FEASIBLE INTERPRETABILITY

Rineke Verbrugge

ITLI Prepublication Series
for Mathematical Logic and Foundations ML-91-11
The ITLI Prepublication Series

1986 86-01  The Institute of Language, Logic and Information
86-02 Peter van Emde Boas  A Semantical Model for Integration and Modularization of Rules
86-03 Johan van Benthem  Categorial Grammar and Lambda Calculus
86-04 Reinhard Muskens  A Relational Formation of the Theory of Types
86-05 Kees Ketelaar, Dick de Jongh  Some A Priori Logic for Branching Time, Part I
86-06 Johan van Benthem  Well-founded Time, Forward looking Operators
86-07 Johan van Benthem  Logical Syntax
87-01 Jeroen Groenendijk, Martin Stokhof  Theorie Type shifting Rules and the Semantics of Interrogatives
87-02 Renate H. Neuenschwander  Models Representations and Discourse Representation
87-03 Jan Willem Kloop, Roel de Vrijer  Unique Normal Forms for Lambda Calculus with Surjective Pairing
87-04 Johan van Benthem  Polyadic quantifiers
87-05 Victor Sánchez Valencia  Traditional Logicians and de Morgan's Example
87-06 Eleonora Oversteegen  Temporal Adverbials in the Two Track Theory of Time
87-07 Johan van Benthem  Categorial Grammar and Type Theory
87-08 Renate Bartsch  The Construction of Properties under Perspectives
87-09 Hans J. Ferwerda  Change in Semantics: The Scope of Quantification and Coordination
1988 LP-88-01 Michiel van Langen  Algorithmic Information Theory
LP-88-02 Yde Venema  Expressiveness and Completeness of an Interval Tense Venga
LP-88-03  Year Report 1987
LP-88-04 Reinhard Muskens  Going partial in Montague Grammar
LP-88-05 Johan van Benthem  Logical Constants across Varying Types
LP-88-06 Johan van Benthem  Semantic Parallels in Natural Language and Computation
LP-88-07 Johan van Benthem  Tenes, Aspects, and their Scoops in Discourse
LP-88-08 Jeroen Groenendijk, Martin Stokhof  Context and Information in Dynamic Semantics
LP-88-09 Theo M.V. Janssen  A mathematical model for the CAT framework of Eurotra
LP-88-10 Anneke Kiepe  A Blissymbols Translation Program
ML-88-02 M.G. Swen  The Arithmetical Fragment of Martin Löf's Type Theories with weak $\Sigma$-elimination
ML-88-03 Dick de Jongh, Frank Veltman  Provability Logics for Relative Interpretability
ML-88-04 A.S. Troelstra  On the Early History of Intuitionistic Logic
ML-88-05 J.A. M. Krom  Truths on Intuitionism and the Philosophy of Mathematics
CT-88-01 Ming Li, Paul M.B. Vantini  Computation and Complexity Theory: Two Decades of Applied Kolmogorov Complexity
CT-88-02 Michiel H.M. Smid  General Lower Bounds for the Partitioning of Range Trees
CT-88-03 Michiel H.M. Smid, Mark H. Overmars, Leen Torenvliet, Peter van Emde Boos  Maintaining Multiple Representations of Dynamic Data Structures
CT-88-04 Dick de Jongh, Lex Hendriks, Gerard R. Renardel de Lavalette  Computations in Fragments of Intuitionistic Propositional Logic
CT-88-05 Peter van Emde Boos  Machine Models and Simulations (revised version)
CT-88-06 Dick de Jongh  A Data Structure for the Union-find Problem having good Single-Operation Complexity
CT-88-07 Johan van Benthem  Time, Logic and Computation
CT-88-08 Michiel H.M. Smid, Mark H. Overmars, Leen Torenvliet, Peter van Emde Boos  Multiple Representations of Dynamic Data Structures
CT-88-09 Dick de Jongh, Frank Veltman  Towards a Universal Parsing Algorithm for Functional Grammar
CT-88-10 Edith Spaan, Leen Torenvliet, Peter van Emde Boos  Nondeterminism, Fairness and a Fundamental Analogy
CT-88-11 Sieger van Dennehuevel, Peter van Emde Boos  Towards implementing RL
X-88-01 Marc Jumelet  Other prepublications: On Solovay's Completeness Theorem
1989 LP-89-01 Johan van Benthem  Logic, Semantics and Philosophy of Language: The Fine-Structure of Categorial Semantics
LP-89-02 Jeroen Groenendijk, Martin Stokhof  Dynamic Predicate Logic: towards a compositional, non-representational semantics of discourse
LP-89-03 Yde Venema  Two-dimensional Modal Logics for Relation Algebras and Temporal Logic of Intervals
LP-89-04 Johan van Benthem  Language in Action
LP-89-05 Johan van Benthem  Modal Logic as a Theory of Information
LP-89-06 Andrej Peletier  Intensional Lambe Calculi: Theory and Application
LP-89-07 Heinrich Wansing  The Adequacy Problem for Sequential Propositional Logic
LP-89-08 Victor Sánchez Valencia  Peirce's Propositional Logic: From Algebra to Graphs
ML-89-01 Dick de Jongh, Albert Visser  Mathematical Logic and Foundations: Explicit Fixed Points for Interpretability Logic
LP-89-02 Roel de Vrijer  Extending the Lambda Calculus with Surjective Pairing is conservative
LP-89-03 Dick de Jongh, Franco Montagna  Rosser Orderings and Free Variables
ML-89-04 Dick de Jongh, Marc Jumelet, Franco Montagna  On the Proof of Solovay's Theorem
LP-89-05 Rinke Verbrugge  $\Sigma$-completeness and Bounded Arithmetic
ML-89-06 Michiel van Langen  The Axiomatization of Randomness
ML-89-07 Dirk Roorda  Elementary Inductive Definitions in HA: from Strictly Positive towards Monotone
ML-89-08 Dirk Rijke  Investigations into Classical Linear Logic
ML-89-09 Alessandra Carbone  Provably fixed points in $\Lambda\Omega+\Omega$
CT-89-01 Michiel H.M. Smid  Computation and Complexity Theory: Dynamic Deferred Data Structures
CT-89-02 Peter van Emde Boos  Machine Models and Simulations
CT-89-03 Jan van Eijck, Herman Neefuželje, Leen Torenvliet, Peter van Emde Boos  On Space Efficient Simulations
CT-89-04 Harry Buhman, Leen Torenvliet  A Comparison of Reductions on Nondeterministic Space
CT-89-05 Peter H. Hartel, Michiel H.M. Smid, Leen Torenvliet, Willem G. Vree  A Parallel Functional Implementation of Range Queries
CT-89-06 H.W. Lenstra, Jr.  Finding Isomorphisms between Finite Fields
CT-89-07 Ming Li, Paul M.B. Vantini  A Theory of Learning Simple Concepts under Simple Distributions and Average Case Complexity for the Universal Distribution (Pre. Version)
CT-89-08 Harry Buhman, Steven Homer, Leen Torenvliet, Pieter Hazewinkel  Random Reductions and Nondeterministic Complexity Classes
CT-89-09 Harry Buhman, Edith Spaan, Leen Torenvliet  On Adaptive Resource Bounded Computations
CT-89-10 Sieger van Dennehuevel  The Rule Language RL/1
CT-89-11 Zhiheng Huang, Sieger van Dennehuevel, Peter van Emde Boos  Towards Functional Classification of Recursive Query Processing
X-89-01 Moritz Kalskot  Other Prepublications: An Orey Sentence for Predicative Arithmetic
X-89-02 G. Wagemakers  New Foundations: a Finding of Quine's Set Theory
X-89-03 A.S. Troelstra  Index of the Heyting Nachlass
X-89-04 Jeroen Groenendijk, Martin Stokhof  Dynamic Montague Grammar, a first sketch
X-89-05 Rijke  The Modal Theory of Inequality
X-89-06 Peter van Emde Boas  Een Relationale Semantiek voor Conceptueel Modelleren: Het RL-project
1990 Logic, Semantics and Philosophy of Language
LP-90-01 Jaap van der Does  Generalized Quantifier Logic for Naked Infinities
LP-90-02 Jeroen Groenendijk, Martin Stokhof  Dynamic Montague Grammar
LP-90-03 Renate Bartsch  Concept Formation and Concept Composition
LP-90-04 Aarne Ranta  Intuitionistic Categorial Grammar
LP-90-05 Patrick Blackburn  Nominal Tense Logic
LP-90-06 Gennaro Chierchia  The Variability of Impersonal Subjects
LP-90-07 Gennaro Chierchia  Anaphora and Dynamic Logic
LP-90-08 Herman Hendriks  Flexible Montague Grammar
LP-90-09 The Principal Markers  Models for Discourse Markers
LP-90-10 Johan van Benthem  General Dynamics
LP-90-12 Serge Lapierre  A Functional Partial Semantics for Intensional Logic
FEASIBLE INTERPRETABILITY

