarrow B$, $\vdash \blacktriangle \top \leftrightarrow (A \vee B)$ and $\vdash \blacktriangle\blacktriangle \bot \leftrightarrow (A \vee B)$.

Suppose $C \vdash D$. It follows that (‡) for some m, $\vdash \blacksquare_m (C \to D)$. Reason in U. Suppose $\blacktriangle C$. In case we have B, we are immediately done. Suppose $\neg B$. In that case, we have A and $\exists x\, (\blacksquare_x (B \to C) \wedge \blacklozenge_x B)$. Suppose $\blacksquare_{x_0} (B \to C)$ and $\blacklozenge_{x_0} B$. We may assume, by (†) that $x_0 \geq m$. By (‡) it follows that A and $\blacksquare_{x_0} (B \to D)$ and $\blacklozenge_{x_0} B$. So $\blacktriangle D$.

Reason in U. Suppose $\blacktriangle C$ and $\blacktriangle (C \to D)$. In case B, we immediately have $\blacktriangle D$. Suppose $\neg B$. It follows that A and for some x_0, x_1, we have $\blacksquare_{x_0} (B \to C)$ and $\blacksquare_{x_1} (B \to (C \to D))$ and $\blacklozenge_{x_0} B$ and $\blacklozenge_{x_1} B$. Let $x := \mathsf{max}(x_0, x_1)$. We find: $\blacksquare_x (B \to C)$ and $\blacksquare_x (B \to (C \to D))$ and $\blacklozenge_x B$. It follows that A and $\blacksquare_x (B \to D)$ and $\blacklozenge_x B$. I.o.w., $\blacktriangle D$. ❑

We have seen, in the previous subsection, that for sententially essentially reflexive theories, interpretability and the HBL jump relation coincide. In the present subsection, we have seen that for extensions of EA, Π_1-conservativity and the HBL_0 jump relations coincide.

In the next section we will provide an example that illustrates that the L jump relation does *not* coincide with interpretability for a wide range of theories.

5. The Kreisel Condition and a Separating Example

Suppose we have a theory U and an interpretation N of EA in U. We assume that the theory U is Δ^{b}_1-axiomatized. As before, the interpretation N provides us the Gödel numbers we use. In this section we want to achieve two things at once. In the first place, we want to produce a Σ_1-predicate $\boxplus$ for U such that $\boxplus^N$ is an L-predicate that satisfies the Kreisel condition: $U \vdash \boxplus^N A$ iff $U \vdash A$, for all U-sentences A. In the second place, we we want $\boxplus^N \bot$ to be a separating example between $\blacktriangleright$ and $\rhd$. Thus, we want: $U \blacktriangleright (U + \boxplus^N \bot)$, but $U \not\rhd (U + \boxplus^N \bot)$.

Let P be any formula defining a set of N-numbers in U. We assume that P starts with a quantifier. Note that we can always add a vacuous quantifier to obtain the desired effect. We treat P as a modal operator, writing $\triangle A$ for $P(\underline{\ulcorner A \urcorner})$. Note that, for the moment, we do not demand any further properties from P.

Consider any set of sentences Z in the language of U. The set Z generates a propositional language as follows. First we define $\mathsf{sub}(Z)$ as the smallest set X such that:

i. $Z \subseteq X$,

ii. if $A \wedge B$ is in X then A and B are in X, and similarly for the other propositional connectives.

iii. if $\triangle A$ is in X, then so is A.

In our set-up, we treat the formulas starting with quantifiers as atoms. Consider any set of sentences Z. We define $Z \vdash_0 C$, if C follows from Z using modus ponens and $\triangle$-necessitation, i.e., the rule that if we have derived A, we may infer $\triangle A$. A $\vdash_0$-proof from Z is simply a sequence of formulas $D_0, \ldots, D_{k-1}$, where that D_i are either in Z or follow from earlier elements of the sequence by our two rules.

Suppose Z is finite. Consider any $\vdash_0$-proof π from Z. Let γ be an occurrence-as-subconclusion of a formula C in π. We note that if C is in $\mathsf{sub}(Z)$, then all formulas occurring above γ as subconclusions are subformulas of formulas in Z. If C is not in $\mathsf{sub}(Z)$, then C is of the form $\triangle D$ and the last rule applied is $\triangle$-necessitation.

Thus, any proof witnessing $Z \vdash_0 A$ has the following form: $A = \triangle^n B$ (n may be 0), where B is subformula of of formula in Z. From B to A we have necessitation inferences, and the proof of B contains only elements of of $\mathsf{sub}(Z)$.

If a $\vdash_0$-proof containing only elements of $\mathsf{sub}(Z)$ is longer than the number of subformulas of formulas of Z, then a certain subconclusion will occur twice sequentially. Thus we can shorten the proof by omitting all but the first occurrence of the subconclusion. Hence, proofs containing only subformulas of formulas in Z can be reduced to proofs with as length at most the number of subformulas of formulas in Z.

We may conclude that $Z \vdash_0 A$ is decidable. We can easily see that our decidability proof can be formalized in EA.

