THE INTERPRETABILITY OF INCONSISTENCY
FEFERMAN'S THEOREM AND RELATED RESULTS

ALBERT VISSER

ABSTRACT. This paper is an exposition of Feferman's Theorem concerning the interpretability of inconsistency and of further insights directly connected to this result. Feferman's Theorem is a strengthening of the Second Incompleteness Theorem. It says, in metaphorical paraphrase, that it is not just the case that a theory fails to prove its own consistency, but that a theory actively holds its own inconsistency for possible. We first give a careful presentation of the result. Then, we provide two versions of the result that are both modal and infinitary. We explain how Feferman's Theorem is connected with two notions of *completion* of a theory. We provide an example of an application of the theorem. Finally, we discuss the failure of the result in a constructive setting.

Date: January 3, 2014.
2000 *Mathematics Subject Classification.* 03A05, 03B25, 03F25, 03F30, 03F45.
Key words and phrases. Interpretability, Second Incompleteness Theorem.
I am grateful to Ali Enayat, Joel Hamkins and Emil Jeřábek for inspiring discussions. Emil Jeřábek sent me a number of corrections and helpful comments for the penultimate version of the paper. Finally, I also thank Emil for his permission to use an argument of his that yields a more general version of Friedman's result that the Disjunction Property implies the Numerical Existence Property.

Contents

Dedicated to Sol Feferman, whose ideas always continued to inspire me.

1. Introduction

Feferman's Theorem is an intriguing result from Sol Feferman's fundamental paper *Arithmetization of metamathematics in a general setting* ([Fef60]). As a first approximation, the theorem says that, under certain conditions, a theory interprets itself plus its own inconsistency. In terms of models this tells us that there is a uniform construction (of a special kind) that yields, for every model of the given theory, an internal model of the theory that satisfies the formalized inconsistency statement of the theory. If, heuristically, we interpret the internal model relation as an epistemic accessibility relation, we could rephrase the theorem by saying: *every theory deems its own inconsistency possible.*

Feferman's Theorem is a strengthening of the Second Incompleteness Theorem. If a theory would prove its own consistency, it would interpret the conjunction of its own consistency statement and its inconsistency statement and, thus, be inconsistent.

Methodologically, Feferman's Theorem is interesting because it is a direct application both of Gödel's Completeness Theorem and of his Second Incompleteness Theorem, thus showing that these two central theorems can very well collaborate.

The present paper is a study of Feferman's Theorem. It is both an exposition of existing results and a presentation of new results.

In Section 3, I give a formulation of Feferman's Theorem in its full generality. I present various proofs of the theorem. In the case of finitely axiomatized theories, the theorem can be strengthened: one can show that, for sufficiently large n, a theory deems its own n-restricted inconsistency possible. Here n-restricted provability means that one only considers proofs where the complexity of the formulas occurring in the proof is below n. We provide a proof of this strengthening. The strengthening is to Feferman's Theorem as Pudlák's version of the Second Incompleteness Theorem for restricted provability is to the ordinary Second Incompleteness Theorem.

In Section 4, we extend Feferman's Theorem to modal and infinitary forms. One considers a *Big Kripke Model* (or: *Possibluum*) with all possible models for finite signature as nodes and with as accessibility relation the internal model relation. If we we heuristically view the modality as epistemic, we can formulate the result as follows. We show that not only does a theory deem its own inconsistency possible, but, what is more, the theory considers it possible that it is inescapably inconsistent. The modal versions give rise to infinitary versions of Feferman's Theorem as a matter of course. We also prove a modal version of Feferman's Theorem for restricted provability which is based on an infinitary version of Feferman's Theorem due to Jan Krajíček. One consequence of the existence of the modal version is a simple proof that the extension of first order predicate logic with the propositional modal logic of the internal model relation is more expressive that first order predicate logic alone.

In Section 5, we study completions of theories, i.e., systematic ways of extending a theory with sentences that are interpretable over it *in a non-arbitrary way.* We study three possible completions: the syntactic completion, the semantic completion and the intrinsic completion. The definitions of the the semantic completion and the intrinsic completion are adaptations of ideas of Emil Jeřábek developed in another context, to wit cut-interpretability. We will prove that the semantic and the intrinsic completion contain all inconsistencies of the infinitary versions of Feferman's Theorem. Thus, we show that inconsistency is highly non-arbitrary. If we did not already know that the inconsistency statements of familiar theories are false it would almost be an argument for adopting them as natural axioms ...

In most of the paper, Feferman's Theorem appears as a tool of understanding the peculiar place of (in)consistency statements in metamathematics. In Section 6, we illustrate that the theorem also has applications to unrelated matters. We show that the Π_3-conservativity of the negation of Σ_1-collection over Elementary Arithmetic is associated with a p-time transformation of proofs. The proof employs a miniaturization of the classical proof of Paris & Kirby ([PK78]). Undoubtedly there are many other ways of implementing such a miniaturization, so the claim is just that Feferman's Theorem 'comes in handy' to do the job, not that it is indispensable.

Finally, we explain, in Section 7, the fact —happy or sad, depending on your perspective– that Feferman's Theorem fails for constructive theories with the disjunction property (as long as we restrict ourselves to parameter-free interpretations). This fact can be proved as an immediate consequence of Harvey Friedman's celebrated theorem that the disjunction property implies the numerical existence property. To find an appropriate adaptation of Feferman's Theorem to the constructive context remains an open question.

Remark 1.1. This paper is intended as a presentation of a classical result and its *Umfeld*. Some parts of it, however, contain new material. Sections 2, 3, and Appendix A are expositions of previously published material. Section 7 is a presentation of Friedman's classical result that the disjunction property implies the numerical existence property. The section adds a few new elements. Specifically, we present some ideas due to Emil Jeřábek (in an unpublished note) to optimize the generality of the result. Section 4 is in part a presentation of known results, e.g. results of Jan Krajíček, but also contains new material, specifically the presentation of the material using modal notions is new. Sections 5, 6 and Appendix B are new. ❑

Remark 1.2. In a companion paper *Jumping in Arithmetic* I will discuss yet another aspect of Feferman's Theorem: the question whether it has a converse. ❑

2. Basic Facts & Definitions

In this section we fix a number of notations and conventions and we remind the reader of basic facts from the literature. In Appendix A, we give a more detailed exposition of some of the notions involved. The reader is advised to go through this section lightly in order to return when some fact or definition is used.

2.1. **Theories and Provability.** Theories in this paper have finite signature. The signature is supposed to be part of the data of the theory. Usually we also take it that a formula representing the axiom set is also part of the data. This is relevant when we consider e.g. the formalization of provability in the given theory. However we will consider some theories that are not recursively enumerable and for these there often is no obvious formula representing the axiom set. We will be slightly sloppy about these things and what is intended will be clear from the context.

The signatures of our theories will be *officially* relational, but we will often treat them as if they also have function symbols. The p-time term-unraveling algorithm guarantees that this confusion is harmless.

A finitely axiomatized theory will be a theory where the axiom set is explicitly given by a disjunction of the form $\bigvee_{i<n} x = \underline{\ulcorner B_i \urcorner}$. Consider the theory that has as axioms the Peano axioms that are larger than the smallest inconsistency proof of Peano Arithmetic. This theory has in fact finitely many axioms, but we will not count it as finitely axiomatized. *Par abus de langage*, we will use A, B, ... to designate a finitely axiomatized theories, thus confusing a sentence axiomatizing a theory with a theory. One disadvantage is that sometimes it is really relevant that we can read off the signature from the theory. The big advantage is that it is immediately clear from the notation that we are looking at something that is finitely axiomatized.

We will use modal notations for arithmetized provability and consistency. E.g., we use $\Box_U A$ for $\mathsf{prov}_U(\underline{\ulcorner A \urcorner})$ and $\Diamond_U \top$ for $\mathsf{con}(U)$. We will also consider *restricted provability*: a sentence is n-provable iff it is provable from axioms with Gödelnumbers below n, where the formulas in the proof have complexity less than n. The notion of complexity we use is *depth of quantifier changes*. [1] We will use $\rho(A)$ for the complexity of A. We write $\Box_{U,x} A$ for the arithmetization of restricted provability.

Some special theories that we will use is Buss' system S^1_2 (see [Bus86] or [HP93]) and Elementary Arithmetic EA, also known as Elementary Function Arithemetic EFA and as $\mathrm{I}\Delta_0 + \mathsf{Exp}$ (see [HP93]).

2.2. **Translations, Interpretations & Interpretability.** We first explain the idea of *a translation* τ between signatures Σ and Θ. More details on translations are given in Appendix A.1 and Appendix A.3. The translation τ sends the predicates of Σ to formulas of the language based on Θ where we represent the argument places by designated variables. Moreover, the translation τ provides a domain formula δ_τ. We may also consider k-dimensional translations. In this case an argument place of a Σ-predicate is represented by a sequence of designated variables of length k. In addition we may allow parameters in our translation: these are variables in the translations of the predicate symbols that do not correspond to an argument place of the translated predicate symbol. We do *not* demand that the identity relation is translated by itself. The translation commutes with the propositional connectives. It also commutes with the quantifiers but it adds a relativization to the domain: e.g., in the 1-dimensional case, $\forall x\, A$ translates to $\forall x\, (\delta_\tau(x) \to A^\tau(x))$.

An interpretation K relates two theories U and V. These theories are part of the data for K. The interpretations provides a translation τ_K. We demand that, for all U-sentences A, if $U \vdash A$, then $V \vdash A^{\tau_K}$. We write $K : U \to V$ or $K : U \lhd V$ or $K : V \rhd U$. The notation $K : U \to V$ is useful when we want to think of

[1] See Appendix A.4 for more information on this complexity measure.

theories and interpretations forming a category. The notations $K : U \lhd V$ and $K : V \rhd U$ function in a context where we think of interpretability as a generalization of provability.

We write $K : U \lhd_{\mathsf{faith}} V$ for: K is *a faithful interpretation* of U in V. This means that: for all U-sentences A, we have: $U \vdash A$ iff $V \vdash A^{\tau_K}$.

A translation τ maps a model $\mathcal{M}$ to an internal model $\widetilde{\tau}(\mathcal{M})$ provided that the $\mathcal{M} \models \exists x\, \delta_\tau(x)$. Thus an interpretation $K : U \to V$ gives us a mapping $\widetilde{K}$ from $\mathsf{MOD}(V)$, the class of models of V to $\mathsf{MOD}(U)$ the class of models of U. If we build a category of theories and interpretations, usually MOD with $\mathsf{MOD}(K) := \widetilde{K}$ will be a contravariant functor.

We have a number of operations of translations and interpretations. First every signature has an identity translation. This induces for every theory an identity interpretation. Secondly, translations and interpretations can be composed in the obvious way. Thirdly we can transform two translations / interpretations into a disjunctive interpretation: given that we have τ_0 and τ_1, we can form the translation $\tau_0\langle A\rangle\tau_1$ that behaves like τ_0 when A and like τ_1 when $\neg A$. Clearly disjunctive translations induce disjunctive interpretations. Uses of disjunctive interpretations will be everywhere dense in this paper.[2]

To make interpretations into a category we need a notion of sameness between interpretations. There are a number of possible choices for what sameness is. We mention four of them. Suppose $K, K' : U \to V$.

a. K is equal$_0$ to K' if V proves that K and K' are extensionally equal, i.e. $V \vdash \forall x\, (\delta_K(x) \leftrightarrow \delta_{K'}(x))$ and $V \vdash \forall \vec{x} \in \delta_\tau\, (P_K\vec{x} \leftrightarrow P_{K'}(\vec{x}))$.

b. K is equal$_1$ to K' if there is a V-definable V-verifiable isomorphism F between K and K'. Equivalently: K is equal$_1$ to K' if , in every V-model $\mathcal{M}$ there is a definable isomorphism between $\widetilde{K}(\mathcal{M})$ and $\widetilde{K}'(\mathcal{M})$. The equivalence between these definitions uses a compactness argument and disjunctive interpretations.[3]

c. K is equal$_2$ to K' if, for every V-model $\mathcal{M}$, we have that $\widetilde{K}(\mathcal{M})$ and $\widetilde{K}'(\mathcal{M})$ are isomorphic.

d. K is equal$_3$ to K' if, for all V-sentences A, $V \vdash A^K \leftrightarrow A^{K'}$. Equivalently: K is equal$_3$ to K' if, for every V-model $\mathcal{M}$, we have that $\widetilde{K}(\mathcal{M})$ and $\widetilde{K}'(\mathcal{M})$ are elementary equivalent. Equivalently: K is equal$_3$ if, for every countable recursively saturated V-model $\mathcal{M}$, we have that $\widetilde{K}(\mathcal{M})$ and $\widetilde{K}'(\mathcal{M})$ are isomorphic.

Each of the notions of equality gives us a different category. Each category in its turn delivers a different notion of isomorphism between theories. Two theories are *definitionally equivalent* or *synonymous* if they are isomorphic in the category of *equal*$_0$. They are *bi-interpretable* if they are isomorphic in the category of *equal*$_1$. Two theories are *iso-congruent* if they are isomorphic in the category of *equal*$_2$. They are *sententially congruent* if they are isomorphic in the category of *equal*$_3$.

---

[2]See Appendix A.1 for more explicit definitions of operations on translations and interpretations.

[3]See Appendices A.2 en A.3 for more details on definable isomorphisms.

We will consider a number of reduction relations between theories based on interpretations.

- We write $U \lhd V$ for: there is a K such that $K : U \lhd V$. We pronounce this as: U is interpretable in V. We write $V \rhd U$ for $U \lhd V$. We pronounce this as: V interprets U. We use $U \equiv V$ for: $U \lhd V$ and $V \lhd U$. So $\equiv$ is the induced equivalence relation of $\lhd$. In this case we say that U and V are *mutually interpretable.*
- We write $U \blacktriangleleft V$ or $V \blacktriangleright U$ for: every V-model has an internal U model. We pronounce this as: U is model-interpretable in V or V model-interprets U. We use $\equiv_{\mathsf{mod}}$ for the induced equivalence relation.
- We write $U \lhd_{\mathsf{loc}} V$ or $V \rhd_{\mathsf{loc}} U$ for: for all finite subtheories U_0 of U, $U_0 \lhd V$. We pronounce this as: U is locally interpretable in V or V locally interprets U. We use $\equiv_{\mathsf{loc}}$ for the induced equivalence relation.
- Suppose the theory W extends U. Then, V *locally interprets* W *over* U, or $V \rhd_{(U,\mathsf{loc})} W$, iff, for all finite subtheories W_0 of W, $V \rhd (U + W_0)$. We use $\equiv_{(U,\mathsf{loc})}$ for the induced equivalence relation.

We sometimes write e.g. $A \rhd_U B$ for $(U + A) \rhd (U + B)$. For finitely axiomatized A we have: $U \rhd A$ iff $U \blacktriangleright A$ iff $U \rhd_{\mathsf{loc}} A$. If follows that:

$$V \rhd U \Rightarrow V \blacktriangleright U \text{ and } V \blacktriangleright U \Rightarrow V \rhd_{\mathsf{loc}} U.$$

In this paper, we present examples that illustrate that neither of these arrows can be reversed.

In this paper we will mainly look at interpretations from the standpoint of kinds of interpretability and not so much from the standpoint of categories that are not just preorders. For this reason, we will be somewhat sloppy w.r.t. the translation / interpretation distinction and w.r.t. the strict regime of source and target that we officially have for interpretations.

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

Theorem 2.1. *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.

We end this subsection with a useful theorem in the style of the Friedman-Goldfarb-Harrington Theorem.