Rineke Verbrugge
Department of Mathematics and Computer Science
University of Amsterdam

ITLI Prepublications
for Mathematical Logic and Foundations
ISSN 0924-2090

Received October 1991
Feasible interpretability

Rineke Verbrugge
Faculty of Mathematics and Computer Science
University of Amsterdam
Plantage Muidergracht 24
1018 TV Amsterdam
the Netherlands

October 15, 1991

Abstract

In $PA$, or even in $I\Delta_0 + EXP$, we can define the concept of feasible interpretability. Informally stated, $U$ feasibly interprets $V$ iff:

for some interpretation, $U$ proves the interpretations of all axioms of $V$ by proofs with Gödel numbers of length polynomial in the length of the Gödel numbers of those axioms.

Here both $U$ and $V$ are $\Sigma^b_1$-axiomatized theories.

Many interpretations encountered in everyday mathematics (e.g., the interpretation of Peano arithmetic into $ZF$) are feasible. However, by fixed point constructions we can find theories that are interpretable in $PA$ in the usual sense but not by a feasible interpretation. By making polynomial analogs of the usual proofs, we show that the bimodal interpretability logic $ILM$ is sound for feasible interpretability over the base theory $PA$. Here, $A \triangleright B$ is translated as $PA + A^* \triangleright_f PA + B^*$, where $^*$ is the translation. Moreover, we can prove in $PA$ a polynomial version of Orey's theorem for feasible interpretability. This paves the way for a polynomial adaptation of Berarducci's proof of arithmetical completeness of $ILM$ with respect to $PA$. Thus, we show that $ILM$ is arithmetically sound and complete with respect to feasible interpretability over $PA$.

1 Introduction

In this paper, we investigate a new concept of interpretability – we call it feasible interpretability – in which the complexity of the interpretation is bounded in a certain way. The concept was invented by Albert Visser, who called it effective interpretability in his paper [Vi].

In order to define this concept, we first review the usual definition of interpretability.

Let $U, V$ be two $\Sigma^b_1$-axiomatized theories, where $V$ is axiomatized by the $\Sigma^b_1$-formula $\alpha_V$. An interpretation $K$ of $V$ into $U$ is given as usual by a formula $\delta(x)$ of $L_U$ defining the universe, and a function from the relation and function symbols of $L_V$ to formulas of $L_U$, respecting the original arities. In the sequel we take the image of $=$ to be $=$, though this is not essential for the results. We can extend $K$ in the usual way to map all formulas
\( \varphi \) of \( L_V \) into formulas \( \varphi^K \) of \( L_U \); in fact we can, in an intentionally correct way, \( \Delta^0_1 \)-define in \( I \Delta_0 + \Omega_1 \) a function \( K \) corresponding to this mapping. For ease of reading, we will write \( s^K \) even if \( s \) is a Gödel number. Thus \( U \triangleright V \) can be defined in \( I \Delta_0 + \Omega_1 \) as follows:

\[
I \Delta_0 + \Omega_1 \vdash U \triangleright V \iff \exists K ("K is an interpretation" \land \forall a (\alpha_N(a) \rightarrow \exists p Prf_U(p, a^K))).
\]

Similarly, we would like to define a concept of feasible interpretability, given half-formally as

\[
U \triangleright_f V \iff \exists K \exists P ("K is an interpretation and \ P is a polynomial" \land \\
\forall a (\alpha_N(a) \rightarrow \exists p (|p| \leq P(|a|) \land Prf_U(p, a^K))))
\]

(1)

If we want to formalize this concept, we need an evaluation function for coded polynomials and we need to be able to prove that the \( exp \) of this function is total. We remind the reader that \( exp \) (the values of polynomials in \(|x|\)) corresponds to the values of \( \# \)-terms in \( x \), where \( x \# y = exp(|x| \cdot |y|) \) as defined in Buss[86]. Thus, since there is an evaluation function for formalized terms containing \( \# \) that is provably total in \( I \Delta_0 + EXP \), we see that the formalization of feasible interpretability can be carried out in \( I \Delta_0 + EXP \). We will not carry out the details, and for ease of reading we will keep using the half-formal definition (1).

However, it is clear that the formula \( U \triangleright_f V \) is \( \Sigma^0_2 \). As we know that, for reasonable theories \( U \) extending \( PA \), \( \{ A \mid U \triangleright U + A \} \) is a \( \Pi^0_2 \)-complete predicate, it would be interesting to find out whether \( \{ A \mid U \triangleright_f U + A \} \) is \( \Sigma^0_2 \)-complete. We haven't yet found the answer to this question.

In [Vil], Visser gave proof sketches to show that \( ILM \) is arithmetically sound with respect to feasible interpretability over \( PA \). Moreover, he gave an Orey-Hájek like characterization for feasible interpretability over \( PA^* \), where \( PA^* \) is defined as follows:

\( C \) is an axiom of \( PA^* \) iff \( C \) is the conjunction of the first \( n \) axioms of \( PA \) for some \( n \).