Let Y_n be the set of A such that, for some $p < n$, $\mathsf{proof}_U(p, A)$. Let $\boxplus^P_x A$ stand for (the arithmetization of) $Y_x \vdash_0 A$, and let $\boxplus^P A$ stand for $\exists x \ \boxplus^P_x A$. We note that P only occurs coded in the definition of $\boxplus^P_x$ and $\boxplus^P$.

Consider any Σ_1-sentence S of the form $\exists x \, S_0(x)$, where S_0 is $\Delta_0(\mathsf{exp})$. Using the Gödel Fixed Point Lemma, we find a formula $\boxtimes$ (or, more explicitly, $\boxtimes^{[S]}$) with:

$$\mathsf{EA} \vdash \boxtimes A \leftrightarrow \ \boxplus^{\boxtimes^N} A < S.$$

Note that we take $P := \boxtimes^N$. We define $\boxtimes^\perp A$ by $S \leq \ \boxplus^{\boxtimes^N} A$.

Theorem 5.1. *We have:*

i. $\mathsf{EA} \vdash \forall A, B \, ((\boxtimes A \wedge \boxtimes (A \to B)) \to \boxtimes B)$,

ii. $\mathsf{EA} \vdash \forall A \, (\boxtimes A \to \boxtimes \boxtimes^N A)$,

iii. $\mathsf{EA} \vdash \forall A \, (\boxtimes A \to \Box \boxtimes^N A)$,

iv. $\mathsf{EA} \vdash \forall A \, (\boxtimes A \to \Box A)$.

v. $\mathsf{EA} + \neg S \vdash \forall A \, (\Box A \leftrightarrow \boxtimes A)$.

vi. EA *verifies that, if* S *is false, then,* $\boxtimes^N$ *is an L-predicate for* U.

vii. $\mathsf{EA} \vdash \Box A \to (\boxtimes A \vee \boxtimes^\perp A)$.

Proof. We reason in EA. We write s for the minimal witness of S. In case $\neg S$, we treat s as ∞ in the obvious way.

Ad (i): Suppose $\boxslash A$ and $\boxslash(A \to B)$. It follows that, for some $x < s$, we have $\boxplus_x^{\boxslash^N} A$ and $\boxplus_x^{\boxslash^N}(A \to B)$. Hence, since $\boxplus_x^{\boxslash^N}$ is closed under modus ponens by construction, we find $\boxplus_x^{\boxslash^N} B$. Ergo, $\boxslash B$.

Ad (ii): Suppose $\boxslash A$. It follows that, for some $x < s$, we have $\boxplus_x^{\boxslash^N} A$. Since $\boxplus_x^{\boxslash^N}$ is closed under $\boxslash^N$-necessitation by construction, we find $\boxplus_x^{\boxslash^N} \boxslash^N A$. Ergo, $\boxslash \boxslash^N A$.

Ad (iii): This is just Σ_1-completeness.

Ad (iv): Suppose $\boxslash A$. Then, for some $x < s$, we have $\boxplus_x^{\boxslash^N} A$. We prove by induction on proof-length that, for every $\vdash_0$-proof p from Y_x of a B, there is a matching ordinary proof q of B. To make the induction possible, we need a multi-exponential bound on the q. We will discuss this bound after describing the transformations. We note that we can consider the $\vdash_0$-proof p as the witness for $\boxplus_x^{\boxslash^N} B$, since the U-proofs needed for verifying that an element of the proof is in Y_x are all bounded by x.

In case B is in Y_x, we are guaranteed a proof $q < x$ of B.

Suppose we have concluded B from C and $C \to B$. Say, we have $\vdash_0$-proofs p_0 of C and p_1 of $C \to B$, then by the induction hypothesis we have proofs q_0 of C and q_1 of $C \to B$. Clearly, we can find a proof q of B with length linear in the lengths of q_0 and q_1.

Suppose we have concluded $B = \boxslash^N C$ from C. Suppose our $\vdash_0$-proof of C is p. Clearly p witnesses $\boxslash C$. So we can construct an ordinary proof of order 2^{2^p} to show $\Box \boxslash^N C$ —following the usual proof of Σ_1-completeness. (Note that we do not need the Induction Hypotheses here.)

On the basis of the two transformations, we can easily see that we can estimate the ordinary proofs q by 2^{2^p}, where p is the $\vdash_0$-proof from which they are derived.

Ad (v): This is immediate using (iv).

Ad (vi): We reason in EA. By (i) and (iii), we have that:

$$\Box((\boxslash^N A \wedge \boxslash^N(A \to B)) \to \boxslash^N B) \text{ and } \Box(\boxslash^N A \to \boxslash^N \boxslash^N A).$$

Suppose $\neg S$ and $\Box A$. Then, by (v), we find that $\boxslash A$. So, by (iii), $\Box \boxslash^N A$.