Theorem 2.2. *Consider any finitely axiomatized theory A and suppose that $N : \mathsf{S}^1_2 \lhd A$. Consider any Σ_1-formula $S(x)$. Then, we can effectively obtain a Σ_1-formula $R(x)$, such that:*

a. $\mathsf{EA} \vdash \forall x\,((A \rhd (A + R^N(x))) \leftrightarrow (S(x) \vee \Box_A \bot))$.

b. $\mathsf{EA} + \Diamond_A \top \vdash \forall x\,(R(x) \leftrightarrow S(x))$.

Proof. By the Gödel Fixed Point Lemma we can find R such that:

$$\mathsf{S}^1_2 \vdash R(x) \leftrightarrow S(x) \leq (A \rhd (A + R^N(x))).$$

We will suppress the parameter x in the reasoning since it just rides along for free. We prove (a). We reason in EA.

From left to right. Suppose $A \rhd (A + R^N)$. Then R or $R^\perp$. In the first case we have S. In the second case, by Σ_1-completeness, $A \rhd (A + R^N + R^{\perp N})$. Hence $A \rhd \perp$ and so $\Box_A \perp$.

From right to left. If we have $\Box_A \perp$ we are immediately done. Suppose S. It follows that R or $R^\perp$. In the first case we have $A \rhd (A + R^N)$, by Σ_1-completeness. In the second case, we have $A \rhd (A + R^N)$, since $R^\perp$ is $(A \rhd (A + R^N)) < S$.

The proof of (b) is left to the reader. ❑

2.3. **The Modal Logic of Internality.** We define the modal language as follows. For any signature Θ we have:

- $\phi_\Theta ::= A_\Theta \mid \neg\, \phi_\Theta \mid (\phi_\Theta \wedge \phi_\Theta) \mid (\phi_\Theta \vee \phi_\Theta) \mid (\phi_\Theta \to \phi_\Theta) \mid \blacksquare_U \phi_{\Sigma_U}$.

Here A_Θ ranges over formulas of predicate logic for signature Θ and U ranges over recursively enumerable theories of ordinary predicate logic, where Σ_U is the signature of U. We use $A, B, \ldots$ for predicate logical formulas and ϕ, ψ for mixed predicate logical and modal formulas. We use $\Gamma, \Delta, \ldots$ for sets of modal formulas. The operator $\blacklozenge_U \phi$ is defined as $\neg \blacksquare_U \neg\, \phi$. The operator $\blacksquare$ is *internal necessity* and $\blacklozenge$ is *internal possibility*. Note that there is no quantifying into modal formulas.

The big Kripke model $\mathbb{K}$ has as nodes all models of finite signature. For any recursively enumerable theory U we have an accessibility relation R satisfying: $\mathcal{M}\ R_U\ \mathcal{K}$ iff $\mathcal{K} \models U$ and $\mathcal{M} \rhd \mathcal{K}$. Here we assume that U is given with a signature Σ and $\mathcal{K}$ has signature Σ.

We define truth-at-a-node and validity.

- Truth at a note is define in the obvious way for the atoms and the truth functional connectives.
- $\mathcal{M} \models \blacksquare_U \phi$ iff, for all $\mathcal{K}$ such that $\mathcal{M}\ R_U\ \mathcal{K}$, we have $\mathcal{K} \models \phi$.
- $\Gamma \models_\Theta \phi$ iff, for all models $\mathcal{M}$ of signature Θ, if $\mathcal{M} \models \Gamma$, then $\mathcal{M} \models \phi$. Here we assume that Γ, ϕ consists of modal sentences of signature Θ. We will often suppress the subscript for the signature.

Note that R_U is reflexive on models of U and that the composition of R_U and R_V is contained in R_V.

We can define model interpretability in terms of the modal logic:

$$V \blacktriangleright U \text{ iff } V \models \blacklozenge_U \top.$$

Remark 2.3. In the present paper we will essentially need modalities corresponding to infinitely axiomatized theories. If we restrict ourselves to finitely axiomatized theories we can simplify the set-up by just having $\blacksquare_\Sigma$ where Σ is a signature, since $\blacksquare_A B$ is equivalent to $\blacksquare_\Sigma(A \to B)$. ❏

Remark 2.4. One can obtain many alternative Big Kripke Models (or Possiblua) by varying the accessibility relation and/or restricting the domain of first order models. Here are some interesting examples.

a. We can restrict ourselves to models of a basic arithmetical theory that is preserved to definable cuts. We take the definable cut relation as accessibility relation.

b. We can restrict ourselves to models of PA with as accessibility relation: is an internal model such with a definable satisfaction predicate such that all axioms of PA are internally true. This structure is studied by Paula Henk in a forthcoming paper. The modal logic of this Big Model is precisely Löb's Logic. It is unknown what happens if e.g. we consider analogues of this idea for finitely axiomatized sequential theories.

c. We can consider models of ZF and the relation: is an internal (parametrically definable) transitive model of ZF. The modal logic of this was characterized by Robert Solovay. See [Sol76]. A detailed exposition is given in [Boo93].

d. We can consider models of ZF and the relation: is an internal (parametrically definable) universe of ZF. The modal logic of this was characterized by Robert Solovay. See [Sol76]. A detailed exposition is given in [Boo93].

e. We can consider models of ZF and consider the relation: is a set forcing extension. The modal logic of this was characterized by Joel Hamkins and Benedikt Löwe. See [HL08].

❏

In this paper we will not study the modal logic of internality. It will rather serve as a language that provides memorable formulations of some results. Two results will be spin-off of versions of Feferman's Theorem. The valid principles involving the box are Π_2-hard and modal definability is stronger than first order definability. Both results use gray boxes for non-finite recursively enumerable theories, so it is open whether we get the same results when we only allow $\blacktriangleright_A$ for A finitely axiomatized. We present one characterization theorem in Appendix B.

2.4. **A Basic Concept.** The modal notions discussed in Subsection 2.3 have a syntactic shadow. In this subsection, we introduce this shadow, to wit the operation $[A]_{U,V}$.

- $[A]_{U,V} := \{A^K \mid K : V \rhd U\}$.
- $[A]_U := [A]_{U,U}$.

We note that, if U and V are recursively enumerable theories then $[A]_{U,V}$ is *prima facie* Σ_3. If U is finitely axiomatized, then $[A]_{U,V}$ is Σ_1. We give two basic insights, also for later reference, concerning $[A]_{U,V}$. The first result is a triviality but very useful.

Theorem 2.5. *Suppose $K : Z \lhd W$ and $M : U \lhd V$. Consider any Z-sentence A. Then, $M^* : (U + [A^K]_{W,U}) \lhd (V + [A]_{Z,V})$. Here M^* is based on the same translation as M.*

We note two salient special cases.

a. If we take $Z := W$ and $K := \mathsf{ID}_W$, then: $M^* : (U + [A]_{W,U}) \lhd (V + [A]_{W,V})$.

b. If we take $U := V$ and $M := \mathsf{ID}_V$, then: $V + [A^K]_{W,V} \subseteq V + [A]_{Z,V}$. If we, in the last case, specialize K to the identical embedding, so that Z is a subtheory of W in the same language, or $Z \subseteq W$, we get: $V + [A]_{W,V} \subseteq V + [A]_{Z,V}$.

So, we have monotonicity in the U-component w.r.t. $\lhd$ and anti-monotonicity in the W-component w.r.t. $\subseteq$.

Proof. For any $L : W \lhd U$, we have $V + [A]_{Z,V} \vdash A^{KLM}$. So it is immediate that $V \vdash (U + \{(A^K)^L \mid L : W \lhd U\})^M$. ❑

Theorem 2.6. *Suppose A is finitely axiomatized. Suppose $U \rhd A$.*

i. Suppose U' is an extension in the same language as U. Then, we have $(U' + [B]_{A,U}) \vdash [B]_{A,U'}$.

ii. Suppose $\mathcal{M} \models U$. Then, $\mathcal{M} \models [B]_{A,U}$ iff $\mathcal{M} \models \blacksquare_A B$.

Proof. We prove (i). Suppose $K : A \lhd U$ and $K' : A \lhd U'$. Then $L := K'\langle A^{K'}\rangle K : A \lhd U$. So $U + [B]_{A,U} \vdash B^L$. It follows that $U' + [B]_{A,U} \vdash B^L$ and, since $U' \vdash A^{K'}$, we have $U' + [B]_{A,U} \vdash B^{K'}$.

The proof of (ii) is similar. ❑

2.5. **Sequential Theories.** The notion of *sequential theory* is an explication of *theory with coding*. Specificaly, a sequential theory provides an interpretation N of 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 a very simple definition discovered by Pavel Pudlák. We first need the definition of Adjunctive Set Theory or AS is a one-sorted theory with a binary relation $\in$.

$\mathsf{AS1}$ $\vdash \exists x \forall y\, y \notin x$,

$\mathsf{AS2}$ $\vdash \forall x, y\, \exists z\, \forall u\, (u \in z \leftrightarrow (u \in x \vee u = y))$.

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 [Vis13a].[4]

We collect a number of basic facts about sequential theories.

In sequential theories, number systems are partly comparable: they share modulo definable isomorphism a definable cut.

Theorem 2.7. *Suppose U is a sequential theory and $N, N' : \mathsf{S}^1_2 \lhd U$. Then there are definable cuts I, I' of N, respectively N' such that there is an U-definable, U-verifiable isomorphism between I and I'.*

Theorem 2.7 is due to Pavel Pudlák. See [Pud85] or [HP93].

A finitely axiomatized sequential theory is mutually interpretable with its own restricted consistency over S^1_2.

Theorem 2.8. *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 consistency. (If this were not the case, we would obtain a contradiction with the Second Incompleteness Theorem.)

We provide an partial analogue of Theorem 2.8 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 2.9. *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 $U \equiv_{\mathsf{loc}} \mho(U)$.

Inspection of the interpretation of U in $\mho(U)$ shows that it can be given a truth-predicate inside $\mho(U)$ for an internally definable language that is downwards closed under taking subformulas and that contains all standard formulas. We do not get, on pain of contradicting the Second Incompleteness Theorem, the truth predicate for a language that is upward closed under the formation rules like forming conjunctions.

[4]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.

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

There is an important connection between interpretability between Π_1-sentences over S^1_2 and provability between Π_1-sentences over EA.

Theorem 2.10. *For any Π^0_1-sentences P, P', we have:*

$$(\mathsf{S}^1_2 + P) \rhd (\mathsf{S}^1_2 + P') \quad \Leftrightarrow \quad \mathsf{EA} \vdash P \to P'.$$

This result is due to Wilkie and Paris. See [WP87]. For a generalization, see: [Vis92].[6]

The following FGH-style result is a variant and refinement of a sequence of FGH theorems proved in [Vis93], [Vis05] and [Vis12a]. This work is in its turn based on ideas and results of Jan Krajíček (see [Kra87]) and Harvey Friedman (see [Smo85]). Krajíček's work is based on results from Alex Wilkie's fundamental paper [Wil86]. Theorem 2.11 is Theorem 10 of [Vis13b].

Theorem 2.11. *Let A be a finitely axiomatized sequential theory. Let k be any number. We can find an interpretation $N_0 : \mathsf{S}^1_2 \lhd A$, such that, for every Σ_1-sentence S with $\rho(S) \leq k$:*

$$\mathsf{EA} \vdash \Box_{A,m} S^{N_0} \leftrightarrow (S \vee \Box_{A,\rho(A)}\bot).$$

Here $m := \mathsf{max}(\rho(A), k + \rho(N_0))$.

We will use the following application of Theorem 2.11. Let A be a finitely axiomatized sequential theory. We note that for some fixed k_0 and for all ℓ we have: $\rho(\Box_{A,\ell}\bot) = k_0$. Substituting $\Box_{A,\ell}\bot$ for S in the statement of Theorem 2.11, we find: there is an interpretation $N_0 : \mathsf{S}^1_2 \lhd A$, such that, for every ℓ:

$$\mathsf{EA} \vdash \Box_{A,m}\Box^{N_0}_{A,\ell}\bot \leftrightarrow (\Box_{A,\ell}\bot \vee \Box_{A,\rho(A)}\bot).$$

Here $m := \mathsf{max}(\rho(A), k + \rho(N_0))$. We note that $\mathsf{EA} \vdash \Box_{A,\ell}\bot \to \Box_{A,\rho(A)}\bot$, since cutelimination for a standard complexity is multi-exponential. It follows that

$$\mathsf{EA} \vdash \Box_{A,m}\Box^{N_0}_{A,\ell}\bot \leftrightarrow \Box_{A,\rho(A)}\bot.$$

From this we have:

$$\mathsf{EA} \vdash \Diamond_{A,m}\Diamond^{N_0}_{A,\ell}\top \leftrightarrow \Diamond_{A,\rho(A)}\top.$$

Ergo, by the Theorems 2.8 and 2.10:

$$A \equiv (\mathsf{S}^1_2 + \Diamond_{A,\rho(A)}\top) \equiv (\mathsf{S}^1_2 + \Diamond_{A,m}\Diamond^{N_0}_{A,\ell}\top) \equiv (A + \Diamond^{N_0}_{A,\ell}\top).$$

Thus, we find:

Theorem 2.12. *Suppose A is a finitely axiomatized sequential theory. Then there is an $N_0 : \mathsf{S}^1_2 \lhd A$ such that, for every ℓ, $A \rhd (A + \Diamond^{N_0}_{A,\ell}\top)$.*

We end with a theorem that is closely related to Theorem 2.11. A theory U is *trustworthy* iff, for all recursively enumerable theories V with $U \rhd V$, we have $U \rhd_{\mathsf{faith}} V$. Harvey Friedman proved that consistent, finitely axiomatized, sequential theories are trustworthy. See [Smo85]. Corollary 5.9 of [Vis05] gives us the following minor but useful strengthening of Friedman's result.

[6] We find the theorem also formulated with Q, PA^- and $\mathrm{I}\Delta_0 + \Omega_1$ in the role of S^1_2. It is easy to see that all these versions are equivalent.

Theorem 2.13. *Suppose A is consistent, finitely axiomatized and sequential. Suppose U is an recursively enumerable theory and U is mutually interpretable with A. Then U is trustworthy.*

3. Proofs of Feferman's Theorem

We present various proofs of Feferman's Theorem. In Subsection 3.1 we present a version of Feferman's own proof. In Subsection 3.2, we adapt a proof strategy due to Kreisel to prove Feferman's Theorem. In Subsection 3.3, we present the simplest known proof of Feferman's Theorem. We prove a variant of the Theorem for restricted provability in Subsection 3.4.

We remind the reader of the full statement of the theorem. Some of the proofs below have less scope.

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 $U + \Box^N_U \bot$ in U.*

3.1. Feferman's Proof.

We work over a theory U which is reflexive with respect to an interpretation $N : \mathsf{S}^1_2 \lhd U$. The Feferman predicate $\Box^*$ is defined by:

$$\Box^*_U A :\leftrightarrow \exists x\, (\Box_{U,x} A \wedge \Diamond_{U,x} \top).$$

We note that we have:

- $U \vdash A \Rightarrow U \vdash \Box^*_U A$ (this uses reflexivity),
- $U \vdash (\Box^*_U (A \to B) \wedge \Box^*_U A) \to \Box^*_U B$,
- $U \vdash \Diamond^{*N}_U \top$,
- $U \vdash \Box^{*N}_U A \to \Box^N_U A$
- $U \vdash S^N \to \Box^{*N}_U S^N$, for $S \in \exists\Sigma^{\mathsf{b}}_1$.