He then surmised that, using this characterization, Berarducci's arguments from [Be 90] could be adapted to show that \( ILM \) is the modal interpretability logic for feasible interpretability over \( PA^* \).

In this paper, we show that \( ILM \) is indeed arithmetically sound and complete with respect to feasible interpretability over \( PA \) itself.

The rest of the paper is organized as follows. In section 2, we show that some well-known interpretations from the contexts of set theory and bounded arithmetic are feasible. For the subsequent sections, the horizon is narrowed down to Peano Arithmetic. Thus we prove in section 3 and section 5 that \( ILM \) is exactly the modal interpretability logic for feasible interpretability over \( PA \). Section 4, meanwhile, gives two counterexamples to show that, for reasonable theories \( U \) extending \( PA \), feasible interpretability over \( U \) is a definitely stricter concept than normal interpretability.

## 2 Feasible interpretations in various settings

For an intuitive introduction to feasible interpretability, it is useful to define feasible interpretability also for settings other than arithmetic. The informal definition is as follows.
$U \models f V$ if and only if there is an interpretation $K$ of $V$ into $U$ which is feasible, i.e. for which there is a polynomial $P$ such that for all axioms $\varphi$ of $V$, there is a proof of length $\leq P(|\varphi|)$ in $U$ of $\varphi^K$.

Here $|\varphi|$ denotes the length of $\varphi$. In this section, we look at some well-known interpretations from different settings and show that they are feasible. As a first remark, it is clear that every interpretation of a finitely axiomatized theory into some other theory is feasible: a constant polynomial, namely the maximum of the lengths of the proofs of the interpreted axioms, suffices. We first prove an easy lemma which can be used to show that many well-known interpretations are feasible.

**Remark 2.1** Of course the definitions of $|\varphi|$ and of the lengths of proofs depend on the setting. For example, it is not always convenient to define $|\varphi|$ as “the length of the binary expression for the Gödel number of $\varphi$”.

However, we have to keep in mind that a few conditions on the definition of the lengths of formulas and proofs are necessary to make lemma 2.2 applicable.

The length of formulas should be defined in such a way that the following conditions hold:

1. $|\neg \psi| \geq |\psi| + 1$,
2. $|\psi \circ \chi| \geq |\psi| + |\chi| + 1$ for $\circ \in \{\land, \lor, \to, \leftrightarrow\}$,
3. $|Qx\psi| \geq |\psi| + 1$ for $Q \in \{\forall, \exists\}$, and
4. for all formulas $\varphi$, $|\varphi| \geq 2$.

The last of these conditions is not necessary, but it just simplifies the computations by allowing us to work with polynomials $P(n)$ of the form $n^d$ only.

Moreover, we suppose that the proof system and the corresponding length of a proof is defined in such a way that applications of $\land$-rules and Modus Ponens do not make the proofs explode to an inordinate length; e.g. we suppose that we do not use a tableau system or a sequent calculus without the cut rule. A sufficient condition is the following.

There is a constant $c$ such that the following conditions hold:

1. if we have a proof of of length $l_A$ of the formula $A$, and a proof of length $l_{A \to B}$ of $A \to B$, then there is a proof of length $\leq l_A + l_{A \to B} + |B|^c$ of the formula $B$; and

2. if we have a proof of length $l_A$ of $A$, a proof of length $l_B$ of $B$ and a proof of length $l_{A \land B \to C}$ of $A \land B \to C$, then we have a proof of length $\leq l_A + l_B + l_{A \land B \to C} + |C|^c$ of the formula $C$.

**Lemma 2.2** Let $L$ be a language and $U$ a theory satisfying the conditions in Remark 2.1. Let $F$ be a function from $L$ into $L_U$ such that

there is a polynomial $P$ such that for all $\varphi \in L$, $|F(\varphi)| \leq P(|\varphi|)$.

Moreover, suppose that $U$ proves the following by proofs of length $\leq P(|\varphi|)$, resp. $\leq P(|\neg \psi|)$, resp. $\leq P(|\psi \circ \chi|)$, resp. $\leq P(|Qx\psi|)$:

1. $F(\varphi)$ for all atomic $\varphi \in L$;
2. \( F(\psi) \rightarrow F(\neg\psi) \) for all \( \psi \in L \);

3. \( F(\psi) \land F(\chi) \rightarrow F(\psi \circ \chi) \) for all \( \psi, \chi \in L \) and \( \circ \in \{\land, \lor, \rightarrow, \leftrightarrow\} \);

4. \( F(\psi) \rightarrow F(Qx\psi) \) for all \( \psi \in L \) and \( Q \in \{\forall, \exists\} \).

Then there is a polynomial \( R \) such that for all \( \varphi \in L, U \vdash \varphi \) by a proof of length \( \leq R(|\varphi|) \).

Proof. Take a constant \( d \geq 2 \) such that

1. for all \( n \geq 2, P(n) \leq n^d \) and

2. for all \( \varphi \in L, |F(\varphi)|^c \leq |\varphi|^d \), where \( c \) is as in Remark 2.1 in the condition on the length of proofs.

Define the polynomial \( R(n) := n^{2d} \). We will prove by induction on the construction of \( \varphi \) that for all \( \varphi \in L, U \vdash F(\varphi) \) by a proof of length \( \leq R(|\varphi|) \).

**Basic step** By the assumption we know that for atomic formulas \( \varphi, U \vdash F(\varphi) \) by a proof of length \( \leq P(|\varphi|) \). But by definition of \( d, P(|\varphi|) \leq |\varphi|^d \leq |\varphi|^{2d} \).

\( \neg \)-step Suppose as induction hypothesis that \( U \vdash F(\psi) \) by a proof of length \( \leq |\psi|^{2d} \). By assumption, \( U \vdash F(\psi) \rightarrow F(\neg\psi) \) by a proof of length \( \leq P(\neg\psi) \leq |\neg\psi|^d \) (where the last inequality holds because of clause 1 of the definition of \( d \)). Therefore by the first clause in the condition on the length of proofs in Remark 2.1, we have \( U \vdash F(\neg\psi) \) by a proof of length \( \leq |\psi|^{2d} + |\neg\psi|^d + |F(\neg\psi)|^c \leq |\psi|^{2d} + |\neg\psi|^d + |\neg\psi|^d \) (where the last inequality holds by clause 2 of the definition of \( d \)). Since we assume that \( |\neg\psi| \geq |\psi| + 1 \), we have \( |\psi|^{2d} + |\neg\psi|^d + |\neg\psi|^d \leq |\neg\psi|^{2d} \) by an easy computation using the binomial theorem and the fact that \( d \geq 2 \). The quantifier steps are analogous to the \( \neg \)-step, so we leave them to the reader.

**Connective step** Let \( \circ \in \{\land, \lor, \rightarrow, \leftrightarrow\} \). Suppose as induction hypothesis that \( U \vdash F(\psi) \) by a proof of length \( \leq |\psi|^{2d} \), and \( U \vdash F(\chi) \) by a proof of length \( \leq |\chi|^{2d} \). By assumption, \( U \vdash F(\psi \circ \chi) \) by a proof of length \( \leq P(|\psi \circ \chi|) \leq |\psi \circ \chi|^d \).