Ad (vii): This is immediate since $\mathsf{EA} \vdash \Box A \to \boxplus^{\boxslash} A$. ❑

Next, we find using the Gödel Fixed Point Lemma, a sentence R such that:

$$\mathsf{EA} \vdash R \leftrightarrow \exists C\,(\Box \boxslash^{[R],N} C \leq \boxplus^{\boxslash^{[R],N}} C).$$

Inspecting the fixed point construction we may arrange it so that R is of the form:

$$\exists p\, \exists C < p\,(\mathsf{proof}(p, \boxslash^{t,N} C) \wedge \forall x < p \neg \boxplus_x^{\boxslash^{t,N}} C),$$

where t is an elementary term that evaluates to (the Gödel number of) R. We note that R is of the form $\exists p\, R_0(p)$, where R_0 is $\Delta_0(\mathsf{exp})$.

Finally we define $\boxplus A :\leftrightarrow \boxminus^{[R]} A$. So,

$$\mathsf{EA} \vdash R \leftrightarrow \exists C\, (\Box\, \boxplus^N C \leq\ \boxplus^{\boxplus^N} C).$$

Theorem 5.2. *We have:*

a. $\mathsf{EA} \vdash R \to \Box\bot$,

b. $\mathsf{EA} \vdash \Diamond\top \to (\boxplus A \leftrightarrow \Box A)$,

c. $\mathsf{EA} \vdash \Box\, \boxplus^N A \leftrightarrow \Box A$,

d. Suppose that U is EA-verifiably sequential and essentially reflexive w.r.t. N. Then, $\mathsf{EA} \vdash \top \rhd \boxplus^N \bot \to \Box\bot$.

Proof. Ad (a): We reason in EA. Suppose R. It follows that, for some C, we have $\Box\, \boxplus^N C \leq\ \boxplus^{\boxplus^N} C$. It follows that $\Box\, \boxplus^N C$, i.e. (a) $\Box(\ \boxplus^{\boxplus^N} C < R)^N$. On the other hand, R implies $R \leq\ \boxplus^{\boxplus^N} C$. So, by Σ_1-completeness, (b) $\Box(R \leq\ \boxplus^{\boxplus} C)^N$. By (a) and (b), we may conclude that $\Box\bot$.

Ad (b): The desired result is immediate by (a) and Theorem 5.1(v).

Ad (c): The right-to-left direction is immediate from Theorem 5.1(iii). We prove left-to-right. We reason in EA. Suppose $\Box\, \boxplus^N A$. We want to show $\Box A$. In case we have R, we are immediately done by (a). If we have $\neg R$, it follows that we cannot have $\Box\, \boxplus^N A <\ \boxplus^{\boxplus^N} A$. So, we must have $\boxplus^{\boxplus^N} A$, and hence $\boxminus A$. By Theorem 5.1(v), we find $\Box A$.

Ad (d): Suppose that U is EA-verifiably sequential and essentially reflexive w.r.t. N. We reason in EA. Suppose $\top \rhd \boxplus^N \bot$. Then also $\top \rhd \neg\ \boxplus^{\bot,N} \bot$. Since $\boxplus^\bot \bot$ is Σ_1, it follows that $\Box\neg\ \boxplus^{\bot,N} \bot$. By Theorem 5.1(vii), $\Box(\Box^N \bot \to (\boxplus^N \bot \vee \boxplus^{\bot,N} \bot))$. Hence, (†) $\Box(\Box^N \bot \to \boxplus^N \bot)$. It follows that $\Box(\Box^N \Box^N \bot \to \Box^N\, \boxplus^N\, \bot)$. Hence, by (c), we find $\Box(\Box^N \Box^N \bot \to \Box^N \bot)$, and so (‡) $\Box\Box^N \bot$. Combining (†) and (‡), we obtain $\Box\, \boxplus^N \bot$. Hence, again by Theorem 5.1(vii), we may conclude $\Box\bot$. ❑

Here (c) gives us the promised result that $\boxplus^N$ has the Kreisel property. Moreover (d) shows that $\blacktriangleright$ strictly extends $\rhd$, since we do have $\top \blacktriangleright \boxplus^N \bot$. In fact $\boxplus^N \bot$ is a Σ_1 Rosser sentence for U. So, we have an example of a Σ_1 Rosser sentence that can be $\blacktriangleright$-reached from $\top$.

References

[AB04] S.N. Artemov and L.D. Beklemishev. Provability logic. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic, 2nd ed.*, volume 13, pages 229–403. Springer, Dordrecht, 2004.

[Bek04] L.D. Beklemishev. Provability algebras and proof-theoretic ordinals. *Annals of Pure and Applied Logic*, 128:103–124, 2004.

[Bek05] L.D. Beklemishev. Reflection principles and provability algebras in formal arithmetic. *Russian Mathematical Surveys*, 60(2):197–268, 2005.

[Bek06] L.D. Beklemishev. The worm principle. In Z. Chatzidakis, P. Koepke, and W. Pohlers, editors, *Logic Colloquium '02*, volume 27 of *Lecture Notes in Logic*, pages 75–95. A.K. Peters and CRC press, Natick, Massachusetts, 2006.