Let G be the ordinary Gödel sentence for U, so $U \vdash G^N \leftrightarrow \neg\Box^N_U G^N$. Here is Feferman's original proof:

$$\begin{array}{lll} \Box^{*N}_U G^N & \vdash_U & \Box^{*N}_U G^N \wedge \Box^N_U G^N \\ & \vdash_U & \Box^{*N}_U (G^N \wedge \Box^N_U G^N) \\ & \vdash_U & \Box^{*N}_U \bot \\ & \vdash_U & \bot \end{array}$$

It follows that $\vdash_U \Diamond^{*N}_U \neg G^N$, and hence $\vdash_U \Diamond^{*N}_U \Box^N_U \bot$. We may conclude, by the Gödel-Hilbert-Bernays-Wang-Henkin-Feferman Theorem (Theorem 2.1), that $\top \rhd_U \Box^N_U \bot$.

A slight variant of the proof is to eliminate the Gödel sentence in favor of the consistency statement:

$$\begin{aligned}\Box^{*N}_U \Diamond^N_U \top &\vdash_U \Box^{*N}_U \Diamond^N_U \top \wedge \Box^N_U \Diamond^N_U \top \\ &\vdash_U \Box^{*N}_U \Diamond^N_U \top \wedge \Box^N_U \bot \\ &\vdash_U \Box^{*N}_U (\Diamond^N_U \top \wedge \Box^N_U \bot) \\ &\vdash_U \Box^{*N}_U \bot \\ &\vdash_U \bot\end{aligned}$$

It follows that $\vdash_U \Diamond^{*N}_U \Box^N_U \bot$. Hence, by the Gödel-Hilbert-Bernays-Wang-Henkin-Feferman Theorem (Theorem 2.1), that $\top \rhd_U \Box^N_U \bot$.

A third variant is to prove that $\vdash_U \Diamond^N_{U,n} \Box^N_U \bot$, for each n and to apply the Orey-Hájek Characterization. We note that this last strategy still needs the Feferman predicate or some related device to prove the Orey-Hájek Characterization.

We note that the Feferman proof works for reflexive theories like PRA, PA and ZF. It still works for theories that are just *sententially* reflexive like $\mathrm{I}\Pi_1^-$, the theory of parameter-free Π_1-induction and the curious theory $\mathsf{PA}^{\mathsf{cor}}$ (see [Vis12b]).

3.2. **A Kreiselian Proof.** Kreisel's entertaining alternative proof of the Second Incompleteness Theorem is reported in [Smo77]. What has not been noted before is that it 'really' is a proof of Feferman's Theorem. Surprisingly, this approach gives Feferman's Theorem in full generality.

Consider $N : \mathsf{S}^1_2 \lhd U$. We assume that U is Δ^{b}_1-axiomatized. The formalized Henkin construction gives us $H : (U + \Diamond^N_U \top) \rhd U$ —this is the Gödel-Hilbert-Bernays-Wang-Henkin-Feferman Theorem (Theorem 2.1).

Let T be the truthpredicate associated with H. We note that T 'works' for sentences on some definable cut J of N. We find a sentence L such that $U \vdash L \leftrightarrow \neg T(L)$. Suppose S is $\exists\Sigma^{\mathsf{b}}_1$. We have $U \vdash S^N \to \Box^N_U S^N$. Hence,

$$(\dagger) \quad H : (U + S^N + \Diamond^N_U \top) \rhd (U + S^N).$$

The construction of T consists of finding a J-path through a binary tree. At each node a *yes-no* choice concerning the consistency of a finite extension of U is made. The *no* decision is $\exists\Sigma^{\mathsf{b}}_1$ in N.

Now suppose we have, in U, $\Diamond^N_U \top$. In this case we may apply H. If, inside H, we have again $\Diamond^N_U \top$, we can repeat this to form H^2. Etc. Since, by (†), the $\exists\Sigma^{\mathsf{b}}_1$-sentences are inward preserved if we iterate H, the path will move to the right. Since the value of L alternates when we iterate H, the path moves in each iteration of H strictly to the right. Since the breadth of the tree at depth L is approximately $n := 2^\ell$, where ℓ is the Gödel number of L, this can happen at most n times. This means that $(H\langle \Diamond^N_U \top \rangle \mathsf{id})^n : U \rhd (U + \Box^N_U \bot)$.

We note that in case we give the proof for e.g. PA, we need not use (†). The fact that Σ^{b}_1-sentences are inward preserved follows from the fact that the internal model construction yields *strict end-extensions*. This, of course, fails in the general case.

The nice feature of the present proof is that it does not presuppose the Second Incompleteness Theorem. On the other hands it uses the same ingredients: $\exists\Sigma_1^{\mathsf{b}}$-completeness and self-reference. (In the case of PA the use of $\exists\Sigma_1^{\mathsf{b}}$-completeness is replaced by upwards preservation of Σ_1-sentences to end-extensions.) The disadvantage is that the witnessing interpretation is rather large.

3.3. **A Simple Proof.** A very simple proof of Feferman's Theorem 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.

Consider any theory U with p-time decidable axiom set and an interpretation $N : \mathsf{S}^1_2 \lhd U$. Clearly, we have $\Diamond^N_U \top \vdash_U \Diamond^N_U \Box^N_U \bot$ and $\Diamond^N_U \Box^N \bot \rhd_U \Box^N_U \bot$, by, respectively, the Second Incompleteness Theorem and the Gödel-Hilbert-Bernays-Wang-Henkin-Feferman Theorem (Theorem 2.1). By composition, $\Diamond^N_U \top \rhd_U \Box^N_U \bot$. Suppose K witnesses that $\Diamond^N_U \top \rhd_U \Box^N_U \bot$. We also have $\mathsf{ID} : \Box^N_U \bot \rhd_U \Box^N_U \bot$. Hence $K\langle \Diamond^N_U \top \rangle \mathsf{ID} : \top \rhd_U \Box^N_U \bot$.

3.4. **Feferman's Theorem for Restricted Provability.** Consider a finitely axiomatized theory A.[7] If we say finitely axiomatized, we mean axiomatized by a formula of the form $x = \underline{\ulcorner A \urcorner}$ or, perhaps, $\bigvee_{i<n} x_i = \underline{\ulcorner A_i \urcorner}$. So nothing like "$x$ is an axiom if $x = \underline{\ulcorner A \urcorner}$ or there is an inconsistency proof of PA below x and x is / codes a Peano axiom."

For finitely axiomatized theories we have Löb's Theorem for restricted provability. This is in essence due to Pudlák [Pud85]. (Pudlák stated the theorem as a form of the Second Incompleteness Theorem, but the fact that Löb follows from the Second Incompleteness Theorem is well known.) For completeness we give the proof. This is just the usual proof where one convinces oneself that one never exceeds the bounds of the chosen restriction for restricted provability. We define $\rho(N)$ as the maximum of the $\rho(P^N\vec{x})$ where P is a relation symbol of a relational version of arithmetic. We note that, for an arithmetical formula, $\rho(A^N)$ is estimated by $\rho(A) + \rho(N) + 1$. We first unravel the terms in a small scope way. This adds 1 to the alternating quantifier depth because we add blocks of existential quantifiers. Then we replace all relations symbols by the corresponding formulas which adds $\rho(N)$.

Theorem 3.1. *We have:*

i. S^1_2 *verifies the following. Suppose* $N : \mathsf{S}^1_2 \lhd A$. *Let* k *be sufficiently large. (In the proof, we discuss what this means.) Then, for any sentence* B *in the language of* A *and for any* $n \geq \mathsf{max}(\rho(B), k)$, *we have: if* $A \vdash_n \Box^N_{A,n} B \to B$, *then* $A \vdash_n B$.

ii. $\mathrm{I}\Delta_0 + \mathsf{supexp}$ *verifies the following. Suppose* $N : \mathsf{S}^1_2 \lhd A$. *Let* k *be sufficiently large. Then, for any sentence* B *in the language of* A *and for any* $n \geq \mathsf{max}(\rho(B), k)$, *we have: if* $A \vdash \Box^N_{A,n} B \to B$, *then* $A \vdash B$.

[7] *Par abus de langage*, the formula-variable 'A' is used to suggest a finitely axiomatized theory. Of course, a finitely axiomatized theory is really a different *kind* of thing than a sentence.

Proof. We note that (ii) is immediate from (i), since we have cut-elimination in $\mathrm{I}\Delta_0 + \mathsf{supexp}$. We treat (i). We work in S^1_2.

Suppose $N : \mathsf{S}^1_2 \lhd A$ and $A \vdash_n \Box^N_{A,n} B \to B$. We note that for this to make sense n must exceed $\rho(B)$, $\rho(A)$, and $\rho(\Box^N_{A,n} B)$, where:

$$\rho(\Box^N_{A,n} B) = \rho(N) + \rho(\mathsf{prov}'(u, v, w)) + 1,$$

since A, n and B only occur as numerals. Here $\mathsf{prov}'(u, v, w)$ is a codification of bounded provability in a finite theory where u represents the theory (in a canonical way), v represents the bound and w represents the conclusion. Clearly, $\rho(\mathsf{prov}(u, v, w))$ is a standard number. We also assume that $n > \rho(p)$, where p is an A-proof of $(\bigwedge \mathsf{S}^1_2)^N$.

We have:

$$(\dagger)\ \mathsf{S}^1_2 \vdash \forall D, E\, \forall x \geq \mathsf{max}(\rho(D), \rho(E), \rho(N) + \rho(\mathsf{prov}'(u, v, w)) + 1) \\ ((\Box_{A,x} D \wedge \Box_{A,x}(D \to E)) \to \Box_{A,x} E).$$

This needs a standard proof, say q. We can also prove:

$$(\ddagger)\ \mathsf{S}^1_2 \vdash \forall D\, \forall x \geq \mathsf{max}(\rho(D), \rho(N) + \rho(\mathsf{prov}'(u, v, w)) + 1)\, (\Box_{A,x} D \to \Box_{A,x}\Box_{A,x} D).$$

This needs a standard proof, say r. We take $n > \mathsf{max}(\rho(q), \rho(r)) + \rho(N)$.

By the Gödel Fixed Point Lemma, we can find a C such that:

$$A \vdash_n C \leftrightarrow (\Box^N_{A,n} C \to B).$$

We note that the proof of the Fixed Point Lemma contains a formula of the form $\mathsf{prov}^N_{A,n}(\mathsf{subst}^N(\underline{m}, \underline{m}))$. The proof is very roughly the computation showing that $\mathsf{subst}(\underline{m}, \underline{m}) = \underline{\ulcorner C \urcorner}$. This amounts to showing that a certain sequence of numbers given as numerals has a desired property. The length of the computation and the numerals occurring in it may be non-standard, but the complexity of the formulas occurring in it clearly has some standard bound not much exceeding $\rho(\mathsf{subst}(v, v)) + \rho(N)$.

Now we reason as follows. Suppose $A \vdash_n \Box^N_{A,n} B \to B$. We reason in A. Suppose $\Box^N_{A,n} C$. Then $\Box^N_{A,n}\Box^N_{A,n} C$, by instantiation of ($\ddagger$). Moreover, by the choice of C and instantiation of ($\dagger$): $\Box^N_{A,n} B$. By our assumption it follows that B. Hence $\Box^N_{A,n} C \to B$, i.e. C.

So we find $A \vdash_n (\Box^N_{A,n} C \to B)$ and $A \vdash_n C$. It follows that $A \vdash_n \Box^N_{A,n} C$ and hence that $A \vdash_n B$. ❑

Theorem 3.2. *Suppose $N : \mathsf{S}^1_2 \lhd A$. Then, we can effectively find a k, such that $A \rhd (A + \Box^N_{A,k} \bot)$.*

Proof. Suppose $N : \mathsf{S}^1_2 \lhd A$. We take k large enough w.r.t. $\rho(N)$ and $\rho(A)$ and $\rho(p)$ where p verifies $N : \mathsf{S}^1_2 \lhd A$. We reason in A. In case we have $\Box^N_{A,k}\bot$, we take the identical interpretation. Otherwise we have $\Diamond^N_{A,k}\top$. By Löb's Theorem, we have $\Diamond^N_{A,k}\Box^N_{A,k}\bot$. From this consistency statement we can build a Henkin interpretation H of $A + \Box^N_{A,k}\bot$. So we take this H. Thus, the disjunctive interpretation $\mathsf{ID}_A\langle\Box^N_{A,k}\bot\rangle H$ does the trick. ❑

Open Question 3.3. Is it possible to prove Theorem 3.2 using a method analogous to the Kreisel-style proof of Feferman's Theorem? ❏

4. Modal and Infinitary Versions of Feferman's Theorem

This section is devoted to modal versions of Feferman's Theorem. Viewed in a different way, these versions are infinitary in the sense that they involve infinitely may inconsistency statements.

Let's briefly look at Feferman's Theorem again:

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 $U + \Box^N_U \bot$ in U.*

A moment's reflection suggests that, unless we substantially enrich the modal framework, there is nog good modal version that reflects what this says. As we will see below, we can formulate and prove a modal version that is in many respects stronger than the original version. It is weaker in that we have to replace interpretability by model interpretability. In other words, the cost is uniformity.

We can also give a modal version for the case of restricted provability. This version is from a technical point of view more interesting than the ordinary one. For example using it we show that definability in the logic of internality is not first-order. In the next section we will see that it also follows that the valid principles of the logic of internality are at least Π^0_2.

4.1. Inconsistency is Possibly Necessary.

Before proving the promised modal result, we first prove an infinitary version of Feferman's Theorem. We remind the reader that $U \rhd_{(W,\mathsf{loc})} V$ iff U and V are extensions of W and, for every finite subtheory V_0 of V, we have $U \rhd (W + V_0)$.

Theorem 4.1. *Suppose U is recursively enumerable. Then, we have:*

$$U \rhd_{(U,\mathsf{loc})} (U + [\Box_U \bot]_{\mathsf{S}^2_1,U}).$$

Sometimes a result has two essentially different proofs. This is true for our lemma. I give both proofs

Proof. The first proof works by iterating Feferman's Theorem. Suppose $N_i : \mathsf{S}^1_2 \lhd U$, for $i \leq k$ and $K : U \rhd (U + \bigwedge_{i<k} \Box^{N_i}_U \bot)$. Clearly $N_k K : \mathsf{S}^1_2 \lhd U$. By Feferman's Theorem, for some M, we have $M : U \rhd (U + \Box^{N_k K}_U \bot)$. Moreover:

$$K : (U + \Box^{N_k K}_U \bot) \rhd (U + \bigwedge_{i \leq k} \Box^{N_i}_U \bot).$$

So $MK : U \rhd (U + \bigwedge_{i \leq k} \Box^{N_i}_U \bot)$.

A second approach is as follows. Suppose $N_i : \mathsf{S}^1_2 \lhd U$, for $i < k$. We want to consider $\Box^\star_U A := \bigwedge_{i<n} \Box^{N_i}_U A$ as a proof predicate. It is important to make clear for oneself that, for every $\Box^{N_i}_U A$, the code of A is given by an N_i-numeral. So $\bigwedge_{i<n} \Box^{N_i}_U A$ does not result from a uniform substitution in some formula of the form $\bigwedge_{i<n} \mathsf{prov}^{N_i}_U(x)$.

It's not that we cannot write down such a formula, but x could be e.g. in N_0 but not in N_1 and even if x was both in N_0 and N_1 it could play the role of, say, 7 in the first case and 114 in the second. In spite of the apparent obstacles, we can prove the Second Incompleteness Theorem for $\Box^\star$.

We first note that do have $\mathsf{K4}$ for $\Box^\star_U A$. The 4 principle is because we have $U \vdash \Box^{N_i}_U A \to \Box^{N_i}_U \Box^\star_U A$. So it is sufficient to prove a relevant version of the Gödel Fixed Point Lemma. We briefly sketch how this works.