The second clause in the condition on the length of proofs in Remark 2.1 now implies that \( U \vdash F(\psi \circ \chi) \) by a proof of length \( \leq |\psi|^{2d} + |\chi|^{2d} + |\psi \circ \chi|^d + |F(\psi \circ \chi)|^c \leq |\psi|^{2d} + |\chi|^{2d} + |\psi \circ \chi|^d + |\psi \circ \chi|^d \) (where the last inequality holds by clause 2 in the definition of \( d \)).

Since we assume that \( |\psi \circ \chi| \geq |\psi| + |\chi| + 1 \), we can again use the binomial theorem to show that \( |\psi|^{2d} + |\chi|^{2d} + |\psi \circ \chi|^d + |\psi \circ \chi|^d \leq |\psi \circ \chi|^{2d} \), as desired.

QED

**Remark 2.3** When we want to prove that some interpretation \( K \) of \( V \) into \( U \) is feasible, we can often use Lemma 2.2. Suppose all axioms of \( V \) have the form \( \Phi(\psi) \), where \( \Phi \) is a formula scheme. The feature we need in order to apply Lemma 2.2 is the fact that both \( |\Phi(\psi)| \) and \( |\psi^K| \) are polynomial in \( |\psi| \).
As a first example, in which we do not yet need lemma 2.2, we will show that the usual interpretation of $\Delta_0 + \Omega_1$ into $\Delta_0$ by a cut is feasible.

**Theorem 2.4** $\Delta_0 \nrightarrow \Delta_0 + \Omega_1$ by a cut.

Proof. Let $J$ be a cut constructed by Solovay's methods such that $\Delta_0$ proves that $J$ is a cut closed under $+$, $\cdot$, and $\omega_1$. Define $\varphi^J$ to be the formula $\varphi$ with all quantifiers restricted to $J$. It is well-known that $J$ is an interpretation of $\Delta_0 + \Omega_1$ into $\Delta_0$; so to show that it is a feasible interpretation, it suffices to find a polynomial $P$ such that for all $\Delta^0_1$-formulas $\varphi$, the following holds by proofs of length $\leq P(|\varphi|)$:

$$\Delta_0 \vdash (\varphi(0) \land \forall x (\varphi(x) \rightarrow \varphi(Sx)) \rightarrow \forall x \varphi(x))^J$$

First, it is easy to see that there is a polynomial $P_1$ such that for all $\Delta^0_1$-formulas $\varphi$, $\Delta_0 \vdash J(a) \rightarrow (\varphi(a) \leftrightarrow \varphi(a))^J$ and $\Delta_0 \vdash \forall x \varphi \rightarrow (\forall x \varphi)^J$ by proofs of length $\leq P_1(|\varphi|)$. Second, there is a polynomial $P_2$ such that for all $\Delta^0_1$-formulas $\varphi$, the following holds by proofs of length $\leq P_2(|\varphi|)$:

$$\Delta_0 \vdash \forall a \ [\varphi(0) \land \forall x \leq a (\varphi(x) \rightarrow \varphi(Sx)) \rightarrow \forall x \leq a \varphi(x)]$$

In fact one uses only the induction axiom for $\forall x \leq a \varphi(x)$, the fact that $\forall a \forall x (Sx \leq a \rightarrow x \leq a)$, and some predicate logic. Combining $P_1$ with $P_2$, we then find a polynomial $P_3$ such that for all $\Delta^0_1$-formulas $\varphi$, the following holds by proofs of length $\leq P_3(|\varphi|)$:

$$\Delta_0 \vdash (\forall a [\varphi(0) \land \forall x \leq a (\varphi(x) \rightarrow \varphi(Sx)) \rightarrow \forall x \leq a \varphi(x)])^J$$

Now it is easy to find a polynomial $P$ from $P_3$ such that for all $\Delta^0_1$-formulas $\varphi$, the following holds by proofs of length $\leq P(|\varphi|)$:

$$\Delta_0 \vdash [\varphi(0) \land \forall x (\varphi(x) \rightarrow \varphi(Sx)) \rightarrow \forall x \varphi(x)]^J$$

We use only the fact that $\forall a (a \leq a)$ and some predicate logic. Thus, $J$ is a feasible interpretation of $\Delta_0 + \Omega_1$ into $\Delta_0$. QED

Next, we will prove that the usual interpretation of $ZF + V = L$ into $ZF$ is feasible. Because $ZF$ consists of a finite list of axioms plus the schemata of separation and replacement, we can restrict our attention to feasibly proving these schemata relativized to the universe $L$ of constructible sets. We will first prove that the schema of separation relativized to $L$ follows feasibly from the reflection theorem for $L$, and then give a feasible proof of the reflection theorem itself. We will try to follow the elegant proof in terms of closed unbounded collections, which unfortunately becomes much less elegant when forced into the straightjacket of the calculation of lengths. We will not stray far from the straightforward presentation given in [Ku 80], where all details about the constructible universe that we omit here can be found. The length $|\varphi|$ of a formula $\varphi$ of $ZF$ is defined as the number of appearances of symbols in $\varphi$; without loss of generality, we can take the length of all variables to be 1. Likewise, we define the length of a proof in $ZF$ to be the total number of symbols appearing in the proof. In the following lemmas, quantifiers in greek letters range over the ordinals, while those in roman letters range over all sets.

The next lemma corresponds to lemma IV.2.5 of [Ku 80].
Lemma 2.5 ZF proves the following by proofs of length polynomial in $|\varphi|$

$$\forall z, \overline{v} \in L \{ x \in z \mid \varphi^L(x, z, \overline{v}) \} \in L \rightarrow \\
\forall z, \overline{v} \in L \exists y \in L \{ x \in y \leftrightarrow x \in z \land \varphi^L(x, z, \overline{v}) \}$$

Proof. Straightforward; we do not even need the fact that $L$ is transitive. Note that by absoluteness of atomic formulas for $L, V$, the succedent is feasibly equivalent to the comprehension schema for $\varphi$, relativized to $L$. QED

The following lemma corresponds to a part of lemma VI.2.1 of [Ku 80].