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

[BJV05] L.D. Beklemishev, J.J. Joosten, and M. Vervoort. A finitary treatment of the closed fragment of Japaridze's provability logic. *Journal of Logic and Computation*, 15(4):447–463, 2005.

[Boo93] G. Boolos. *The logic of provability*. Cambridge University Press, Cambridge, 1993.

[BS91] G. Boolos and G. Sambin. Provability: the emergence of a mathematical modality. *Studia Logica*, 50:1–23, 1991.

[Bus86] S.R. Buss. *Bounded Arithmetic*. Bibliopolis, Napoli, 1986.

[Bus11] S.R. Buss. Cut elimination *in situ*. `http://math.ucsd.edu/~sbuss/`, 2011.

[BV05] L.D. Beklemishev and A. Visser. On the limit existence principles in elementary arithmetic and Σ^0_n-consequences of theories. *Annals of Pure and Applied Logic*, 136(1–2):56–74, 2005.

[DJ94] G. Dzhaparidze (Japaridze). A simple proof of arithmetical completeness for Π_1-conservativity logic. *Notre Dame Journal of Formal Logic*, 35:346–354, 1994.

[Fef60] S. Feferman. Arithmetization of metamathematics in a general setting. *Fundamenta Mathematicae*, 49:35–92, 1960.

[Fef97] S. Feferman. My route to arithmetization. *Theoria*, 63(3):168–181, 1997.

[Ger03] P. Gerhardy. Refined Complexity Analysis of Cut Elimination. In Matthias Baaz and Johann Makovsky, editors, *Proceedings of the 17th International Workshop CSL 2003*, volume 2803 of *LNCS*, pages 212–225. Springer-Verlag, Berlin, 2003.

[Ger05] P. Gerhardy. The Role of Quantifier Alternations in Cut Elimination. *Notre Dame Journal of Formal Logic*, 46(2):165–171, 2005.

[GJ08] E. Goris and J.J. Joosten. Modal matters in interpretability logic. *Logic Journal of IGPL*, 16(4):371–412, 2008.

[Göd33] K. Gödel. Ein Interpretation des intuitionistischen Aussagenkalküls. In *Ergebnisse eines mathematischen Kolloquiums*, volume 4, pages 39–40. 1933. Reprinted as: *An interpretation of the intuitionistic propositional calculus*, in: Feferman, S., ed., Gödel Collected Works I, publications 1929-1936, 300-303.

[HB39] D. Hilbert and P. Bernays. *Grundlagen der Mathematik II*. Springer, Berlin, 1939. second edition: 1970.

[HC96] G.E. Hughes and M.J. Cresswell. *A new introduction to modal logic*. Burns & Oates, 1996.

[HL01] Leon Horsten and Hannes Leitgeb. No future. *Journal of philosophical logic*, 30(3):259–265, 2001.

[HLW03] V. Halbach, H. Leitgeb, and P. Welch. Possible-worlds semantics for modal notions conceived as predicates. *Journal of Philosophical Logic*, 32(2):179–223, 2003.

[HM90] P. Hájek and F. Montagna. The logic of Π_1-conservativity. *Archiv für Mathematische Logik und Grundlagenforschung*, 30:113–123, 1990.

[HM92] P. Hájek and F. Montagna. The logic of Π_1-conservativity continued. *Archiv für Mathematische Logik und Grundlagenforschung*, 32:57–63, 1992.

[HP93] P. Hájek and P. Pudlák. *Metamathematics of First-Order Arithmetic*. Perspectives in Mathematical Logic. Springer, Berlin, 1993.

[Ign91] K.N. Ignatiev. Partial conservativity and modal logics. Technical Report X-91-04, ILLC, University of Amsterdam, 1991.

[Jap85] G. Japaridze. The polymodal logic of provability. In *Intensional Logics and Logical Structure of Theories: Material from the fourth Soviet–Finnish Symposium on Logic, Telavi*, pages 16–48, 1985.

[JdJ98] G. Japaridze and D. de Jongh. The logic of provability. In S. Buss, editor, *Handbook of proof theory*, pages 475–546. North-Holland Publishing Co., Amsterdam, 1998.

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

[Kal91] M.B. Kalsbeek. Towards the interpretability logic of $\mathrm{I}\Delta_0 + \mathsf{EXP}$. Logic Group Preprint Series 61, Faculty of Humanities, Philosophy, Utrecht University, Janskerkhof 13A, 3512 BL Utrecht, `http://www.phil.uu.nl/preprints/lgps/`, 1991.

[KM60] D. Kaplan and R. Montague. A paradox regained. *Notre Dame Journal of Formal Logic*, 1, 1960.

[Kre53] G. Kreisel. On a problem of Henkin's. *Indagationes mathematicae*, 15:405–406, 1953.

[Lin96] P. Lindström. Provability logic – a short introduction. *Theoria*, 62(1-2):19–61, 1996.

[Löb55] M.H. Löb. Solution of a problem of Leon Henkin. *Journal of Symbolic Logic*, 20:115–118, 1955.