Let $v_0, \ldots, v_{k-1}$ be a sequence of designated variables. The idea is that v_i ranges over N_i. We define a substitution function $\mathsf{subst}(x, y)$ that substitutes simultaneously, for each v_i the N_i-numeral of x in y. Consider any formula $B(v_0, \ldots, v_{k-1})$. Let $C(v_0, \ldots, v_{k-1}) := B(\mathsf{subst}^{N_0}(v_0, v_0), \ldots, \mathsf{subst}^{N_0}(v_{k-1}, v_{k-1}))$. Let c be the Gödel number of C. Let $\underline{c}_{(i)}$ be the N_i-numeral of c and let $D := C(\underline{c}_{(0)}, \ldots, \underline{c}_{(k-1)})$. It is easy to see that $U \vdash D \leftrightarrow B(\underline{\ulcorner D \urcorner}_{(0)}, \ldots, \underline{\ulcorner D \urcorner}_{(k-1)})$. Given the ingredients we collected, we can now prove the Second Incompleteness Theorem for $\Box^\star_U$.

Using a disjunctive interpretation one can show that $(U + \Diamond^\star_U E) \rhd (U + E)$. Using this we can repeat the usual proof of Feferman's Theorem for $\Box^\star_U$. ❑

Remark 4.2. If U is sequential, the above result has a third proof. By an insight due to Pudlák, we can find an $N : \mathsf{S}^1_2 \lhd U$ that is verifiably definably initially embeddable in N_0, ..., N_{k-1}. By Feferman's Theorem, we have $U \rhd (U + \Box^N_U \bot)$ and, hence, by upward persistence of Σ_1-sentences, $U \rhd (U + \bigwedge_{i<k} \Box^{N_i}_U \bot)$. ❏

Here is the promised modal version of Feferman's Theorem.

Theorem 4.3. *Suppose U is recursively enumerable and for some N, we have $N : \mathsf{S}^1_2 \lhd U$. Then, $U \models \blacklozenge_U \blacksquare_{\mathsf{S}^1_2} \Box_U \bot$, or, equivalently, $U \blacktriangleright (U + [\Box_U \bot]_{\mathsf{S}^2_1, U})$.*

We note that, the equivalence of $U \models \blacklozenge_U \blacksquare_{\mathsf{S}^1_2} \Box_U \bot$ and $U \blacktriangleright (U + [\Box_U \bot]_{\mathsf{S}^2_1, U})$ follows from Theorem 2.6.

Proof. Consider any model $\mathcal{M} \models U$. If, for each internal S^1_2-model $\mathcal{N}$ of $\mathcal{M}$, we have $\mathcal{N} \models \Box_U \bot$, we are done. Otherwise, for some internal S^1_2-model $\mathcal{N}^\star$, we have $\mathcal{N}^\star \models \Diamond_U \top$. Hence, *a fortiori*, $\mathcal{N}^\star \models \mho(U)$. Since $U \rhd_{\mathsf{loc}} (U + [\Box_U \bot]_{\mathsf{S}^1_2, U})$, it follows, by Theorem 2.9, that $\mho_U \rhd (U + [\Box_U \bot]_{\mathsf{S}^1_2, U})$. (We remind the reader that $U + [\Box_U \bot]_{\mathsf{S}^1_2, U}$ is a recursively enumerable theory.) Let K be the interpretation witnessing this. Then, $\mathcal{M}^\star := \widetilde{K}(\mathcal{N}^\star)$ satisfies $U + [\Box_U \bot]_{\mathsf{S}^1_2, U}$. We note that $\mathcal{M}^\star$ is an internal U-model of $\mathcal{M}$. By Theorem 2.6, we find that $\mathcal{M}^\star \models \blacksquare_{\mathsf{S}^1_2} \Box_U \bot$. ❑

We show that the model-interpretability of $(U + [\Box_U \bot]_{\mathsf{S}^2_1, U})$ in U is, for certain theories, optimal.

Theorem 4.4. *Suppose A is finitely axiomatized, consistent and sequential. Then, $A \not\rhd (A + [\Box_A \bot]_{\mathsf{S}^2_1, A})$.*

Proof. Suppose A is finitely axiomatized, consistent and sequential. Let $W := (A + [\Box_A \bot]_{\mathsf{S}^2_1, A})$. Suppose $A \rhd W$, then clearly $A \equiv W$. By Theorem 2.13, the theory W is trustworthy. It follows that there is a faithful interpretation N of S^1_2 in

W. Let N_0 be any interpretation of S^1_2 in A. (Such an interpretation exists because A is sequential.) Then, $M := N\langle(\bigwedge \mathsf{S}^1_2)^N\rangle N_0$ is an interpretation of S^1_2 in A. Thus, $\Box^M_A\bot$ is an axiom of W and hence $W \vdash \Box^N_A\bot$, contradicting the faithfulness of N. ❑

Theorem 4.4 provides a separating example between interpretability and model-interpretability.

4.2. **Krajíček Theories.** In this subsection we show that if A is finitely axiomatized, then we can model-interpret a *Krajíček theory* for A in A. A Krajíček theory for A is axiomatized by A plus, for every $N : \mathsf{S}^1_2 \lhd A$, a statement of the form $\Box^N_{A,n}\bot$, where n varies with N. In other words, a Krajíček theory is axiomatized by $A + \{\Box^N_{A,\nu(N)}\bot \mid N : \mathsf{S}^1_2 \lhd A\}$, where ν maps each N to some standard number. The possibility of such theories was first noted in [Kra87].

Theorem 4.5. *Let A be finitely axiomatized. Let $N_i : \mathsf{S}^1_2 \lhd A$ enumerate the number systems of A. We can effectively construct a theory $\mathsf{kraj}(A)$ of the form $A + \{\Box^{N_i}_{A,n_i}\bot \mid i \in \omega\}$ and $A \rhd_{\mathsf{loc}} \mathsf{kraj}(A)$.*

Note that $\mathsf{kraj}(A)$ is a *specific* Krajíček theory.

Proof. Let $N_i : \mathsf{S}^1_2 \lhd A$ enumerate the number systems of A. Then, we can find a sequence $n_0, n_1, \ldots$ such that, for every k, we have: $A \rhd (A + \bigwedge_{i<k} \Box^{N_i}_{n_i}\bot)$.

We construct interpretations K_j and numbers n_i in stages. At stage k we produce $K_k : A \rhd (A + \bigwedge_{i<k} \Box^{N_i}_{n_i}\bot)$ and at stage $k+1$ we construct n_k.

At stage 0, we take $K_0 := \mathsf{ID}_A$. We consider stage $k+1$. We are given $K_k : A \rhd (A + \bigwedge_{i<k} \Box^{N_i}_{n_i}\bot)$. Clearly, $N_k K_k : \mathsf{S}^1_2 \lhd A$. By Theorem 3.2, we can effectively find an n_k and an M such that $M : A \rhd (A + \Box^{N_k K_k}_{n_k}\bot)$. It follows that:

$$K_{k+1} := K_k M : A \rhd (A + \bigwedge_{i<k+1} \Box^{N_i}_{n_i}\bot).$$

Hence it follows that $A \rhd_{\mathsf{loc}} (A + \{\Box^{N_i}_{n_i}\bot \mid i \in \omega\})$. ❑

Before proceeding we need some definitions.

- A model $\mathcal{M}$ of A is *a Krajíček model* for A if, for all internal models $\mathcal{N}$ of S^1_2 in $\mathcal{M}$, there is an n such that $\mathcal{N} \models \Box_{A,n}\bot$. The class of all Krajíček models for A is $\mathfrak{K}_A$. The predicate logical theory of all Krajíček models for A is $\mathsf{Th}(\mathfrak{K}_A)$.
- Consider an interpretation $N : \mathsf{S}^1_2 \lhd A$ and an A-model $\mathcal{K}$. We say that J is an *infinite initial segment* of N in $\mathcal{K}$ if, in $\mathcal{K}$, the set given by J is a downward closed subset of δ_N and if, for each standard n, $\mathcal{K}$ satisfies $J(\underline{n})$.
- We define $\mathsf{S}^1_{2,\jmath}$ as the theory in the language of arithmetic extended by a unary predicate $\jmath$ axiomatized by:
$$\mathsf{S}^1_2 + \forall x, y((\jmath(x) \wedge y \leq x) \rightarrow \jmath(y)) + \{\jmath(\underline{n}) \mid n \in \omega\}.$$
So $\mathsf{S}^1_{2,\jmath}$ is the theory of an infinite initial segment.

- We define $\Box_{A,\jmath} B :\leftrightarrow \exists x\,(\jmath(x) \wedge \Box_{A,x} B)$.

We give a modal characterization of a Krajíček model.

Theorem 4.6. *Let $\mathcal{K}$ be an A-model. We have:*

$$\mathcal{K} \models \blacksquare_{\mathsf{S}^1_{2},\jmath} \Box_{A,\jmath}\bot \text{ iff } \mathcal{K} \text{ is a Krajíček model for } A.$$

Proof. From right to left is immediate. Suppose $\mathcal{K} \models \blacksquare_{\mathsf{S}^1_{2},\jmath} \Box_{A,\jmath}\bot$. Consider any $N : \mathsf{S}^1_2 \lhd A$. In $\mathcal{K}$, we define $J^\star := \{a \in N \mid \Diamond^N_{A,a}\top\}$. In case $J^\star$ contains all standard natural numbers, it is an infinite initial segment. This contradicts that we have $\Box^N_{A,J^\star}\bot$. So J^* must be finite and, thus, for some n, we have $\Box^N_{A,n}\bot$. ❑

The following theorem is our infinitary version of Feferman's Theorem for restricted interpretability.

Theorem 4.7. *Suppose $N_0 : \mathsf{S}^1_2 \lhd A$. Then, every model of A has an internal Krajíček model for A. In modal terms: $A \models \blacklozenge_A \blacksquare_{\mathsf{S}^1_{2},\jmath} \Box_{A,\jmath}\bot$.*

Proof. Consider any model $\mathcal{M}$ of A. In case $\mathcal{M}$ is itself a Krajíček model, we are done.

Otherwise, there is an internal model $\mathcal{N}$ such that $\mathcal{N} \models \mho(A)$. Since $A \rhd_{\mathsf{loc}} \mathsf{kraj}(A)$, we have, by Theorem 2.9, that $\mho(A) \rhd \mathsf{kraj}(A)$. It follows that $\mathcal{N}$ has an internal model $\mathcal{K}$ that satisfies $\mathsf{kraj}(A)$. By transitivity, $\mathcal{K}$ is an internal model of $\mathcal{M}$. We claim that $\mathcal{K}$ is a Krajíček model. Consider any internal S^1_2-model $\mathcal{N}'$ of $\mathcal{K}$. Suppose this model is given by the interpretation N'. Clearly $N'' := N'\langle(\bigwedge \mathsf{S}^1_2)^{N'}\rangle N_0$ is an interpretation of S^1_2 in A. So, for some k, we have $\mathsf{kraj}(A) \vdash \Box^{N''}_{A,k}\bot$. Since, in $\mathcal{K}$, the interpretation N'' defines the same internal model as N', we find that $\mathcal{N}'$ satisfies $\Box_{A,k}\bot$. It follows that $\mathcal{K}$ is a Krajíček model. ❑

Remark 4.8. We note that for consistent, finitely axiomatized, sequential A, we have $A \rhd_{\mathsf{loc}} \mho(A)$. The existence of Krajíček models shows that we cannot have $A \blacktriangleright \mho(A)$. So, we have a separating example between model interpretability and local interpretability. ❏

We can use Krajíček models to show that internal modal logic is more expressive than predicate logic.

Theorem 4.9. *Let A be any finitely axiomatized, consistent, sequential theory. Then $\mathsf{Th}(\mathfrak{K}_A)$, the theory of all Krajíček models has a model that is not itself a Krajíček model.*

Proof. Let A be any consistent finitely axiomatized sequential theory. Let $N_0 : \mathsf{S}^1_2 \lhd A$, be the interpretation promised in Theorem 2.12, such that, for any k, we have $A \rhd (A + \Diamond^{N_0}_{A,k}\top)$. Consider the theory $U := \mathsf{Th}(\mathfrak{K}_A) + \Diamond^{N_0}_{A,c}\top + \{c \neq \underline{n} \mid n \in \omega\}$. Here c is a fresh constant and the $\underline{n}$ are N_0-numerals. We claim that U is consistent. By compactness, it is sufficient to show that, for any n, $U_n := \mathsf{Th}(\mathfrak{K}_A) + \Diamond^{N_0}_{A,\underline{n}}\top$ is consistent. Consider any Krajíček model $\mathcal{K}$. Let $M : A \rhd (A + \Diamond^{N_0}_{A,n}\top)$. Then $\mathcal{M} := \widetilde{M}(\mathcal{K})$ is again a Krajíček model that satisfies $\Diamond^{N_0}_{A,\underline{n}}\top$. Thus, $\mathcal{M} \models U_n$.

Let $\mathcal{K}^\star$ be any model of U. Clearly $\mathcal{K}^\star$ is not a Krajíček model. ❑

We turn to the syntactical trace of $\blacksquare_{\mathsf{S}^1_{2,\jmath}} \Box_{A,\jmath}\bot$.

- We define $\digamma(A) := A + [\Box_{A,\jmath}\bot]_{\mathsf{S}^1_{2,\jmath},A}$. So, $\digamma(A)$ contains the $\Box^N_{A,J}\bot$ such that N is an interpretation of S^1_2 in A and J is an infinite initial segment of N.

Since $\mathsf{S}^1_{2,\jmath}$ is not finitely axiomatized we cannot conclude that, for any A-model $\mathcal{M}$ we have: $\mathcal{M} \models \blacksquare_{\mathsf{S}^1_{2,\jmath}} \Box_{A,\jmath}\bot$ iff $\mathcal{M} \models \digamma(A)$.

Open Question 4.10. Suppose A is a consistent, finitely axiomatized, sequential theory. Is $\mathsf{Th}(\mathfrak{K}_A)$ axiomatized by $\digamma(A)$? ❑

Here is a characterization of $\digamma(A)$.

Theorem 4.11. *Let A be a finitely axiomatized theory. We have: $\digamma(A) \vdash B$ iff $(A + \neg B) \rhd \mho(A)$.*

Proof. Suppose $N : \mathsf{S}^1_{2,\jmath} \lhd A$. Then,

$$A + \forall x \in \jmath^N \Diamond^N_{A,x}\top \vdash \mho^N(A).$$

Clearly, if $\digamma(A) \vdash B$, then $A + \neg B$ implies a disjunction of sentences of the form $\forall x \in \jmath^N \Diamond^N_{A,x}\top$, so $(A + \neg B) \vdash \mho^N(A)$.

Conversely, suppose $N' : (A + \neg B) \rhd \mho(A)$. We extend N' to N by interpreting $\jmath$ as $\{x \in N \mid \neg B \to \Diamond^{N'}_{A,x}\top\}$. Since $A + \exists x\,((\neg B \to \Diamond^{N'}_{A,x}\top) \wedge \Box^{N'}_{A,x}\bot)$ implies B, we have $A + \Box^N_{A,\jmath}\bot \vdash B$. So, $\digamma(A) \vdash B$. ❑

We note that Theorem 4.11 makes it perspicuous that the set of theorems from $\digamma(A)$ is Σ_3.

Open Question 4.12. I conjecture that, for consistent, finitely axiomatized, sequential A, the set of theorems of $\digamma(A)$ is complete Σ_3. (In this paper we show that it is Π_2-hard. See Theorem 5.16.)

We note that it would follow, via Theorem 4.11, that interpretability between recursively enumerable theories is complete Σ_3. This last fact is already known. It was proven by Volodya Shavrukov in [Sha97]. Still it would not hurt to have an alternative proof, in the light of the fact that Shavrukov's proof is quite intricate. ❑

As an immediate consequence of theorem 4.7, we find that $\digamma(A)$ is model-interpretable in A.