Lemma 2.6 ZF proves the following by proofs of length polynomial in $|\varphi|$

$$\forall \alpha \exists \beta \forall z, x, \overline{v} \in L_{\beta} \left[ \varphi^L(x, z, \overline{v}) \leftrightarrow \varphi^{L_{\beta}}(x, z, \overline{v}) \right] \rightarrow \\
\forall z, \overline{v} \in L \{ x \in z \mid \varphi^L(x, z, \overline{v}) \} \in L$$

Proof. It is easy to see that the usual proof in ZF is feasible: suppose

1. $\forall \alpha \exists \beta \forall z, x, \overline{v} \in L_{\beta} \left[ \varphi^L(x, z, \overline{v}) \leftrightarrow \varphi^{L_{\beta}}(x, z, \overline{v}) \right]$ and

2. $z, \overline{v} \in L$

From 2 it follows that there is an $\alpha$ such that $z, \overline{v} \in L_{\alpha}$. Now let $\beta > \alpha$ be such that $\forall x \in L_{\beta} \left[ \varphi^L(x, z, \overline{v}) \leftrightarrow \varphi^{L_{\beta}}(x, z, \overline{v}) \right]$. Then, using the fact that $L$ is transitive and that $x \in z$ is absolute for $L_{\beta}, L$, we find that

$$\{ x \in z \mid \varphi^L(x, z, \overline{v}) \} = \{ x \in L_{\beta} \mid (x \in z \land \varphi(x, z, \overline{v})) \} \in Def(L_{\beta}) = L_{\beta+1},$$

so $\{ x \in z \mid \varphi^L(x, z, \overline{v}) \} \in L$. QED

From lemma 2.5 and lemma 2.6, we conclude that in order to feasibly prove the comprehension schema, we only need polynomial length proofs of

$$\forall \alpha \exists \beta \forall z, x, \overline{v} \in L_{\beta} \left[ \varphi^L(x, z, \overline{v}) \leftrightarrow \varphi^{L_{\beta}}(x, z, \overline{v}) \right].$$

For a proof of this reflection theorem, we need a few more definitions.

Definition 2.7 A collection $C$ of ordinals is

- *unbounded* iff $\forall \alpha \exists \beta > \alpha (\beta \in C)$;
- *closed* iff $\forall \alpha (\alpha \not= 0 \land a \subseteq C \rightarrow \sup a \in C)$;
- *closed unbounded (c.u.b.)* iff $C$ is both closed and unbounded.

Lemma 2.8 $ZF \vdash \text{"If } C \text{ and } D \text{ are c.u.b., then } C \cap D \text{ is c.u.b. as well"}$

Proof. An easy application of lemma II.6.8 of [Ku 80]. QED

Definition 2.9 A collection $C$ of ordinals is *closed unbounded for $\varphi$* iff
1. $C$ is closed unbounded, and

2. $C$ consists of ordinals $\alpha$ such that $L_\alpha$ reflects $\varphi$, i.e.
   \[ \forall \alpha \ (\alpha \in C \rightarrow \forall \bar{v} \in L_\alpha \ [\varphi^L(\bar{v}) \leftrightarrow \varphi^{L_\alpha}(\bar{v})]) \]

Suppose $\varphi$ is a formula and $D$ is a first-order definable collection of ordinals. Using definition 2.7, we are able to construct new first-order formulas $CUB_D$, $CUB_{D,\varphi}$ and $REF_{\varphi}$ with the following intended meanings:

1. $CUB_D := \ "D \text{ is closed unbounded}"$

2. $CUB_{D,\varphi} := \ "D \text{ is closed unbounded for } \varphi"$

3. $REF_{\varphi} := \ "\text{there is some collection of ordinals that is closed unbounded for } \varphi"$

The next lemma roughly corresponds to theorem IV.7.5 of [Ku 80].

Lemma 2.10 (Reflection theorem) $ZF$ proves the following by proofs of length polynomial in $|\varphi|$: 

\[ \forall \alpha \exists \beta \forall z, x, \bar{v} \in L_\beta \ [\varphi^L(z, x, \bar{v}) \leftrightarrow \varphi^{L_\beta}(z, x, \bar{v})] \]

Proof. First we note that $ZF$ proves ($\alpha < \beta \rightarrow L_\alpha \subseteq L_\beta$), "if $\gamma$ is a limit ordinal, then $L_\gamma = \bigcup_{\alpha < \gamma} L_\alpha" \text{ and } L = \bigcup_{\alpha < \omega} L_\alpha.$

We will prove the reflection theorem by induction on the construction of $\varphi$. A straightforward application of lemma 2.2 implies that for the reflection theorem to have a proof of length polynomial in $|\varphi|$, it is sufficient to find a polynomial bounding the lengths of the induction steps. Thus, we need to find a polynomial $P$ such that by proofs of length $\leq P(|\varphi|)$, resp. $\leq P(|\neg \psi|)$, resp. $\leq P(|\psi \circ \chi|)$, resp. $\leq P(|Qz\psi|)$, ZF proves the following:

1. for atomic $\varphi$:
   \[ \forall \alpha \exists \beta > \alpha \forall z, x \in L_\beta \ [\varphi^L(z, x) \leftrightarrow \varphi^{L_\beta}(z, x)] \wedge CUB_{OR,\varphi} \]

2. the $\neg$-step:
   \[ \forall \alpha \exists \beta > \alpha \forall \bar{v} \in L_\beta \ [\psi^L(\bar{v}) \leftrightarrow \psi^{L_\beta}(\bar{v})] \wedge REF_\psi \rightarrow \forall \alpha \exists \beta > \alpha \forall \bar{v} \in L_\beta \ [\neg \psi^L(\bar{v}) \leftrightarrow \neg \psi^{L_\beta}(\bar{v})] \wedge REF_{\neg \psi} \]

3. the connective step, where $\circ \in \{\wedge, \vee, \rightarrow, \leftrightarrow\}$:
   \[ \forall \alpha \exists \beta > \alpha \forall \bar{v} \in L_\beta \ [\psi^L(\bar{v}) \leftrightarrow \psi^{L_\beta}(\bar{v})] \wedge REF_\psi \wedge \forall \alpha \exists \beta > \alpha \forall \bar{v} \in L_\beta \ [\chi^L(\bar{v}) \leftrightarrow \chi^{L_\beta}(\bar{v})] \wedge REF_\chi \rightarrow \forall \alpha \exists \beta > \alpha \forall \bar{v} \in L_\beta \ [\psi^L \circ \chi^L(\bar{v}, \bar{w}) \leftrightarrow \psi^{L_\beta} \circ \chi^{L_\beta}(\bar{v}, \bar{w})] \wedge REF_{\psi \circ \chi} \]

4. the quantifier step, where $Q \in \{\exists, \forall\}$:
   \[ \forall \alpha \exists \beta > \alpha \forall z, \bar{v} \in L_\beta \ [\psi^L(z, \bar{v}) \leftrightarrow \psi^{L_\beta}(z, \bar{v})] \wedge REF_\psi \rightarrow \forall \alpha \exists \beta > \alpha \forall \bar{v} \in L_\beta \ [Qz \in L \psi^L(z, \bar{v}) \leftrightarrow Qz \in L_\beta \psi^{L_\beta}(z, \bar{v})] \wedge REF_{Qz_\psi} \]
Finding polynomials bounding the lengths of the proofs of 1, 2 and 3 is very easy: we can use the feasibly provable fact that atomic formulas are absolute for any $L_\alpha, L$, some propositional reasoning independent on the specific $\psi, \chi$, and an application of lemma 2.8 for step 3. We will show how the proofs of the $\exists$-case in step 4 can be bounded by a polynomial; a bound for the $\forall$-step then follows by some uses of the bounds for the $\neg$-step and the $\exists$-step.