[Mon63] Richard Montague. Syntactical treatments of modality, with corollaries on reflexion principles and finite axiomatizability. *Acta philosophica fennica*, 16:153–167, 1963.

[Mon78] F. Montagna. On the algebraization of a Feferman's predicate (the algebraization of theories which express Theor; X). *Studia Logica*, 37:221–236, 1978.

[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.

[Pud83] P. Pudlák. Some prime elements in the lattice of interpretability types. *Transactions of the American Mathematical Society*, 280:255–275, 1983.

[Pud85] P. Pudlák. Cuts, consistency statements and interpretations. *The Journal of Symbolic Logic*, 50(2):423–441, 1985.

[RL86] J. des Rivières and H.J. Levesque. The consistency of syntactical treatments of knowledge. In *Proceedings of the 1986 Conference on Theoretical aspects of reasoning about knowledge*, pages 115–130. Morgan Kaufmann Publishers Inc., 1986.

[SF13] J. Stern and M. Fisher. Paradoxes of interaction. Unpublished manuscript, 2013.

[Sha88] V.Yu. Shavrukov. The logic of relative interpretability over Peano arithmetic (in Russian). Technical Report Report No.5, Stekhlov Mathematical Institute, Moscow, 1988.

[Sha94] V.Yu. Shavrukov. A smart child of Peano's. *Notre Dame Journal of Formal Logic*, 35:161–185, 1994.

[Sol76] R.M. Solovay. Provability interpretations of modal logic. *Israel Journal of Mathematics*, 25:287–304, 1976.

[Šve00] V. Švejdar. On provability logic. *Nordic Journal of Philosophical Logic*, 4(2):95–116, 2000.

[Tho80] R. H. Thomason. A note on syntactical treatments of modality. *Synthese*, 44(3):391–395, 1980.

[Vis89] A. Visser. Peano's smart children: A provability logical study of systems with built-in consistency. *Notre Dame Journal of Formal Logic*, 30(2):161–196, 1989.

[Vis90] A. Visser. Interpretability logic. In P.P. Petkov, editor, *Mathematical logic, Proceedings of the Heyting 1988 summer school in Varna, Bulgaria*, pages 175–209. Plenum Press, Boston, 1990.

[Vis93] A. Visser. The unprovability of small inconsistency. *Archive for Mathematical Logic*, 32(4):275–298, 1993.

[Vis98] A. Visser. An Overview of Interpretability Logic. In M. Kracht, M. de Rijke, H. Wansing, and M. Zakharyaschev, editors, *Advances in Modal Logic*, volume 1, 87 of *CSLI Lecture Notes*, pages 307–359. Center for the Study of Language and Information, Stanford, 1998.

[Vis09] A. Visser. Cardinal arithmetic in the style of Baron von Münchhausen. *Review of Symbolic Logic*, 2(3):570–589, 2009. doi: 10.1017/S1755020309090261.

[Vis11] A. Visser. Can we make the Second Incompleteness Theorem coordinate free. *Journal of Logic and Computation*, 21(4):543–560, 2011. First published online August 12, 2009, doi: 10.1093/logcom/exp048.

[Vis12] A. Visser. Peano Basso and Peano Corto. Logic Group Preprint Series 298, Faculty of Humanities, Philosophy, Utrecht University, Janskerkhof 13A, 3512 BL Utrecht, http://www.phil.uu.nl/preprints/lgps/, 2012.

[Vis13] A. Visser. What is the right notion of sequentiality? In P. Cégielski, C. Charampolas, and C. Dimitracopoulos, editors, *New Studies in Weak Arithmetics*, volume 211 of *CSLI Lecture Notes*, pages 229–272. CSLI Publications and Presses Universitaires du Pôle de Recherche et d'Enseingement Supérieur Paris-est, Stanford, 2013.

Appendix A. Basic Facts and Definitions

In this appendix we explain some basic notions.

The appendix still should be checked. Do we have everything here? Are the notations coherent with the rest?

A.1. **Theories.** Theories are, in this paper, theories of first-order predicate logic, that have a finite signature and that are axiomatized by an axiom set that is represented by a Δ_1^{b}-formula.[3] Theories will usually be one-sorted, but we will consider a few times two-sorted theories.

The formula specifying the axiom set is part of the data for the theory. Thus, we treat theories *intensionally* and not as mere sets of theorems.

We say that a theory is *finitely axiomatized* if its axiomatization has the form $\bigvee_{i<n} x = \underline{\ulcorner A_i \urcorner}$. Note that S^1_2 may prove that a theory has an axiom-set of, say, less than two axioms, without being able to prove the equivalence of the formula defining the axiom set with any formula of the prescribed form.

Our official signatures are relational, however, via the term-unwinding algorithm, we can also accommodate signatures with functions.

A.2. **Translations and Interpretations.** We present the notion of *m-dimensional interpretation without parameters*. There are two extensions of this notion: we can consider piecewise interpretations and we can add parameters. We will not treat these extensions in this paper.