Theorem 4.13. $A \blacktriangleright \digamma(A)$.

We cannot generally improve on Theorem 4.13. Suppose A is finitely axiomatized, consistent and sequential. Theorem 4.4 tells us that $A \not\rhd (A + [\Box_A\bot]_{\mathsf{S}^2_1,A})$. So, *a fortiori*, $A \not\rhd \digamma(A)$.

In the next section we will explain that Theorem 4.7 tells us that $\digamma(A)$ is in the semantic completion of A.

5. Completions

In this section, we discuss constructions of certain natural completions of theories. The section is like a walk at the rim of a vast ocean most of which is *mare incognitum*. There are many such completions and most questions concerning them are open.

We introduce the three notions of completion that we will study. Let a theory U be given. Let M and M' range over interpretations of U in U.

I. $\mathsf{synco}(U) := \{A \mid \exists M\,\forall M'\,U \vdash A^{M'M}\}$.

II. $\mathsf{semco}(U) := \{A \mid U \models \blacklozenge_U \blacksquare_U A\}$.

III. $\mathsf{intco}(U) := \{A \mid \forall B\,(U \rhd (U+B) \Rightarrow U \rhd (U+A+B))\}$.

We note that, for finitely axiomatized A, the completion $\mathsf{synco}(A)$ is *prima facie* Σ_3 and $\mathsf{intco}(A)$ is Π_2. To classify semco we note that we can replace the quantification over the external models by a quantification over complete theories and the quantifications over internal models by quantifications over translations. Thus we get:

$$\begin{aligned} A \in \mathsf{semco}(U) \quad \Leftrightarrow \quad & \forall X \subseteq \mathsf{sent}_U\ (\ (X \text{ is a complete extension of } U) \Rightarrow \\ & \exists \tau\ (\ \forall B\,(U \vdash B \Rightarrow B^{\tau} \in X) \wedge \\ & \quad \forall \tau'\,(\forall B\,(U \vdash B \Rightarrow B^{\tau'\tau} \in X) \rightarrow A^{\tau'\tau} \in X)\ )\ ) \end{aligned}$$

Thus, the semantic completion of an recursively enumerable theory is *prima facie* Π^1_1. We will show in Subsection 5.4 that all three completions are Π_2-hard for any consistent, finitely axiomatized, sequential theory A.

5.1. The Syntactic Completion.

The theory $\mathsf{synco}(U)$ is *the syntactic completion of* U. It is easy to see that the syntactic completion contains U, is deductively closed and is closed under conjunction. Thus it is a theory.

Let M, M', M'' range over interpretations of U in U. If we define $M' \leq M :\Leftrightarrow \exists M''\ M' = M''M$, then we have:

$$A \in \mathsf{synco}(U) \Leftrightarrow \exists M\,\forall M' \leq M\ U \vdash A^{M'}.$$

We clearly have:

$$A \in \mathsf{synco}(U) \Leftrightarrow U \rhd (U + [A]_U).$$

We note that $U + [A]_U$ need not be enumerable. If U is a finitely axiomatized theory it clearly is. Moreover, for a finitely axiomatized theory A, the set $[B]_A$ has an natural p-time decidable axiomatization over A to wit:

$$\{B^{\tau\langle A^{\tau}\rangle \mathsf{id}_{\Sigma_A}} \mid \tau : \Sigma_A \to \Sigma_A\}.$$

We show that synco preserves various notions of sameness of theories. So we may consider it as an operation on the various structures of theories modulo one of these equivalence relations.

Theorem 5.1. *The operation* synco *preserves mutual interpretability, sentential congruence, iso-congruence, bi-interpretablity and definitional equivalence.*

Proof. Suppose $K : U \lhd V$ and $M : V \lhd U$ and $A \in \mathsf{synco}(U)$. Then, by Theorem 2.5:

$$V \rhd U \rhd (U + [A]_U) \rhd (V + [A^K]_V).$$

Hence, $A^K \in \mathsf{synco}(V)$. Hence, $K^\star : \mathsf{synco}(V) \lhd \mathsf{synco}(U)$, where $K^\star$ is based on the same translation as K. Similarly, $M^\star : \mathsf{synco}(U) \lhd \mathsf{synco}(V)$.

Now if K, M form e.g. a sentential congruence, then, for any B, $U \vdash B^{KM} \leftrightarrow B$. It follows that $U + [A]_U \vdash B^{K^\star M^\star} \leftrightarrow B$. Similarly for the MK case. So, $K^\star$ and $M^\star$ form a sentential congruence. Similarly, for iso-congruence, bi-interpretablity and definitional equivalence. ❑

In case A is finitely axiomatized and sequential, $\mathsf{semco}(A)$ trivializes as is shown by the following theorem.

Theorem 5.2. *Suppose A is finitely axiomatized and sequential. We have:*
$B \in \mathsf{synco}(A)$ iff $A \vdash B$.

Proof. The right-to-left direction is trivial. We treat left-to-right. If A is inconsistent this is immediate. Suppose A is consistent. Suppose $B \in \mathsf{semco}(A)$. Then, $A \equiv (A + [B]_A)$. By Theorem 2.13, the theory $A + [B]_A$ is trustworthy. This means that if W is interpretable in $A + [B]_A$, then W is faithfully interpretable in $A + [B]_A$. It follows that there is a faithful interpretation M of A in $A + [B]_K$. Let $M^* := M\langle A^M \rangle \mathsf{ID}_A$. We have $M^* : A \lhd A$ and hence $A + [B]_A \vdash B^{M^*}$. Since $A + [B]_A \vdash A^M$, it follows that $A + [B]_A \vdash B^M$. But M is faithful, so $A \vdash B$. ❑

Remark 5.3. Let us restrict ourselves to arithmetical theories A like S^1_2 that are preserved to definable (ω_1-)cuts. Suppose that we replace, in the definition of synco, interpretations by cut-interpretations, i.o.w. by relativization to a definable cut. Let us call the resulting notion $\mathsf{synco}_{\mathsf{cut}}$. Then, $\mathrm{I}\Delta_0 + \Omega_1 + \mathrm{B}\Sigma_1 + \mho(A)$ will be in $\mathsf{synco}_{\mathsf{cut}}(A)$. Hence, we do not have an analogue of Theorem 5.2 in the case of cut-interpretability. ❍

Remark 5.4. There is a model theoretic variant of the definition that works as follows. Let $\mathcal{M}$, $\mathcal{M}'$ range over U-models and let M range over interpretations of U in U. We remind the reader that $\widetilde{M}$ is the functor that associates an internal U-model $\mathcal{M}'$ to an U-model $\mathcal{M}$ using the translation τ_M. We define:

$$A \in \mathsf{synco}^+(U) \Leftrightarrow \exists M\, \forall \mathcal{M}\, \forall \mathcal{M}' \lhd \widetilde{M}(\mathcal{M})\; \mathcal{M}' \models A.$$

It is easy to see that $\mathsf{synco}^+(U)$ is contained in $\mathsf{synco}(U)$. In case U is finitely axiomatized, the converse is also true. ❍

5.2. **The Semantic Completion.** The notion of semantic completion was introduced in the context of cut-interpretability by Emil Jeřábek. In fact Jeřábek's notion is not entirely analogous to ours since his formulation was in terms of initial sub-cuts and not in terms of internal subcuts.

The following theorem connects $B \in \mathsf{semco}(U)$ to the model-interpretability of $[B]_U$.

Theorem 5.5. *We have:*

i. Suppose U is any theory and suppose B is an U-sentence. Then,
$B \in \mathsf{semco}(U) \Rightarrow U \blacktriangleright (U + [B]_U)$,

ii. Suppose A is any finitely axiomatized theory and suppose B is an A-sentence. Then, $B \in \mathsf{semco}(A) \Leftrightarrow A \blacktriangleright (A + [B]_A)$.

Proof. Ad (i). Suppose $B \in \mathsf{semco}(U)$ and $\mathcal{M}$ is an U-model. Let $\mathcal{M}'$ be an internal U-model of $\mathcal{M}$ such that $\mathcal{M}' \models \blacksquare_U B$. Consider any $K : U \rhd U$. Then, clearly $\widetilde{K}(\mathcal{M}') \models B$ and, hence, $\mathcal{K}' \models B^K$.

Ad (ii). From left-to-right is by (i). Suppose $A \blacktriangleright (A + [B]_A)$. Consider any A-model $\mathcal{M}$. Let $\mathcal{M}'$ be the promised internal A-model of $\mathcal{M}$ such that $\mathcal{M}' \models [B]_A$. By Theorem 2.6, we find that $\mathcal{M}' \models \blacksquare_A B$. ❑

It follows immediately that, for finitely axiomatized A, the theory $\mathsf{synco}(A)$ is contained in $\mathsf{semco}(A)$.

Remark 5.6. We note that $\mathsf{synco}^+(U)$ of Remark 5.4 is contained in $\mathsf{semco}(U)$, also in the infinitely axiomatized case. ❑

The operation semco preserves all good notions of sameness that one can think of. Thus it can be seen as a good operation from the standpoint of a more abstract view of theories.

Theorem 5.7. *The operation* semco *preserves mutual interpretability, sentential congruence, iso-congruence, bi-interpretablity and definitional equivalence.*

Proof. Suppose $K : U \lhd V$ and $M : V \lhd U$. Suppose $A \in \mathsf{semco}(U)$. We prove that $A^K \in \mathsf{semco}(V)$. Consider any V-model $\mathcal{M}$. Let $\mathcal{K} := \widetilde{K}(\mathcal{M})$. By our assumption, $\mathcal{K}$ has an internal U-model $\mathcal{K}'$ such that $\mathcal{K}' \models \blacksquare_U A$. Clearly, also $\mathcal{M}' := \widetilde{M}(\mathcal{K}') \models \blacksquare_U A$. It follows that $\mathcal{M}' \models \blacksquare_V \blacksquare_U A$, and hence $\mathcal{M}' \models \blacksquare_V A^K$. Clearly $\mathcal{M}'$ is an internal model of $\mathcal{M}$ and we are done.

We have shown that K lifts to an interpretation $K^\star$ with the same underlying translation of $\mathsf{semco}(U)$ in $\mathsf{semco}(V)$. Similarly we can lift M to an interpretation $M^\star$ of $\mathsf{semco}(V)$ in $\mathsf{semco}(U)$. It is easy to see that the properties that make K, M into a definitional equivalence, a bi-interpretation, an iso-congruence, or a sentential congruence are preserved from K, M to $K^\star, M^\star$. ❑

We show that inconsistencies are highly non-arbitrary by the lights of the semantic completion.

Theorem 5.8. *The theory* $\mathsf{semco}(U)$ *contains* $U + [\Box_U \bot]_{\mathsf{S}^1_2, U}$.

Proof. By Theorem 4.3, we have $U \models \blacklozenge_U \blacksquare_{\mathsf{S}^1_2} \Box_U \bot$. Hence, $U \models \blacklozenge_U \blacksquare_U \blacksquare_{\mathsf{S}^1_2} \Box_U \bot$, and, thus, $U \models \blacklozenge_U \blacksquare_U [\Box_U \bot]_{\mathsf{S}^1_2, U}$. ❑

We remind the reader that, for finitely axiomatized A, we have $\digamma(A) := A + [\Box_{A,\jmath} \bot]_{\mathsf{S}^1_{2,\jmath}, A}$.

Theorem 5.9. *Suppose A is finitely axiomatized. Then,* $\mathsf{semco}(A)$ *contains* $\digamma(A)$.

Proof. By Theorem 4.7, we have $A \models \blacklozenge_A \blacksquare_{\mathsf{S}^1_{2,\jmath}} \Box_{A,\jmath} \bot$. Hence, $A \models \blacklozenge_A \blacksquare_A \blacksquare_{\mathsf{S}^1_{2,\jmath}} \Box_{A,\jmath} \bot$, and, thus, $A \models \blacklozenge_A \blacksquare_A [\Box_{A,\jmath} \bot]_{\mathsf{S}^1_{2,\jmath}, A}$. ❑

We end this subsection by a remark in which we reflect on the meaning of a piece of $\mathsf{S4}$ reasoning.

Remark 5.10. Suppose $U \models B \to \blacksquare_U B$. We claim that:

$$U \models \blacklozenge_U \blacksquare_U (\neg B \to \blacksquare_U \neg B).$$

This follows from the following two inferences:

$$\begin{array}{lllll} U \models & \blacklozenge_U B & \to & \blacklozenge_U \blacksquare_U B \\ & & \to & \blacklozenge_U \blacksquare_U (\neg B \to \blacksquare_U \neg B) \end{array}$$

$$U \models \ \blacksquare_U \neg B \ \to \ \blacklozenge_U \blacksquare_U (\neg B \to \blacksquare_U \neg B)$$

It follows that for B such that $U \models B \to \blacksquare_U B$, we have $(\neg B \to [\neg B]_U)$ in $\mathsf{semco}(U)$. Here $(\neg B \to [\neg B]_U) := \{\neg B \to C \mid C \in [B]_U\}$.

If we want to apply the above insight to finitely axiomatized sequential theories we are in for a disappointment. Consider a finitely axiomatized, sequential theory A. For which B do we have: $A \models B \to \blacksquare_A B$ or equivalently $A + B \vdash [B]_A$? We certainly have this when $A + B$ is inconsistent. Suppose $A + B$ is consistent. By Theorem 2.13, $A + B$ is trustworthy. Let K be a faithful interpretation of A in B. Let $K' := K\langle A^K \rangle \mathsf{ID}_A$. Clearly $K' : A \rhd A$. So, if $A + B \vdash [B]_A$, we find $A + B \vdash B^{K'}$ and hence $A + B \vdash B^K$. Since K is faithful, it follows that $A \vdash B$. So $A + B \vdash [B]_A$ if either $A \vdash \neg B$ or $A \vdash B$. Thus, in the finitely axiomatized, sequential case the above observation does not have an interesting application.

In case we change the interpretation of $\blacksquare$ by supposing that A is an arithmetical theory that is preserved to definable cuts and by taking as accessibility relation the definable cut relation, we have a completely different situation. Let's signal our change of meaning using a superscript cut. We have $A \models^{\mathsf{cut}} P \to \blacksquare^{\mathsf{cut}}_A P$, for any Π_1 sentence P. So it follows that, for any Σ_1-sentence S we have $A + S \to [S]^{\mathsf{cut}}_A$,is in $\mathsf{semco}^{\mathsf{cut}}(A)$. If we take $A := \mathsf{PA}^-$ this gives us precisely that the theory Peano Basso is in the semantic completion. See [Vis12b].

We are in the following interesting situation: our full present knowledge of sentences in $\mathsf{semco}(A)$ beyond A itself comes from Feferman style reasoning. This does not give us anything when we look at $\mathsf{semco}^{\mathsf{cut}}(A)$, since restriction to cuts cannot introduce Σ_1-unsoundness. On the other hand our full knowledge of extra principles beyond A in $\mathsf{semco}^{\mathsf{cut}}(A)$ comes from the above $\mathsf{S4}$ reasoning. As we have shown this reasoning is completely powerless for the semco case. Thus in the present stage of knowledge $\mathsf{semco}(A)$ and $\mathsf{semco}^{\mathsf{cut}}(A)$ seem to be orthogonal. ❑

5.3. **The Intrinsic Completion.** The idea for $\mathsf{intco}(U)$ is an adaptation of an idea that Emil Jeřábek formulated in the context of cut-interpretability. We note that we have $U \rhd_{(U,\mathsf{loc})} \mathsf{intco}(U)$.