Define

$$D := \{ \beta \mid \forall \bar{v} \in L_\beta \ [\exists z \in L \psi^L(z, \bar{v}) \rightarrow \exists z \in L_\beta \ psi^L(z, \bar{v})] \}. $$

It is easy to see that $ZF$ proves the following by proofs of length polynomial in $|\exists z \psi|$

$$\forall \alpha \exists \beta > \alpha \forall z, \bar{v} \in L_\beta [\psi^L(z, \bar{v}) \leftrightarrow \psi^L(z, \bar{v})] \wedge C U B_{C_\psi}, \psi \wedge C U B_D \rightarrow
\forall \alpha \exists \beta > \alpha \forall \bar{v} \in L_\beta [\exists z \in L \psi^L(z, \bar{v}) \leftrightarrow \exists z \in L_\beta \ psi^L(z, \bar{v})] \wedge C U B_{C_\psi} \wedge D, \exists z \psi
$$

In fact, we only use lemma 2.8 and the fact that $\forall \beta (L_\beta \subseteq L)$. Thus we need to find a polynomial $P$ such that $ZF \vdash C U B_D$ by a proof of length $\leq P(|\exists z \psi|)$. Immediately from the definition, it is clear that $ZF \vdash "D is closed"$ by a proof of length polynomial in $|\exists z \psi|$. Thus, it suffices to show by a proof of length polynomial in $|\exists z \psi|$ that $ZF$ proves that $D$ is unbounded, that is:

$$\forall \alpha \exists \beta > \alpha \forall \bar{v} \in L_\beta [\exists z \in L \psi^L(z, \bar{v}) \rightarrow \exists z \in L_\beta \ psi^L(z, \bar{v})],
$$

i.e.

$$\forall \alpha \exists \beta > \alpha \forall \bar{v} \in L_\beta \exists z \in L_\beta [\exists z \in L \psi^L(z, \bar{v}) \rightarrow \psi^L(z, \bar{v})].$$

We will reason in $ZF$, taking care that all steps are applications of general $ZF$-theorems that do not depend on the specific formula $\psi$. Take any ordinal $\alpha$. We know using only predicate logic that

$$\forall \bar{v} \in L_\alpha \exists z \in L [\exists z \in L \psi^L(z, \bar{v}) \rightarrow L\psi^L(z, \bar{v})];$$

therefore,

$$\forall \bar{v} \in L_\alpha \exists \alpha_{\bar{v}} (\alpha_{\bar{v}} = \{ \beta > \alpha \mid \exists z \in L_\beta [\exists z \in L \psi^L(z, \bar{v}) \rightarrow \psi^L(z, \bar{v})] \}).$$

by the unrelativized replacement and union axioms, there is a $\beta_1$ such that $\beta_1 = \sup \{ \alpha_{\bar{v}} | \bar{v} \in L_\alpha \}$. Continuing in this way, we can define by recursion a sequence $\beta_p$ for $p \in \omega$, where for all $p \in \omega$,

$$\forall \bar{v} \in L_{\beta_p} \exists z \in L_{\beta_{p+1}} [\exists z \in L \psi^L(z, \bar{v}) \rightarrow \psi^L(z, \bar{v})] \tag{2}$$

Define $\beta := \sup \{ \beta_p \mid p \in \omega \}$. Because $\alpha = \beta_0 < \beta_1 < \beta_2 < \ldots$, we infer that $\beta$ is a limit ordinal $> \alpha$. Now using (2) and the fact that $L_\beta = \bigcup_{\gamma \subseteq \beta} L_\gamma$, we find that

$$\forall \bar{v} \in L_\beta \exists z \in L_\beta [\exists z \in L \psi^L(z, \bar{v}) \rightarrow \psi^L(z, \bar{v})],$$

as desired. QED
Lemma 2.11 For all $\varphi$, ZF feasibly proves the comprehension schema for $\varphi$, relativized to $L$; i.e. by proofs of length polynomial in $|\varphi|$, ZF proves the following:

$$\forall z, \overline{v} \in L \exists y \in L \forall x \in L \left[ x \in y \leftrightarrow x \in z \land \varphi^L(x, z, \overline{v}) \right]$$

Proof. Combine lemmas 2.5, 2.6 and 2.10. QED

Lemma 2.12 For all $\varphi$, ZF feasibly proves the replacement schema for $\varphi$, relativized to $L$; i.e. by proofs of length polynomial in $|\varphi|$, ZF proves the following:

$$\forall a, \overline{v} \in L \left[ \forall x \in a \exists y \in L \varphi^L(x, y, \overline{v}) \rightarrow \exists c \in L \forall y \in L \left( y \in c \leftrightarrow \exists x \in a \varphi^L(x, y, \overline{v}) \right) \right]$$

Proof. We already have feasible proofs of the relativized comprehension schema for the formula $y \in b \land \exists x \in a \varphi(x, y, \overline{v})$. So we can (feasibly) prove that it suffices to show the following by proofs of length polynomial in $|\varphi|$:

$$ZF \vdash \forall a, \overline{v} \in L \left[ \forall x \in a \exists y \in L \varphi^L(x, y, \overline{v}) \rightarrow \exists b \in L \left( \forall x \in a \exists y \in L \varphi^L(x, y, \overline{v}) \right) \right]$$

The last proof works, as in lemma 2.10, by general theorems of ZF that do not depend on the specific $\varphi$. Work in $ZF$ and suppose $a, \overline{v} \in L$ and $\forall x \in a \exists y \in L \varphi^L(x, y, \overline{v})$. Now

$$\forall x \in a \exists y \in L \left( \beta_x = \bigcap \{ \alpha \mid \exists y \in L \varphi^L(x, y, \overline{v}) \} \right);$$

then by replacement and the union axiom we find $\beta$ such that $\beta = \bigcup \{ \beta_x \mid x \in a \}$, and we let $b$ be $L_\beta$. Then

$$\forall x \in a \exists y \in b \varphi^L(x, y, \overline{v}).$$

QED

Contrary to our expectations, the usual interpretation of $ZF + V \neq L$ into $ZF(M)$ (by forcing with generic extensions), although much more complex, is still feasible. We checked this following the lines of the proof in [Ku 80]. Our proof relies so heavily on the many details of Kunen’s proof, that it would be incomprehensible to readers not conversant with that book. Therefore, we do not give it here.

In the literature there are also proofs of $ZF \not\models ZF + V \neq L$ and $ZF + AC \not\models ZF + AC + \neg CH$ which entirely avoid the use of the transitive countable collection $M$. A sketch of such a proof can be found in [Co 66, Section IV.11], and a completely different full proof in [VH 72, Ch. V, VI]. It appears that these proofs can also be analyzed to show that the interpretations in question are feasible.

Other well-known interpretations, such as the one of $PA$ into $ZF$, are also feasible, as the reader may check for her/himself. All in all it seems that the only examples of theories $U$ and $V$ such that $U \models V$ but not $U \not\models V$ are contrived theories obtained by fixed-point constructions like the ones in section 4. It would be nice to find a more natural counterexample.