Consider two signatures Σ and Θ. An m-dimensional translation $\tau : \Sigma \to \Theta$ is a quadruple $\langle \Sigma, \delta, \mathcal{F}, \Theta \rangle$, where $\delta(v_0, \ldots, v_{m-1})$ is a Θ-formula and where for any n-ary predicate P of Σ, $\mathcal{F}(P)$ is a formula $A(\vec{v}_0, \ldots, \vec{v}_{n-1})$ in the language of signature Θ, where $\vec{v}_i = v_{i0}, \ldots, v_{i(m-1)}$. Both in the case of δ and A all free variables are among the variables shown. Moreover, if $i \neq j$ and $k \neq \ell$, then v_{ik} is syntactically different from $v_{j\ell}$.

We demand that we have $\vdash \mathcal{F}(P)(\vec{v}_0, \ldots, \vec{v}_{n-1}) \to \bigwedge_{i<n} \delta(\vec{v}_i)$. Here $\vdash$ is provability in predicate logic. This demand is inessential, but it is convenient to have.

We define B^τ as follows:

- $(P(x_0, \ldots, x_{n-1}))^\tau := \mathcal{F}(P)(\vec{x}_0, \ldots, \vec{x}_{n-1})$.
- $(\cdot)^\tau$ commutes with the propositional connectives.
- $(\forall x\, A)^\tau := \forall \vec{x}\, (\delta(\vec{x}) \to A^\tau)$.
- $(\exists x\, A)^\tau := \exists \vec{x}\, (\delta(\vec{x}) \wedge A^\tau)$.

There are two worries about this definition. First, what variables $\vec{x}_i$ on the side of the translation A^τ correspond with x_i in the original formula A? The second worry is that substitution of variables in δ and $\mathcal{F}(P)$ may cause variable clashes. These worries are never important in practice: we choose 'suitable' sequences $\vec{x}$ to correspond to variables x, and we avoid clashes by α-conversions. However, if we want to give precise definitions of translations and, for example, of composition of translations these problems come into play. These problems are clearly solvable, but they are beyond the scope of this paper.

We allow identity to be translated to a formula that is not identity. There are several important operations on translations.

[3] See [Bus86] or [HP93] for an explanation of the relevant formula classes.

- id_Σ is the identity translation. We take $\delta_{\mathsf{id}_\Sigma}(v) := v = v$ and $\mathcal{F}(P) := P(\vec{v})$.
- We can compose translations. Suppose $\tau : \Sigma \to \Theta$ and $\nu : \Theta \to \Lambda$. Then $\nu \circ \tau$ or $\tau\nu$ is a translation from Σ to Λ. We define:
  - $\delta_{\tau\nu}(\vec{v}_0, \ldots, \vec{v}_{m_\tau - 1}) := \bigwedge_{i<m_\tau} \delta_\nu(\vec{v}_i) \wedge (\delta_\tau(v_0, \ldots, v_{m_\tau - 1}))^\nu$.
  - $P_{\tau\nu}(\vec{v}_{0,0}, \ldots, \vec{v}_{0,m_\tau-1}, \ldots \vec{v}_{n-1,0}, \ldots, \vec{v}_{n-1,m_\tau-1}) :=$ $\bigwedge_{i<n, j<m_\tau} \delta_\nu(\vec{v}_{i,j}) \wedge (P(v_0, \ldots, v_{n-1})^\tau)^\nu$.
- Let $\tau, \nu : \Sigma \to \Theta$ and let A be a sentence of signature Θ. We define the disjunctive translation $\sigma := \tau\langle A\rangle\nu : \Sigma \to \Theta$ as follows. We take $m_\sigma := \mathsf{max}(m_\tau, m_\nu)$. We write $\vec{v} \restriction n$, for the restriction of $\vec{v}$ to the first n variables, where $n \leq \mathsf{length}(\vec{v})$.
  - $\delta_\sigma(\vec{v}) := (A \wedge \delta_\tau(\vec{v} \restriction m_\tau)) \vee (\neg A \wedge \delta_\nu(\vec{v} \restriction m_\nu))$.
  - $P_\sigma(\vec{v}_0, \ldots, \vec{v}_{n-1}) := (A \wedge P_\tau(\vec{v}_0 \restriction m_\tau, \ldots, \vec{v}_{n-1} \restriction m_\tau)) \vee$ $(\neg A \wedge P_\nu(\vec{v}_0 \restriction m_\nu, \ldots, \vec{v}_{n-1} \restriction m_\nu))$

Note that in the definition of $\tau\langle A\rangle\nu$ we used a padding mechanism. In case, for example, $m_\tau < m_\nu$, the variables $v_{m_\tau}, \ldots, v_{m_\nu - 1}$ are used 'vacuously' when we have A. If we had piecewise interpretations, where domains are built up from pieces with possibly different dimensions, we could avoid padding by building the domain of disjoint pieces with different dimensions.

A translation relates signatures; an interpretation relates theories. An interpretation $K : U \to V$ is a triple $\langle U, \tau, V\rangle$, where U and V are theories and $\tau : \Sigma_U \to \Sigma_V$. We demand: for all axioms A of U, we have $V \vdash A^\tau$. Here are some further definitions.