Remark 5.11. One fanciful way to think about the intrinsic completion is as follows. Hilbert's program for foundations was very crudely: justify a theory by showing its consistency. One problem of Hilbert's approach was the non-uniqueness problem: mutually contradictory extensions of a given theory may be consistent. One solution to this problem is to say that meaning is theory internal, so that the extensions do not *really* contradict each other since the content of the A in extension 1 is not the content of A in $\neg A$ in extension 2.

Nelson's foundational program ([Nel86]) can be viewed as replacing consistency proofs by relative interpretability. (It is not quite clear if he wants general interpretability or just interpretability by relativization to definable cuts. Both notions are interesting.) Here we still have non-uniqueness because of the Orey phenomenon that we can interpret mutually contradictory extensions.[8] In the light of the Orey phenomenon, we can make two moves. Just as in the Hilbert case we can say that the meaning of the extensions is different. In fact we can view the meaning as *given* by the interpretation. The other way is to restrict oneself to extensions that are compatible with all other extensions: i.e. to opt for the intrinsic completion. ❑

We start with a characterization of $\mathsf{intco}(U)$ in terms of the U-local interpretability of $[A]_U$.

Theorem 5.12. *We have:* $A \in \mathsf{intco}(U) \Leftrightarrow U \rhd_{(U,\mathsf{loc})} (U + [A]_U)$.

The idea of the proof is due to Joel Hamkins (in e-mail correspondence).

Proof. Suppose $A \in \mathsf{intco}(U)$. Let $K_0, \ldots, K_{n-1}$ be interpretations of U in U. Clearly, $U \rhd (U + (\neg A \vee \bigwedge_{i<n} A^{K_i}))$, by the interpretation:

$$K := K_0\langle\neg A^{K_0}\rangle(K_1\langle\neg A^{K_1}\rangle(\ldots(K_{n-1}\langle\neg A^{K_{n-1}}\rangle\mathsf{ID}_U)\ldots)).$$

It follows that $U \rhd (U + A + (\neg A \vee \bigwedge_{i<n} A^{K_i}))$. Ergo, $U \rhd (U + \bigwedge_{i<n} A^{K_i})$.

Conversely, suppose $U \rhd_{(U,\mathsf{loc})} (U + [A]_U)$ and $L : U \rhd (U + B)$. We find:

$$U \rhd (U + A^L) \rhd (U + A + B).$$

So, $U \rhd (U + A + B)$. ❑

We note that we have:

Theorem 5.13. *Let A be finitely axiomatized. Then,*

- $B \in \mathsf{synco}(A)$ *iff* $A \rhd (A + [B]_A)$.
- $B \in \mathsf{semco}(A)$ *iff* $A \blacktriangleright (A + [B]_A)$.
- $B \in \mathsf{intco}(A)$ *iff* $A \rhd_{\mathsf{loc}} (A + [B]_A)$.

As a consequence we have: $\mathsf{synco}(A) \subseteq \mathsf{semco}(A) \subseteq \mathsf{intco}(A)$.

Next we show that intco is a good operation w.r.t. more abstract views of theories.

Theorem 5.14. *The operation* intco *preserves mutual interpretability, sentential congruence, iso-congruence, bi-interpretablity and definitional equivalence.*

Proof. The proof is analogous to the proof of Theorem 5.1. ❑

Finally we show that $[\Box_A\bot]_{\mathsf{S}^1_2,U}$ is a subtheory of $\mathsf{intco}(U)$.

Theorem 5.15. $[\Box_A\bot]_{\mathsf{S}^1_2,U}$ *is a subtheory of* $\mathsf{intco}(U)$.

[8] Solovay found a variant of the Orey phenomenon for cut-interpretability. Here the sentences are not strictly contradictory but their conjunction implies exp, i.e. the totality of exponentiation, which is a *taboo* statement in Nelson's program.

We note that we cannot conclude our theorem immediately from Theorem 5.8, since, for non-finitely axiomatized U, we do not know whether $\mathsf{semco}(U)$ is included in $\mathsf{intco}(U)$.

Proof. By Theorem 4.1, we have $U \rhd_{(U,\mathsf{loc})} (U + [\Box_U \bot]_{\mathsf{S}^2_1,U})$. We clearly have:

$$(U + [\Box_U \bot]_{\mathsf{S}^2_1,U}) \vdash (U + [[\Box_U \bot]_{\mathsf{S}^2_1,U}]_{U,U}).$$

Hence, $U \rhd_{(U,\mathsf{loc})} (U + [[\Box_U \bot]_{\mathsf{S}^2_1,U}]_{U,U})$. So we are done by Theorem 5.12. ❑

5.4. **Complexity.** We show that for finitely axiomatized, sequential, consistent A, the theories $\digamma(A)$, $\mathsf{semco}(A)$ and $\mathsf{intco}(A)$ are Π_2-hard. We note that $\digamma(A)$ is *prima facie* Σ_3, $\mathsf{semco}(A)$ is *prima facie* Π^1_1 and $\mathsf{intco}(A)$ is *prima facie* Π_2. So we find that $\mathsf{intco}(A)$ is Π_2-complete.

Theorem 5.16. *Suppose A is a consistent, finitely axiomatized sequential theory. Then, $\digamma(A)$, $\mathsf{semco}(A)$ and $\mathsf{intco}(A)$ are Π_2-hard.*

Proof. Let A be a consistent, finitely axiomatized sequential theory. We will provide a p-time computable function Φ from Σ_1-formulas in one variable $S(x)$ to A-sentences such that the following are equivalent:

i. $\forall x\, S(x)$ is true.
ii. $\Phi(S)$ is in $\digamma(A)$.
iii. $\Phi(S)$ is in $\mathsf{semco}(A)$.
iv. $\Phi(S)$ is in $\mathsf{intco}(A)$.

We first construct Φ. Let N_0 be the interpretation given by Lemma 2.12. Consider any Σ_1-formula $S(x)$. By Lemma 2.2, we can effectively find a Σ_1-formula $R(x)$ such that:

$$(\dagger) \quad \{n \in \omega \mid S(n)\} = \{n \in \omega \mid R(n)\} = \{n \in \omega \mid A \rhd (A + R^{N_0}(n))\}.$$

We define $J(x) :\leftrightarrow x \in N_0 \wedge \forall y \leq x\, R^{N_0}(x)$. We take $\Phi(S) := \Box^{N_0}_{A,J}\bot$.

Suppose $\forall n\, S(n)$ is true. Then, by Lemma 2.2, $\forall n\, R(n)$. It follows, by Σ_1-completeness that J is an infinite initial segment for A, N_0. Hence $\Box^{N_0}_{A,J}\bot$ is in $\digamma(A)$. So (i) implies (ii).

Suppose $\Box^{N_0}_{A,J}\bot$ is in $\digamma(A)$, then by Theorem 5.13, $\Box^{N_0}_{A,J}\bot$ is in $\mathsf{semco}(A)$. So (ii) implies (iii). Similarly, (iii) implies (iv).

We show that (iv) implies (i). Suppose $\Box^{N_0}_{A,J}\bot$ is in $\mathsf{intco}(A)$. Consider any n. Since, by Lemma 2.12, we have $A \rhd (A + \Diamond^{N_0}_{A,n}\top)$, it follows by the definition of intco, that:

$$A \rhd (A + \Diamond^{N_0}_{A,n}\top + \exists x \in J\, \Box^{N_0}_{A,x}\bot).$$

Hence, $A \rhd (A + R(n))$. By (†) we may conclude that $S(n)$. Since n was arbitrary, we find: $\forall n\, S(n)$. ❑

6. Conservativity of the Negation of Σ_1-Collection

We present a well-known construction of Paris & Kirby ([PK78]) to show the conservativity of the negation of Σ_1-collection. In this section we work in extensions of $\mathrm{I}\Delta_0$. For the purposes of this section, a Σ_1-formula is a formula of the form $\exists\vec{x}\, S_0\vec{x}$, where S_0 is Δ_0, i.e. S_0 contains only bounded quantifiers. We use that over EA we have a Σ_1-predicate $\mathsf{def}_{\vec{x}}(y,z)$ such that whenever an element a is Σ_1-definable in parameters $\vec{b}$, then, for some numeral k, $\mathsf{def}_{\vec{b}}(\underline{k},z)$ defines a. We follow Paris & Kirby in defining def as follows. Let $\mathsf{T}(e,w,x)$ is Kleene's T-predicate where T is Δ_0. We take:

$$\mathsf{def}_{\vec{x}}(y,z) :\Leftrightarrow \exists v\, (\mathsf{T}(y,\langle\vec{x},z\rangle,v) \wedge \forall w' < \langle z,v\rangle \neg\, \mathsf{T}(y,\langle\vec{x},(w')_0\rangle,(w')_1)).$$

Consider any model $\mathcal{N}$ of $\mathrm{I}\Delta_0$. Let $\vec{m}$ be a finite set of elements of $\mathcal{N}$. Let M be the set of $\Sigma_{1,0}(\vec{m})$-definable elements of $\mathcal{N}$. Clearly, M is closed under the arithmetical operations 0, successor, plus and times. Let $\mathcal{M}$ be the restriction of $\mathcal{N}$ to M. For any Π_2-formula $A(\vec{k})$, with parameters $\vec{k}$ from $\mathcal{M}$, we have, as is easily seen, that, whenever $\mathcal{N} \models A(\vec{k})$, then $\mathcal{M} \models A(\vec{k})$. Thus, $\mathcal{M}$ will satisfy $\mathrm{I}\Delta_0$. If $\mathcal{N} \models \mathsf{EA}$, then $\mathcal{M} \models \mathsf{EA}$, etc.

Suppose that $\mathcal{N} \models \mathsf{EA}$. Let $\mathcal{M}$ be the model constructed above for any $\vec{m}$. Suppose $\mathcal{M}$ is non-standard and that $m^\star$ is a non-standard element of $\mathcal{M}$. Then, we have: $\mathcal{M} \models \forall x < m^\star + 1\, \exists y < m^\star\, \mathsf{def}_{\vec{m}}(y,x)$. Hence $\mathcal{M}$ satisfies the negation of Σ_1-coll. (If $\mathcal{M}$ did satisfy Σ_1-coll, this would give us a bound b for the relevant witnesses of def. Thus we could replace def by a Δ_0-formula. This would contradict the well-known fact that we have the Δ_0-pigeon hole principle in EA. See e.g. [HP93], p42.)

We prove that $\neg\,\Sigma_1$-coll is Π_3-conservative over EA.

Suppose A is Π_3 and $\mathsf{EA} \nvdash A$. Let $\mathcal{N}$ be a non-standard model of EA plus $\neg\, A$. Suppose A is of the form $\forall\vec{x}\, A_0(\vec{x})$, where A_0 is Σ_2. Pick $\vec{m}$ such that $\mathcal{N} \models \neg\, A_0(\vec{m})$. Let n be any non-standard element of $\mathcal{N}$. We now construct the submodel $\mathcal{M}$ of $\mathcal{N}$ by restriction to the $\Sigma_1(n,\vec{m})$-definable elements. Then, by our previous considerations, $\mathcal{N}$ is a model of $\neg\,\Sigma_1$-coll and $\mathcal{N} \models \neg\, A_0(\vec{m})$. Thus, $\mathsf{EA} + \neg\,\Sigma_1$-$\mathsf{coll} \nvdash A$.

Remark 6.1. It is unknown whether the presence of the totality of exponentiation can be eliminated from the argument above. In fact we do not know whether $\mathrm{I}\Delta_0 + \neg\,\mathsf{exp} + \neg\,\mathrm{B}\Sigma_1$ is consistent. See [AKP12] for a discussion of the state-of-the-art. ❑

Our purpose is now to show that this conservativity result is verifiable in a weak theory like $\mathrm{I}\Delta_0 + \Omega_1$. There is a p-time transformation of a proof of a Π_3-sentence A from $\mathsf{EA} + \neg\,\Sigma_1$-$\mathsf{coll}$ into a proof of A from EA. Our strategy is to transmute the above model construction into the construction of an interpretation with similar properties.

We will construct, for every Σ_3-sentence B, an interpretation

$$(\ddagger) \quad Q_B : (\mathsf{EA} + \neg\,\Sigma_1\text{-}\mathsf{coll} + B) \rhd (\mathsf{EA} + B)$$

such that both Q_B and the proof witnessing (‡) are polynomial in B. Then we can reason as follows. Suppose C is Π_3 and (i) $\mathsf{EA} + \neg\,\Sigma_1\text{-}\mathsf{coll} \vdash C$. We have: (ii) $\mathsf{EA} + \neg C \vdash (\mathsf{EA} + \neg\,\Sigma_1\text{-}\mathsf{coll} + \neg\, C)^{Q_{\neg C}}$. On the other hand, by (i) and (ii), we have (iii) $\mathsf{EA} + \neg C \vdash C^{Q_{\neg C}}$. The proof witnessing (iii) is p-time in the original proof witnessing (i). Combining (ii) and (iii), we find $\mathsf{EA} + \neg\, C \vdash \bot$, and, hence, (iv) $\mathsf{EA} \vdash C$. Of course, the proof witnessing (iv) is p-time in the proof witnessing (i).

In the Paris–Kirby construction the standard numbers play an important role: the Σ_1-definitions we use are standard. We need a suitable substitute for the standard numbers when we internalize the construction. Our substitute will be a *strict cut*: a definable cut of our numbers such that we can provably produce an element above the cut. Clearly, true arithmetical theories have no strict cuts, so we have to pre-process our theory to insert a strict cut. This is where Feferman's Theorem for restricted provability comes in.

We use that, for any finitely axiomatized theory A and any $N : \mathsf{S}^1_2 \lhd A$, we have: $K : A \rhd (A + \Box^N_{A,n}\bot)$. Inspection of the construction shows that (the Gödelnumber of K) is polynomial in n and (the Gödel numbers of) A and N. By the results of [Pud85] (see also [Vis93]), we can find a cut J of N such that $A \vdash \Diamond^J_{A,n}\top$. Using the technique of writing short formulas (see [Pud85] and [FR79]), we can show that the size of J just depends polynomially on n. We find that, in $A + \Box^N_{A,n}\bot$, the cut J is a strict cut of N.

We now consider $A_0 := \bigwedge \mathsf{EA} + B$, where B is Σ_3. In this theory, we interpret $A_1 := A_0 + \Box_{A_0,n}\bot$, for sufficiently large n. We proceed in A_1. Suppose B is of the form $\exists \vec{x}\, B_0(\vec{x})$, where B_0 is Π_2. Using an interpretation with parameters we can now interpret $A_2 := \bigwedge \mathsf{EA} + B_0(\vec{c}) + \Box_{A_0,n}\bot$, for fresh constants $\vec{c}$. In A_2, we have the cut J that is below the smallest A_0-proof of p of $\bot$. Since in A_2 we have a truth-predicate true for Σ_1-sentences, we can define the set M of numbers that are $\Sigma_1(\vec{c})$-definable by a definition in J. Relativization to M gives us an interpretation of $A_3 := \mathsf{EA} + \neg\,\Sigma_1\text{-}\mathsf{coll} + B$ in A_2. Composing all our interpretations, we get an interpretation $Q_B : A_3 \to A_0$. This interpretation is p-time in B and so is the witnessing proof.

We see that we have the promised p-time transformation. Moreover, every step in the argument is verifiable in $\mathrm{I}\Delta_0 + \Omega_1$.

7. Feferman's Theorem Fails in the Constructive Case

Feferman's Theorem for parameter-free interpretations fails in the constructive setting. To my knowledge the most elegant proof of this is to use Harvey Friedman's result that the disjunction property implies the numerical existence property. See [Fri75]. The main point of our application of the theorem here is that, since the disjunction property is 'coordinate-free', we have the numerical existence property for any any interpretation of number theory in the given theory. Throughout this section we consider a theory U, where we assume that U is Δ^{b}_1-axiomatized. By the results of [Bus86], we can find a Δ^{b}_1-definition $\mathsf{prf}_U(x,y)$ of the proof-predicate for U.