It would also be interesting to investigate severely restricted kinds of interpretability which do distinguish between interpretations used in everyday mathematics. For example, one could restrict the complexity of formulas allowed to occur in the proofs of the interpreted axioms.
Sam Buss suggested the following restricted definition of feasible interpretability to us:

\[
U \vdash_{fm} V \iff \exists K \exists M \left( \text{"K is an interpretation and M is a deterministic polynomial time Turing Machine"} \land \forall a (\alpha_V(a) \rightarrow \text{Prf}_U(M(a), a^K)) \right). \tag{3}
\]

This definition is more in line with the conventional use of the word "feasible" in the context of polynomial time computability. The clause \(\text{Prf}_U(M(a), a^K)\) in (3) is a \(P\)-like formula, while the clause \(\exists p \left( |p| \leq P(|a|) \right) \land \text{Prf}_U(p, a^K)\) in the definition of feasible interpretability used in this paper is an \(NP\)-like formula. However, all interpretations considered in this section can also be shown to be feasible in Buss's sense: we only need an easy analogue of lemma 2.2.

3 Soundness of ILM for feasible interpretability over PA

In this section, we restrict our attention to feasible interpretability over \(PA\). We show that the modal interpretability logic \(ILM\) is \(PA\)-sound even if the intended meaning of \(A \triangleright B\) is "\(PA + A\) feasibly interprets \(PA + B\)".

**Definition 3.1** The modal interpretability logic \(ILM\) contains, besides all formulas having the form of a propositional tautology, the usual axioms for the provability logic \(L\) and the rules modus ponens and necessitation, the following axioms:

\[
\begin{align*}
\text{J1} & \quad \Box (A \rightarrow B) \rightarrow (A \triangleright B) \\
\text{J2} & \quad (A \triangleright B) \land (B \triangleright C) \rightarrow (A \triangleright C) \\
\text{J3} & \quad (A \triangleright C) \land (B \triangleright C) \rightarrow (A \lor B \triangleright C) \\
\text{J4} & \quad (A \triangleright B) \rightarrow (\Diamond A \rightarrow \Diamond B) \\
\text{J5} & \quad \Diamond A \triangleright A \\
\text{M} & \quad (A \triangleright B) \rightarrow (A \land \Box C \triangleright B \land \Box C)
\end{align*}
\]

**Definition 3.2** A feasibility interpretation is a map \(\ast\) which assigns to every propositional variable \(p\) a sentence \(p^\ast\) of the language of \(PA\), and which is extended to all modal formulas as follows:

1. \((A \triangleright B)^\ast = PA + A^\ast \triangleright_f PA + B^\ast\)
2. \((\Box A)^\ast = \text{Prov}_{PA}(A^\ast)\)
3. \(\ast\) commutes with the propositional connectives.

Here \(\triangleright_f\) abbreviates the formalization of feasible interpretability.

We will prove that \(ILM\) is arithmetically sound for feasible interpretability, i.e. that for all modal formulas \(A\), if \(ILM \vdash A\), then for all feasibility interpretations \(\ast\), \(PA \vdash A^\ast\). Thus, we have to check that the axioms J1 to J5 are valid in \(PA\) when \(A \triangleright B\) is read as \(PA + A \triangleright_f PA + B\). Whenever possible, we will prove generalizations of these axioms to theories \(U, V \supseteq PA\). Also we prove a generalization of the property \(M\), where an infinite set of \(\Sigma^0_1\)-sentences can be added on both sides instead of one \(\Box\)-sentence only.
Lemma 3.3 PA proves all feasibility translations of J1 to J5.

Proof. The proofs for J1 through J4 can be found almost verbatim in [Vi]. We reason in PA.

J1 Suppose for some theory V and some p that Prf\(_V\)(p, A\(^\circ\)). Then by the identity interpretation and the polynomial bound P(n) = n + 3 \cdot |p|, V \(\triangleright\) V + A. So in particular, if \(\square_{PA}(A \rightarrow B)\), then PA + A \(\triangleright\) APA + A + B, and surely PA + A \(\triangleright\) APA + B.

J2 Suppose

- U \(\triangleright\) V by interpretation K\(_1\) and polynomial P\(_1\), and
- V \(\triangleright\) W by the interpretation K\(_2\) and polynomial P\(_2\).

As in the usual case, U \(\triangleright\) W by the interpretation K\(_2\) \circ K\(_1\). We need to show that there is a polynomial bound for the proofs of the translated axioms. So let b code an axiom of W, and p a proof in V of b\(^{K_2}\) with |p| \(\leq\) P\(_2\)(|b|).

If we take the K\(_1\)-translations of all formulas appearing in the proof coded by p, and add some intermediate steps, we can construct a U-proof of (b\(^{K_2}\))\(^{K_1}\) from K\(_1\)-translations of axioms of V as assumptions; this proof will be of length \(\leq k \cdot |p|\), where k is a constant depending on the translation K\(_1\). Now we only have to add proofs of the translated V-axioms; the axioms themselves have codes of length \(\leq |p|\), so their K\(_1\)-translations have proofs with codes of length \(\leq P_1(|p|) \leq P_1(P_2(|b|))\).

All in all, even in the worst case where the U-proof of (b\(^{K_2}\))\(^{K_1}\) consists wholly of assumptions, there is a q with |q| \(\leq k \cdot P_2(|b|) \cdot P_1(P_2(|b|))\) such that Prf\(_U\)(q, (b\(^{K_2}\))\(^{K_1}\)).

In particular, if PA + A \(\triangleright\) APA + B and PA + B \(\triangleright\) APA + C, then PA + A \(\triangleright\) APA + C.

J3 Suppose

- U + A \(\triangleright\) V by interpretation K\(_1\) and polynomial P\(_1\), and
- U + B \(\triangleright\) V by interpretation K\(_2\) and polynomial P\(_2\).

As in the usual case, we have U + A \(\lor\) B \(\triangleright\) V by the disjunctive interpretation M which equals K\(_1\) in case A holds and equals K\(_2\) in case \(\neg\)A holds. To find a polynomial bound, we observe that for all C, \(\vdash A \rightarrow (C^M \leftrightarrow C^{K_1})\) and \(\vdash \neg A \rightarrow (C^M \leftrightarrow C^{K_1})\) by proofs of length \(\leq P(|C|)\), where the polynomial P depends on K\(_1\) and K\(_2\). Now suppose that c codes an axiom of V, that p\(_1\) codes a U + A-proof of c\(^{K_1}\) with |p\(_1\)| \(\leq P_1(|c|)\), and that p\(_2\) codes a U + B-proof of c\(^{K_2}\) with |p\(_2\)| \(\leq P_2(|c|)\). But then there is a constant k such that

- we can find p\(_1\)' such that Prf\(_U\)(p\(_1\)', A \(\rightarrow\) c\(^M\)) with |p\(_1\)'| \(\leq P(|c|) + P_1(|c|) + k \cdot |c|\); and
- we can find p\(_2\)' such that Prf\(_U\)(p\(_2\)', \(\neg\)A \(\land\) B \(\rightarrow\) c\(^M\)) with |p\(_2\)'| \(\leq P(|c|) + P_2(|c|) + k \cdot |c|\).