- $\mathsf{ID}_U : U \to U$ is the interpretation $\langle U, \mathsf{id}_{\Sigma_U}, U\rangle$.
- Suppose $K : U \to V$ and $M : V \to W$. Then, $KM := M \circ K : U \to W$ is $\langle U, \tau_M \circ \tau_K, W\rangle$.
- Suppose $K : U \to (V + A)$ and $M : U \to (V + \neg A)$. Then $K\langle A\rangle M : U \to V$ is the interpretation $\langle U, \tau_K\langle A\rangle\tau_M, V\rangle$. In an appropriate category $K\langle A\rangle M$ is a special case of a product.

The notation $K : U \to V$ is inspired by the idea of interpretations as arrows in a category. There is also an intuition of interpretability as a generalization of provability. The traditional notations and notions associated to this intuition are:

- $K : U \lhd V$ stands for $K : U \to V$.
- $K : V \rhd U$ stands for $K : U \to V$.
- $U \lhd V$ stands for $\exists K\ K : U \lhd V$. We say: U is *interpretable* in V.
- $V \rhd U$ stands for $\exists K\ K : V \rhd U$. We say: V *interprets* U.
- $U \equiv V$ stands for $U \rhd V$ and $V \rhd U$. We say: V and U are *mutually interpretable*.

A basic insight in concerning interpretability is the Gödel-Hilbert-Bernays-Wang-Henkin-Feferman Theorem.

Theorem A.1. *Consider $N : \mathsf{S}^1_2 \lhd U$. We assume that U is Δ^{b}_1-axiomatized. Then, we can construct an interpretation $H : (U + \Diamond^N_U \top) \rhd U$. We call H: the Henkin interpretation. This interpretation has the additional feature that we can construct inside U a truth-predicate T such that for some definable cut I of N the commutation conditions for the language coded in I are U-verifiable.*

The proof uses the formalized Henkin construction to produce an interpretation $H : (U + \Diamond^N_U \top) \rhd U$. The basic intuition here is, of course, that an interpretation is a uniform internal model construction. The lack of induction in our setting has to be systematically compensated by going to shorter and shorter definable cuts of N.

A.3. **Sequential Theories.** A sequential theory provides an interpretation N of a weak number theory, say S^1_2, and sequences of all objects of the domain of the theories with projections in N. We can use these sequences to develop partial satisfaction predicates. Using these we can prove restricted consistency statements of U in U.

The notion of sequential theory has an very simple definition discovered by Pavel Pudlák. We first need the definition of a very weak set theory. The theory Adjunctive Set Theory or AS is a one-sorted theory with a binary relation $\in$.

$\mathsf{AS}1 \vdash \exists x \, \forall y \; y \notin x$,

$\mathsf{AS}2 \vdash \forall x, y \, \exists z \, \forall u \, (u \in z \leftrightarrow (u \in x \vee u = y))$.

We note that we do not demand extensionality. For example, in AS we could have lots of 'empty sets'.

An interpretation is *direct* iff it is one-dimensional, unrelativised (that is, it has the trivial domain) and identity preserving (that is, it translates identity to identity).

A theory U is sequential iff it directly interprets AS. By a substantial bootstrap, we can define, in a sequential theory U, an interpretation N of a weak number theory, sequences of all objects, etc.

For details see, for example, [Pud83], [Pud85], [MPS90], [HP93], [Vis09] and [Vis13].

We can generalize the notion of sequentiality a bit to *poly-sequentiality* by replacing *direct interpretation* in the definition by its obvious generalization to the m-dimensional case.

A.4. **Complexity Measures.** In sequential theories we can define partial satisfaction predicates for formulas with complexity below n, for any n. The presence of these predicates has as a consequence that for any sequential theory U and for any n, we can find an interpretation N of a weak arithmetic like Buss' S^1_2 in U such that $U \vdash \mathsf{con}^N_n(U)$. See, for example, [Vis93] for more details. We give the relevant definitions of complexity notions.

Restricted provability plays an important role in this paper. An n-proof is a proof from axioms with Gödel number smaller or equal than n only involving formulas of complexity smaller or equal than n. To work conveniently with this notion, a good complexity measure is needed. This should satisfy three conditions. (i)

Eliminating terms in favour of a relational formulation should raise the complexity only by a fixed standard number. (ii) Translation of a formula via the translation corresponding to an interpretation K should raise the complexity of the formula by a fixed standard number depending only on K. (iii) The tower of exponents involved in cut-elimination should be of height linear in the complexity of the formulas involved in the proof.

Such a good measure of complexity together with a verification of desideratum (iii) —a form of nesting degree of quantifier alternations— is supplied in the work of Philipp Gerhardy. See [Ger03] and [Ger05]. It is also provided by Samuel Buss in his preliminary draft [Bus11]. Buss also proves that (iii) is fulfilled.