The theories $\mathsf{i\text{-}S}^1_2$, $\mathsf{i\text{-}T}^1_2$, $\mathsf{i\text{-}EA}$ and $\mathsf{i\text{-}I}\Sigma_1$ are the obvious constructive counterparts of respectively S^1_2, T^1_2, EA and $\mathrm{I}\Sigma_1$. In $\mathsf{i\text{-}S}^1_2$ we have the decidability of Δ^{b}_1-formulas. In $\mathsf{i\text{-}T}^1_2$ we have the Δ^{b}_1-minimum principle.

We first prove a theorem that already blocks the Feferman result (restricted to parameter-free interpretations) for a wide range of theories. After that we will prove a better result using ideas derived from an unpublished note by Emil Jeřábek that we were allowed to use with Emil's gracious permission.

We write $\lhd_{\mathsf{pf}}$ for parameter-free interpretability.

Theorem 7.1 ($\mathsf{i\text{-}EA}$). *Let S be an $\exists\Delta^{\mathsf{b}}_1$-sentence. Suppose $N : \mathsf{i\text{-}T}^1_2 \lhd_{\mathsf{pf}} U$. The following can be verified in* $\mathsf{i\text{-}EA}$. *Suppose that U has the disjunction property and $U \vdash S^N$. Then, S is true or U is inconsistent.*

Proof. Let S be $\exists\Delta^{\mathsf{b}}_1$, say S is of the form $\exists x\, S_0x$, where S_0 is Δ^{b}_1. Let $N : \mathsf{i\text{-}T}^1_2 \lhd_{\mathsf{pf}} U$. We find R with $\mathsf{i\text{-}S}^1_2 \vdash R \leftrightarrow [S \vee \Box_U \neg R^N] \leq \Box_U R^N$. We write $R^\perp$ for the opposite of R, to wit: $\Box_U R^N < [S \vee \Box_U \neg R^N]$.

From this point on, we work in $\mathsf{i\text{-}EA}$. Since we are working in $\mathsf{i\text{-}EA}$, we will use $\Box$ for $\vdash$, etc. Suppose U satisfies the disjunction property and $\Box_U S^N$.

Since in N we have both $\mathsf{i\text{-}T}^1_2$ and S, there is, inside N, a minimal u such that $S(u) \vee \mathsf{prf}_U(u, R^N) \vee \mathsf{prf}_U(\neg R^N)$. It follows that: $\Box_U(R^N \vee R^{\perp N})$.

By the disjunction principle, we find (a) $\Box_U R^N$ or (b) $\Box_U R^{\perp N}$. In case (a), we have (aa) R or (ab) $R^\perp$. If (aa) R, then (aaa) S or (aab) $\Box_U \neg R^N$. In case (aab) we have both $\Box_U R^N$ and $\Box_U \neg R^N$. Hence $\Box_U \perp$. So, in case (aa) we have S or $\Box_U \perp$. In case (ab) we have $R^\perp$, and, hence, by Σ_1-completeness, $\Box_U R^{\perp N}$. Combining this with $\Box_U R^N$, we find $\Box_U \perp$. We may conclude that in case (a) we have S or $\Box_U \perp$.

In case (b) we have $\Box_U \neg R^N$. It follows that (ba) R or (bb) $R^\perp$. If we have (ba) R, we find, by Σ_1-completeness, $\Box_U R^N$ and, hence $\Box_U \perp$. In case (bb), we have $\Box_U R^N$ and, hence, $\Box_U \perp$. So in case (b) we have $\Box_U \perp$.

We may conclude that $\Box_U S^N$ implies either S or $\Box_U \perp$. ❑

We note that Theorem 7.1 is sufficient to block Feferman's Theorem in the parameter-free case. If we had $U \rhd_{\mathsf{pf}} (\mathsf{i\text{-}T}^1_2 + \Box_U \perp)$, then we would also have $\Box_U \perp$.

Remark 7.2. Let us view the mapping $C \mapsto C^N$ not as an interpretation but just as some p-time function from sentences to sentences. Analyzing the proof, we see that the following principles are used:

I. $\mathsf{i\text{-}S}^1_2 \vdash C \quad \Rightarrow \quad U \vdash C^N$

II. $U \vdash (C \to D)^N \to (C^N \to D^N)$

III. $U \vdash (C \vee D)^N \to (C^N \vee D^N)$

IV. $U \vdash \neg \perp^N$

Thus the proof also works e.g. for Boolean morphisms. ❑

Next we present Emil Jeřábek's variant of Theorem 7.1

Theorem 7.3 (i-EA)**.** *Let S be an $\exists\Delta_1^{\mathsf{b}}$-sentence. Say S is of the form $\exists x\, S_0 x$, where S_0 is Δ_1^{b}. Suppose $N : \text{i-S}_2^1 \lhd_{\mathsf{pf}} U$. The following can be verified in* i-EA. *Suppose that U has the disjunction property and $U \vdash (\exists x\, S_0|x|)^N$, then S is true or U is inconsistent.*

Proof. We note that all steps in the proof of Theorem 7.1 go through except one. This is the step where we conclude $\Box_U R^N \vee \Box_U \neg R^N$. For this step we now use that, inside N, there is a minimal $u \leq |x|$ such that $S(u) \vee \mathsf{prf}_U(u, R^N) \vee \mathsf{prf}_U(\neg R^N)$, where x is the promised witness of $S_0|x|$. ❑

Using Theorem 7.3, we are now ready to prove a first approximation of the numerical existence property.

Theorem 7.4 (i-EA)**.** *Suppose $N : \text{i-S}_2^1 \lhd_{\mathsf{pf}} U$. Consider any sentence A of the form $\exists x \in \delta_N\, A_0 x$. The following can be verified in* i-EA. *Suppose that U has the disjunction property and $U \vdash A$. Then, for some n, we have $U \vdash \exists x \leq \underline{n}\ A_0 x$. (Here the numeral $\underline{n}$ is defined relative to N.)*

Proof. Consider any sentence A of the form $\exists x \in \delta_N\, A_0 x$. We define:

- $\boxslash_U B :\leftrightarrow \exists x\, \mathsf{prf}_U(|x|, \underline{\ulcorner B \urcorner})$.

We find a sentence Q with $U \vdash Q \leftrightarrow A \leq \boxslash_U^N Q$.

We reason in i-EA. Suppose U has the disjunction property. Suppose $\Box_U A$. We claim $\Box_U (Q \vee \boxslash_U^N Q)$. To see this, reason inside $\Box_U$. Let x witness A. Either there is a U-proof of Q below $|x|$ or there isn't, since Δ_1^{b}-formulas are provably decidable in i-S_2^1. In the first case, we have $\boxslash_U Q$ and, in the second case, we have Q. We exit from $\Box_U$.

It follows, by the disjunction property, that $\Box_U Q$ or $\Box_U \boxslash_U^N Q$. Applying Theorem 7.3 to the second disjunct, we find $\boxslash_U Q$ and, hence, $\Box_U Q$. So, in all cases, we have $\Box_U Q$. Suppose p is a U-proof of Q. Clearly, it follows that $\Box_U \mathsf{prf}_U(\underline{p}, \underline{\ulcorner Q \urcorner})$. We also have $\Box_U (A \leq \boxslash_U^N Q)$. Let $n := 2^p$. Then, $\Box_U \exists x {\leq} \underline{n}\, A_0 x$. ❑

We are now ready to prove a better version of Σ_1-reflection than Theorem 7.1.

Theorem 7.5 (i-EA)**.** *Let S be an Σ_1-sentence (or, if you wish, a $\Sigma_1(\omega_1)$-sentence). Suppose $N : U \rhd_{\mathsf{pf}} \text{i-S}_2^1$. The following can be verified in* i-EA. *Suppose that U has the disjunction property and $U \vdash S^N$. Then, S is true or U is inconsistent.*

Proof. Let S be an Σ_1-sentence. Say $S = \exists x\, S_0(x)$, where S_0 is Δ_0 (or $\Delta_0(\omega_1)$). Suppose $N : U \rhd_{\mathsf{pf}} \text{i-S}_2^1$. We reason in i-EA. Suppose U has the disjunction property. By Theorem 7.4, for some n, we have $\Box_U \exists x \leq \underline{n}\, S_0 x$. In case $\exists x \leq \underline{n}\, S_0 x$, we have S. In case $\forall x \leq \underline{n} \neg S_0(x)$, we find $\Box_U \forall x \leq \underline{n} \neg S_0(x)$, and, hence, $\Box_U \bot$. ❑

We note that it follows that, if $U \rhd_{\mathsf{pf}} (\text{i-S}_2^1 + \Box_U \bot)$, then U is inconsistent.

Theorem 7.6 (i-IΣ_1)**.** *Suppose $N : \text{i-S}_2^1 \lhd_{\mathsf{pf}} U$. Consider any formula $A_0 x$ with only x free. We can verify the following in* i-IΣ_1. *Suppose U has the disjunction property and $\Box_U \exists x \leq \underline{n}\, A_0 x$. Then, $\exists m\, \Box_U A_0 \underline{m}$.*

Proof. Suppose $N : \text{i-S}^1_2 \lhd_{\mathsf{pf}} U$. Consider A_0x with only x free. We reason in i-IΣ_1. Suppose U has the disjunction property. and $\Box_U \exists x \leq \underline{n}\, A_0x$. The desired result follows by induction on k for the formula $\Box_U \exists y \leq \underline{(n-k)}\, A_0y \vee \exists m\, \Box_U A\underline{m}$. ❑

Theorem 7.7 (i-IΣ_1)**.** *Consider any formula A_0x with only x free. Suppose $N : U \rhd_{\mathsf{pf}} \text{i-S}^1_2$. The following can be verified in* i-IΣ_1. *Suppose that U has the disjunction property and $U \vdash \exists x \in \delta_N\, A_0x$. Then, for some n, we have $U \vdash A_0\underline{n}$.*

Proof. The result is immediate by Theorem 7.4 and Theorem 7.6. ❑

Remark 7.8. What happens when we drop the restriction to parameter-free interpretations? We only have a very limited result.

Suppose U has the disjunction property and $N(\vec{x}) : \text{i-S}^1_2 \lhd U$. Suppose the parameters of $N(\vec{x})$ are taken from the numbers of a parameter-free interpretation M of i-S$^1_2$. Say the parameter-domain is α. We assume that:

i. $U \vdash \forall \vec{x} \in \alpha\;\; \vec{x} \in \delta_M$.

ii. $U \vdash \exists \vec{x}\;\; \vec{x} \in \alpha$.

iii. $U \vdash \forall \vec{x} \in \alpha\, (A^{N(\vec{x})} \wedge \Box_U \bot)$, where A is the conjunction of the axioms of i-S$^1_2$.

Applying Friedman's theorem to M we obtain M-numerals $\underline{\vec{m}}$ such that: $U \vdash \underline{\vec{m}} \in \alpha$. Subtituting $\underline{\vec{m}}$ in N we obtain a parameter-free interpretation $N' := N(\underline{\vec{m}})$ of i-S$^1_2 + \Box_U \bot$. From this it follows that U is inconsistent.

The general question whether it is possible that U has the disjunction property, U is consistent and $U \rhd (\text{i-S}^1_2 + \Box_U \bot)$, where parameters are allowed, is open. ❑

References

[AKP12] Z. Adamowicz, L.A. Kolodziejczyk, and J. Paris. Truth definitions without exponentiation and the Σ_1-collection scheme. *Journal of Symbolic Logic*, 77(2):649, 2012.

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

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

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

[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.

[FR79] J. Ferrante and C.W. Rackoff. *The computational complexity of logical theories*, volume 718 of *Lecture Notes in Mathematics.* Springer, Berlin, 1979.

[Fri75] Harvey Friedman. The disjunction property implies the numerical existence property. *Proceedings of the National Academy of Sciences*, 72(8):2877–2878, 1975.

[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.

[HL08] Joel Hamkins and Benedikt Löwe. The modal logic of forcing. *Transactions of the American Mathematical Society*, 360(4):1793–1817, 2008.

[Hod93] W. Hodges. *Model theory.* Encyclopedia of Mathematics and its Applications, vol. 42. Cambridge University Press, Cambridge, 1993.

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

[Kra87] J. Krajíček. A note on proofs of falsehood. *Archiv für Mathematische Logik und Grundlagenforschung*, 26(1):169–176, 1987.

[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.

[Nel86] E. Nelson. *Predicative arithmetic*. Princeton University Press, Princeton, 1986.

[PK78] J.B. Paris and L.A.S. Kirby. Σ_n-collection schemas in arithmetic. In A. Macintyre, L. Pacholski, and J.B. Paris, editors, *Logic Colloquium '77*, pages 199–209. North–Holland, 1978.

[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.

[Sha97] V.Yu. Shavrukov. Interpreting reflexive theories in finitely many axioms. *Fundamenta Mathematicae*, 152:99–116, 1997.

[Smo77] C. Smoryński. The Incompleteness Theorems. In J. Barwise, editor, *Handbook of Mathematical Logic*, pages 821–865. North-Holland, Amsterdam, 1977.

[Smo85] C. Smoryński. Nonstandard models and related developments. In L.A. Harrington, M.D. Morley, A. Scedrov, and S.G. Simpson, editors, *Harvey Friedman's Research on the Foundations of Mathematics*, pages 179–229. North Holland, Amsterdam, 1985.

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

[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.

[Vis92] A. Visser. An inside view of EXP. *The Journal of Symbolic Logic*, 57(1):131–165, 1992.

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

[Vis05] A. Visser. Faith & Falsity: a study of faithful interpretations and false Σ^0_1-sentences. *Annals of Pure and Applied Logic*, 131(1–3):103–131, 2005.

[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`.

[Vis12a] A. Visser. The arithmetics of a theory. Logic Group Preprint Series 293, Faculty of Humanities, Philosophy, Utrecht University, Janskerkhof 13A, 3512 BL Utrecht, `http://www.phil.uu.nl/preprints/lgps/`, 2012.

[Vis12b] 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.

[Vis13a] 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.

[Vis13b] Albert Visser. Interpretability degrees of finitely axiomatized sequential theories. *Archive for Mathematical Logic*, pages 1–20, 2013.

[Wil86] A.J. Wilkie. On sentences interpretable in systems of arithmetic. In *Logic Colloquium '84*, volume 120 of *Studies in Logic and the Foundations of Mathematics*, pages 329–342. Elsevier, 1986.

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

Appendix A. Further Details on Definitions

In this appendix we explain some basic notions.

A.1. **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 give a bit more details on parameters in Appendix A.3. We will not describe piecewise interpretations here.

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.

- 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.

A.2. **i-morphisms.** Consider an interpretation $K : U \to V$. We can view this interpretation as a uniform way of constructing internal models $\tau_K(\mathcal{M})$ of U from models $\mathcal{M}$ of V. This construction gives us the contravariant model functor as soon as we have defined an appropriate category of interpretations.

Now consider two interpretations $K, M : U \to V$. Between the inner models $\tau_K(\mathcal{M})$ and $\tau_M(\mathcal{M})$ we have the usual structural morphisms of models. We are interested in the case where these morphisms are V-definable and uniform over models. This idea leads to the following definition. An i-morphism $M : K \to M$ is a triple $\langle K, F(\vec{u}, \vec{v}), M\rangle$, where $F(\vec{u}, \vec{v})$ is a V-formula and where $\vec{u}$ has length m_K and $\vec{v}$ has length m_M. We demand:

- $V \vdash F(\vec{u}, \vec{v}) \to (\delta_K(\vec{u}) \wedge \delta_M(\vec{v}))$,
- $V \vdash \delta_K(\vec{u}) \to \exists \vec{v}\, (\delta_M(\vec{v}) \wedge F(\vec{u}, \vec{v}))$,
- $V \vdash (\vec{u}_0 =_K \vec{u}_1 \wedge F(\vec{u}_0, \vec{v}_0) \wedge F(\vec{u}_1, \vec{v}_1)) \to \vec{v}_0 =_M \vec{v}_1$,
- $V \vdash (\vec{u}_0 =_K \vec{u}_1 \wedge \vec{v}_0 =_M \vec{v}_1 \wedge F(\vec{u}_0, \vec{v}_0)) \to F(\vec{u}_1, \vec{v}_1)$,
- $V \vdash (P_K(\vec{u}_0, \ldots \vec{u}_{n-1}) \wedge \bigwedge_{i<n} F(\vec{u}_i, \vec{v}_i)) \to P_M(\vec{v}_0, \ldots \vec{v}_{n-1})$.

Clearly, $F : K \to M$ is an i-morphism iff, for all models $\mathcal{M}$ of V, $F^{\mathcal{M}}$ represents a morphism of models from $\tau_K(\mathcal{M})$ to $\tau_M(\mathcal{M})$.

Two i-morphisms $F, G : K \to M$ are *i-equal*, when $V \vdash \forall \vec{u}, \vec{v}\, (F(\vec{u}, \vec{v}) \leftrightarrow G(\vec{u}, \vec{v}))$.

In the obvious way, we can define the identity i-morphism $\mathsf{Id}_K : K \to K$, composition of i-morphisms, i-isomorphisms, etc. One can show that these operations preserve i-equality. Moreover, i-isomorphisms really are isomorphisms in the categories given by these operations.

We will say that two interpretations K, M are *i-equivalent* when there is an i-isomorphism between them, that is, they are i-isomorphic.

We will *not* divide out i-equivalence of interpretations. This enables us to use the notation τ_M meaningfully, to speak about the dimension of an interpretation, etc. However, we demand that operations on interpretations preserve i-equivalence. It is easy to see that, for example, the operation $K, M \mapsto K\langle A\rangle M$ preserves i-equivalence. Moreover, if K and M are i-equivalent, then $\overline{K} = \overline{M}$.

One can show, by a simple compactness argument, that K and M are i-isomorphic iff, for every $\mathcal{M} \models V$, there is an F such that $F^{\mathcal{M}}$ represents an isomorphism between $\tau_K(\mathcal{M})$ and $\tau_M(\mathcal{M})$.

The category INT_1 is the category of theories (as objects) and interpretations modulo i-equivalence (as arrows). One may show that we have indeed defined a category. The relation of i-equivalence is preserved by composition, etcetera. Two theories U and V are *bi-interpretable* if they are isomorphic in INT_1. Wilfrid Hodges calls this notion: *homotopy*. See [Hod93], p222.

Thus, U and V are bi-interpretable if there are interpretations $K : U \to V$ and $M : V \to U$, so that $M \circ K$ is i-isomorphic to ID_U and $K \circ M$ is i-isomorphic to ID_V. We call the pair K, M a *bi-interpretation* between U and V. One can show that the components of a bi-interpretation are faithful interpretations. Many good properties of theories like finite axiomatizability, decidability, κ-categoricity are preserved by bi-interpretations.

A.3. **Parameters.** In general interpretations are allowed to have parameters. We will briefly sketch how to add parameters to our framework. We first define a translation with parameters. The parameters of the translation are given by a fixed sequence of variables $\vec{w}$ that we keep apart from all other variables. A translation is defined as before, but for the fact that now the variables $\vec{w}$ are allowed to occur in the domain and in the translations of the predicate symbols in addition to the variables that correspond to the argument places. Officially, we represent a translation $\tau_{\vec{w}}$ with parameters $\vec{w}$ as a quintuple $\langle \Sigma, \delta, \vec{w}, F, \Theta\rangle$. The parameter sequence may be empty: in this case our interpretation is parameter-free.

An interpretation with parameters $K : U \to V$ is a quadruple $\langle U, \alpha, E, \tau_{\vec{w}}, V\rangle$, where $\tau_{\vec{w}} : \Sigma_U \to \Sigma_V$ is a translation and α is a V-formula containing at most $\vec{w}$ free. The formula α represents the parameter domain. For example, if we interpret the Hyperbolic Plain in the Euclidean Plain via the Poincaré interpretation, we need two distinct points to define a circular disk. These points are parameters of the construction, the parameter domain is $\alpha(w_0, w_1) = (w_0 \neq w_1)$. (For this specific example, we can also find a parameter-free interpretation.) The formula E represents an equivalence relation on the parameter domain. In practice this is always pointwise identity for parameter sequences, but for reasons of theory one must admit other equivalence relations too. We demand:

- $\vdash \delta_{\tau,\vec{w}}(\vec{v}) \to \alpha(\vec{w})$,
- $\vdash P_{\tau,\vec{w}}(\vec{v}_0, \ldots, \vec{v}_{n-1}) \to \alpha(\vec{w})$.
- $V \vdash \exists \vec{w}\, \alpha(\vec{w})$;

- $V \vdash E(\vec{w}, \vec{z}) \to (\alpha(\vec{w}) \wedge \alpha(\vec{z}))$;
- V proves that E represents an equivalence relation on the sequences forming the parameter domain;
- $\vdash E(\vec{w}, \vec{z}) \to \forall \vec{x}\, (\delta_{\tau,\vec{w}}(\vec{x}) \leftrightarrow \delta_{\tau,\vec{z}}(\vec{x}))$;
- $\vdash E(\vec{w}, \vec{z}) \to \forall \vec{x}_0, \ldots, \vec{x}_{n-1}\, (P_{\tau,\vec{w}}(\vec{x}_0, \ldots, \vec{x}_{n-1}) \leftrightarrow P_{\tau,\vec{z}}(\vec{x}_0, \ldots, \vec{x}_{n-1}))$;
- for all U-axioms A, $V \vdash \forall \vec{w}\, (\alpha(\vec{w}) \to A^{\tau,\vec{w}})$.

We can lift the various operations in the obvious way. Note that the parameter domain of $N := M \circ K$ and the corresponding equivalence relation should be:

- $\alpha_N(\vec{w}, \vec{u}_0, \ldots, \vec{u}_{k-1}) := \alpha_M(\vec{w}) \wedge \bigwedge_{i<k} \delta_{\tau_M}(\vec{w}, \vec{u}_i) \wedge (\alpha_K(\vec{u}))^{\tau_M,\vec{w}}$.
- $E_N(\vec{w}, \vec{u}_0, \ldots, \vec{u}_{k-1}, \vec{z}, \vec{v}_0, \ldots, \vec{v}_{k-1}) :=$ $E_M(\vec{w}, \vec{z}) \wedge \bigwedge_{i<k} \delta_{\tau_M}(\vec{w}, \vec{u}_i) \wedge \bigwedge_{i<k} \delta_{\tau_M}(\vec{w}, \vec{v}_i) \wedge (E_K(\vec{u}, \vec{v}))^{\tau_M,\vec{w}}$.

Consider interpretations $K, M : U \to V$. An i-morphism $\phi : K \to M$ is a triple $\langle K, G, F, M\rangle$, where $G(\vec{u}, \vec{w})$ and $F(\vec{u}, \vec{w}, \vec{x}, \vec{y})$ are V-formulas.[9] We write $F^{\vec{u};\vec{w}}(\vec{x}, \vec{y})$ for F. We demand that:

- V proves that G is a surjective relation between α_K/E_K and α_M/E_M;
- $V \vdash F^{\vec{u};\vec{w}}(\vec{x}, \vec{y}) \to G(\vec{u}, \vec{w})$;
- V proves that, if $G(\vec{u}, \vec{w})$, then $F^{\vec{u};\vec{w}}$ is a function from $\delta_K/{=_K}$ to $\delta_M/{=_M}$.
- V proves that if $E_K(\vec{u}_0, \vec{u}_1)$ and $E_M(\vec{w}_0, \vec{w}_1)$, then $F^{\vec{u}_0,\vec{w}_0}$ is the same function is $F^{\vec{u}_1,\vec{w}_1}$.

Finally, we say that two i-maps ϕ_0 and ϕ_1 are *i-equal* if V proves that G_{ϕ_0} and G_{ϕ_1} and F_{ϕ_0} and F_{ϕ_1} are the same.

The definitions of the identity i-morphism and of composition of i-morphisms are as is to be expected. We can compute what an i-isomorphism is: G is, V-verifiably, a bijection between α_K/E_K and α_M/E_M, and V proves that, if $G(\vec{u}, \vec{w})$, then $F^{\vec{u};\vec{w}}$ is a bijection between $\delta_K/{=_K}$ and $\delta_M/{=_M}$.

A.4. **Complexity Measures.** *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

[9] In G and F we could allow extra parameters, $\vec{z}$, the *eigenparameters* of G and F. We will refrain from doing that here to unburden the presentation a bit.

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 $\mathsf{N}^\star_n$.[10]

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.

Appendix B. Finite Necessity in a Sequential Environment

In this Appendix, we provide characterization of necessity for finitely axiomatized theories in terms of restricted provability. The characterization needs an ambient sequential model.

Suppose $\mathcal{M}$ is a sequential model. Modulo isomorphism, the internal S^1_2-models of $\mathcal{M}$ have a unique intersection $\mathcal{J}_{\mathcal{M}}$. To see this, first consider any internal S^1_2-model $\mathcal{N}$ of $\mathcal{M}$. We take the intersection $\mathcal{J}^{\mathcal{N}}_{\mathcal{M}}$ of all $\mathcal{M}$-definable cuts of $\mathcal{N}$. Consider any other internal S^1_2-model $\mathcal{N}'$ of $\mathcal{M}$ and let $\mathcal{J}^{\mathcal{N}'}_{\mathcal{M}}$ be the intersection of all $\mathcal{M}$-definable cuts of $\mathcal{N}'$. By a theorem of Pudlák, we can find definable cuts $\mathcal{I}$ of $\mathcal{N}$ and $\mathcal{I}'$ of $\mathcal{N}'$ such that there is an $\mathcal{M}$-definable isomorphism between $\mathcal{I}$ and $\mathcal{I}'$. The restriction of that isomorphism to $\mathcal{J}^{\mathcal{N}}_{\mathcal{M}}$ is an isomorphism between $\mathcal{J}^{\mathcal{N}}_{\mathcal{M}}$ and $\mathcal{J}^{\mathcal{N}'}_{\mathcal{M}}$. Thus all $\mathcal{J}^{\mathcal{N}}_{\mathcal{M}}$ are isomorphic. This justifies the notation $\mathcal{J}_{\mathcal{M}}$.

We note that the isomorphisms we produced between $\mathcal{J}^{\mathcal{N}}_{\mathcal{M}}$ $\mathcal{J}^{\mathcal{N}'}_{\mathcal{M}}$ are independent of the chosen cuts: the restrictions to $\mathcal{J}^{\mathcal{N}}_{\mathcal{M}}$ of all definable isomorphisms between cuts of $\mathcal{N}$ resp. $\mathcal{N}'$ are identical. This means that there is precisely one definable isomorphism between $\mathcal{J}^{\mathcal{N}}_{\mathcal{M}}$ and $\mathcal{J}^{\mathcal{N}'}_{\mathcal{M}}$.

One can show that $\mathcal{J}_{\mathcal{M}}$ is a model of (at least) $\mathsf{EA} + \mathrm{B}\Sigma_1 + \mho(A)$.

Theorem B.1. *Let $\mathcal{M}$ be a sequential model and let S be a Σ_1-sentence. Then $\mathcal{M} \models \blacksquare_{\mathsf{S}^1_2} S$ iff $\mathcal{J}_{\mathcal{M}} \models S$.*

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

Proof. From left to right is trivial. Suppose $\mathcal{M} \models \blacksquare_{\mathsf{S}^1_2} S$. Suppose S has the form $\exists x\, S_0 x$, where $S_0 x$ is Δ_0. Consider any internal S^1_2-model $\mathcal{N}$ of $\mathcal{M}$. Let $\mathcal{X} := \{a \in \mathcal{N} \mid \mathcal{N} \models \forall y < a\, \neg S_0(y)\}$. If $\mathcal{X}$ is closed under successor it can be shortened to a definable ω_1-cut $\mathcal{I}$ of $\mathcal{N}$. On this cut we have $\neg S$, contradicting the fact that $\mathcal{I} \models \mathsf{S}^1_2$ and $\mathcal{M} \models \blacksquare_{\mathsf{S}^1_2} S$. So $\mathcal{X}$ is not closed under successor. It follows that, for some b, $\mathcal{N} \models S_0(b) \wedge \forall y < b\, \neg S_0(y)$. Clearly b must be in all definable ω_1-cuts of $\mathcal{N}$. Hence, $\mathcal{J}_{\mathcal{M}} \models S$. ❑

Remark B.2. It seems to me that we can make sense of Theorem B.1 if we allow parameters in S from $\mathcal{J}_{\mathcal{M}}$, since, in every internal S^1_2-model of $\mathcal{M}$, these parameters have unique representatives. So, $\mathcal{M} \models \blacksquare_{\mathsf{S}^1_2} S(\vec{a})$ would mean: for all internal S^1_2-models $\mathcal{N}$ and for the unique representatives $\vec{b}$ in $\mathcal{N}$ of $\vec{a}$, we have $\mathcal{N} \models S(\vec{b})$. ❏

We define $\triangle_A B :\leftrightarrow \Box_{A,\mathsf{max}(\rho(A),\rho(B))} B$.

Theorem B.3. *Let $\mathcal{M}$ be a sequential model. Let A be any finitely axiomatized theory. Then, the following are equivalent:*

i. $\mathcal{M} \models \blacksquare_A B$,

ii. $\mathcal{M} \models \blacksquare_{\mathsf{S}^1_2} \triangle_A B$,

iii. $\mathcal{J}_{\mathcal{M}} \models \triangle_A B$.

Proof. The equivalence between (ii) and (iii) is immediate from Theorem B.1.

We prove: (i) $\Rightarrow$ (ii) by contraposition. Suppose $\mathcal{M} \models \blacklozenge_{\mathsf{S}^1_2} \triangledown_A C$. Then, for some internal S^1_2-model $\mathcal{N}$ of $\mathcal{M}$, we have $\mathcal{N} \models \Diamond_{A,\mathsf{max}(\rho(A),\rho(C))} C$. Using the Henkin-Feferman construction, we can find an internal model $\mathcal{K}$ of $\mathcal{N}$ with $\mathcal{K} \models (A \wedge C)$. By the transitivity of the internal model relation, we have $\mathcal{M} \models \blacklozenge_A C$. (Note that this direction does not use sequentiality.)

We prove (ii) $\Rightarrow$ (i) by contraposition. Suppose $\mathcal{M} \models \blacklozenge_A C$. This means that, for some interpretation K, we have $\mathcal{M} \models (A \wedge C)^K$. Since $\mathcal{M}$ is sequential, for any sufficiently large k, we can find an internal S^1_2-model $\mathcal{N}$ of $\Diamond_k (A \wedge C)^K$. By the usual properties of interpretations, our model $\mathcal{N}$ also satisfies $\Diamond_{A,\mathsf{max}(\rho(A),\rho(C))} C$. So $\mathcal{M} \models \blacklozenge_{\mathsf{S}^1_2} \triangledown_A C$. ❑

Philosophy, Faculty of Humanities, Utrecht University, Janskerkhof 13, 3512BL Utrecht, The Netherlands

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