Gerhardy's measure corresponds to the following formula classes:

- AT is the class of atomic formulas.
- $\mathsf{N}^\star_{-1} = \Sigma^\star_{-1} = \Pi^\star_{-1} := \emptyset$.
- $\mathsf{N}^\star_n ::= \mathsf{AT} \mid \neg \mathsf{N}^\star_n \mid (\mathsf{N}^\star_n \wedge \mathsf{N}^\star_n) \mid (\mathsf{N}^\star_n \vee \mathsf{N}^\star_n) \mid (\mathsf{N}^\star_n \to \mathsf{N}^\star_n) \mid \forall\, \Pi^\star_n \mid \exists\, \Sigma^\star_n$.
- $\Sigma^\star_n ::= \mathsf{AT} \mid \neg \Pi^\star_n \mid (\mathsf{N}^\star_{n-1} \wedge \mathsf{N}^\star_{n-1}) \mid (\Sigma^\star_n \vee \Sigma^\star_n) \mid (\Pi^\star_n \to \Sigma^\star_n) \mid \forall\, \Pi^\star_{n-1} \mid \exists\, \Sigma^\star_n$.
- $\Pi^\star_n ::= \mathsf{AT} \mid \neg \Sigma^\star_n \mid (\Pi^\star_n \wedge \Pi^\star_n) \mid (\mathsf{N}^\star_{n-1} \vee \mathsf{N}^\star_{n-1}) \mid (\mathsf{N}^\star_{n-1} \to \mathsf{N}^\star_{n-1}) \mid \forall\, \Pi^\star_n \mid \exists\, \Sigma^\star_{n-1}$.

We may define $\rho(A)$ as the minimal n such that A is in $\mathsf{N}^\star_n$.[4]

Samuel Buss gives the following formula classes.

- $\Sigma^*_0 = \Pi^*_0 =$ the class of quantifier-free formulas.
- $\Sigma^*_n ::= \Sigma^*_{n-1} \mid \Pi^*_{n-1} \mid \neg\, \Pi^*_n \mid (\Sigma^*_n \wedge \Sigma^*_n) \mid (\Sigma^*_n \vee \Sigma^*_n) \mid (\Pi^*_n \to \Sigma^*_n) \mid \exists\, \Sigma^*_n$.
- $\Pi^*_n ::= \Sigma^*_{n-1} \mid \Pi^*_{n-1} \mid \neg\, \Sigma^*_n \mid (\Pi^*_n \wedge \Pi^*_n) \mid (\Pi^*_n \vee \Pi^*_n) \mid (\Sigma^*_n \to \Pi^*_n) \mid \forall\, \Pi^*_n$.

We may define $\rho(A)$ as the smallest n such that A is in Σ^*_n. This is the same measure, as was employed in [Vis93]. For our purposes it does not matter whether we use Gerhardy's of Buss' definition.

We use $\mathsf{proof}_{U,n}$ for the proof predicate where only U-axioms with Gödel numbers $\leq n$ are allowed and where the formulas occurring in the proof are in the complexity class Γ_n of all formulas of complexity $\leq n$. Similarly we use $U \vdash_n A$, $\mathsf{con}_n(U)$, $\Box_{U,m} A$, etc.

We end with some basic facts concerning sequential theories and restricted provability. A finitely axiomatized sequential theory is mutually interpretable with its own restricted consistency over S^1_2.

Theorem A.2. *Suppose A is finitely axiomatized and sequential. We have:*

$$A \equiv (\mathsf{S}^1_2 + \Diamond_{A,\rho(A)} \top).$$

For a proof, see, [Pud85] or [HP93]. We note that the right-to-left direction of the result is a variant of the Gödel-Hilbert-Bernays-Wang-Henkin-Feferman Theorem. An important point here is that the existence of a truth-predicate for the witnessing Henkin interpretation is lost when we switch from ordinary consistency to restricted

[4]Vincent van Oostrom gave a variant of this formulation of Gerhardy's measure in conversation.

consistency. (If this were not the case, we would obtain a contradiction with the Second Incompleteness Theorem.)

We provide an partial analogue of Theorem A.2 for infinitely axiomatized theories. The $\mho$-functor is given as follows.[5]

- $\mho(U) := \mathsf{S}^1_2 + \{\Diamond_{U,n}\top \mid n \in \omega\}$.

The central fact about the $\mho$-functor is as follows:

Theorem A.3. *Suppose U is sequential. We have: $U \rhd_{\mathsf{loc}} V \Leftrightarrow \mho(U) \rhd V$.*

If we restrict ourselves to sequential theories, the theorem tells us that $\mho$ is the right adjoint of the embedding functor of $\lhd$ considered as a preorder category into $\lhd_{\mathsf{loc}}$ considered as a preorder category. For a proof, see [Vis11] We note that it follows that $U \equiv \mho(U)$.

Department of Philosophy, Utrecht University, Janskerkhof 13, 3512BL Utrecht, The Netherlands

E-mail address: a.visser@uu.nl

[5]We pronounce $\mho$ as 'mho' is such a way that it rhymes with 'joe'.