Combining p\(_1\)' and p\(_2\)' and their respective polynomial bounds, we find p and P\(_p\)' such that Prf\(_U\)(p, A \(\lor\) B \(\rightarrow\) c\(^M\)) with |p| \(\leq P'(|c|)\). In particular, we have: if PA + A \(\triangleright\) APA + C and PA + B \(\triangleright\) APA + C, then PA + A \(\lor\) B \(\triangleright\) APA + C.
Because \((PA + A \not\vdash_f PA + B) \rightarrow (PA + A \not\vdash PA + B)\), we have by the soundness of J4 for normal interpretability immediately \((PA + A \not\vdash_f PA + B) \rightarrow (\Box A \rightarrow \Box B)\).

In an easier variation of lemma 5.11, we use a claim proved in [Vi 89], which is stated in this paper as lemma 5.10. Suppose \(\beta\) is a \(\Sigma^b_1\)-formula axiomatizing a subset \(U\) of a \(\Sigma^b_1\)-language \(L\). We will prove that \(Q + \Box_\beta \top \not\vdash_f U\) i.e. \(Q + \Box_U \top \not\vdash_f U\).

By lemma 5.10, we have
\[
PA \vdash \Box_{Q + \text{Con}(\beta)} \text{Con}(\beta) \rightarrow \exists \forall \alpha \in \text{Sent}(L) \text{Polprov}_{Q + \text{Con}(\beta), \alpha} (r \Box \alpha \rightarrow \gamma \alpha^K).
\]

Of course we also know that \(PA \vdash \Box_{Q + \text{Con}(\beta)} \text{Con}(\beta)\), so
\[
PA \vdash \exists \forall \alpha \in \text{Sent}(L) \text{Polprov}_{Q + \text{Con}(\beta), \alpha} (r \Box \alpha \rightarrow \gamma \alpha^K).
\]

On the other hand, we have by provable \(\Sigma^b_1\)-completeness
\[
PA \vdash \forall \alpha (\beta(\alpha) \rightarrow \text{Polprov}_{Q + \text{Con}(\beta), \alpha} (r \Box \alpha \gamma)).
\]

Combining the last two results, we have
\[
PA \vdash \exists \forall \alpha \in (\beta(\alpha) \rightarrow \text{Polprov}_{Q + \text{Con}(\beta), \alpha} (\gamma \alpha^K)),
\]
so \(PA \vdash (Q + \Box_\beta \top) \not\vdash_U\). In particular, we have for any sentence \(A\):
\[
PA \vdash (Q + \Box_PA A) \not\vdash_f PA + A,
\]
so especially
\[
PA \vdash PA + \Box_PA A \not\vdash_f PA + A.
\]

QED

We want to prove that Montagna’s property \(M\) holds for feasible interpretability over \(PA\) in its general version, where we can add an infinite set of \(\Sigma^b_1\)-sentences on both sides. In order to ensure that the usual arguments can indeed be polynomialized, we do not formulate the proof in the usual model-theoretic way, and we give many details that are not given in most proofs of Montagna’s property for normal interpretability over \(PA\). The example we give in theorem 4.1 of a set \(S\) of formulas such that \(PA \vdash PA \not\vdash PA + S\) but \(\omega \not\not\vdash PA \not\vdash_f PA + S\) also relies heavily on these details.

Suppose \(U \supseteq PA\), \(V \supseteq PA\). Now suppose \(U \not\vdash_f V\) by the interpretation \(K\) (preserving \(=\)) with domain \(\delta\), and polynomial \(P\). We want to find a polynomial \(Q\) such that for every \(\Sigma^b_1\)-sentence \(\sigma\) there is a \(U + \sigma\)-proof \(p\) of \(\sigma^K\) with \(|p| \leq Q(|\sigma|)\). First, we need some definitions and lemmas. Fix \(U, V, K, P\) as given above.

**Definition 3.4** Define \(pism(s)\) for “\(s\) is a partial isomorphism” and the function \(G(j, y)\) as follows:

\[
pism(s) \quad := \quad \text{seq}(s) \land (s)_0 = 0^K \land \forall i < \text{lh}(s) - 1(s)_{i+1} = S^K(s)_i
\]

\[
G(j, y) \quad := \quad \exists s(pism(s) \land \text{lh}(s) = j + 1 \land (s)_j = y)
\]

12
Lemma 3.5 $U \vdash \forall j \exists ! s (\text{pism}(s) \land \text{lh}(s) = j + 1)$ and thus $U \vdash \forall j \exists y G(j,y)$. Therefore, there is a function $g$ corresponding to $G$.

Proof. By induction. QED

Lemma 3.6 $U$ proves that $g$ is injective, and $U \vdash \forall j \forall y (G(j,y) \to \delta(y))$.

Proof. By induction. QED

Lemma 3.7 $U$ proves that $g$ preserves $0, S, +, \cdot$, and $\leq$.

Proof. We will give some of the preservation proofs. It follows immediately from the definition of $\text{pism}(s)$ that $U \vdash g(0) = 0^K$ and $U \vdash \forall x (g(Sx) = S^K(g(x)))$.

We now prove by induction that $g$ preserves $+$; the other proofs are analogous. We have $U \vdash g(x + 0) = g(x) = g(x) + K 0^K = g(x) + K g(0)$ and $U \vdash g(x + y) = g(x) + K g(y) \to g(x + Sy) = g(S(x + y)) = S^K(g(x + y)) = S^K(g(x) + K g(y)) = g(x) + K S^K(g(y)) = g(x) + K g(Sy)$, So by induction (with $x$ as parameter) $U \vdash \forall x \forall y (g(x + y) = g(x) + K g(y))$.

QED

Lemma 3.8 The range of $g$ is ‘closed downwards’, i.e. $U \vdash \forall x \forall y (\delta(u) \land u < K g(x) \to \exists y < x(u = g(y)))$.

Proof. Before we start the proof proper, we note a useful fact. $V$ includes $PA$ and $K$ is an interpretation of $V$ into $U$. Thus, as

1. $PA \vdash \forall x \forall y (u < x + 1 \to u < x \lor u = x)$ and

2. $U \vdash \forall x (g(x) + K 1^K = g(x + 1))$, we also have

3. $U \vdash \forall x \forall y (\delta(u) \land u < K g(x + 1) \to u < K g(x) \lor u = g(x))$.

Now we can start with the proof by induction on $x$ of $U \vdash \forall x \forall y (\delta(u) \land u < K g(x) \to \exists y < x(u = g(y)))$.

$x = 0$ We have $U \vdash \neg \exists u (\delta(u) \land u < K g(0))$, so $U \vdash \forall u (\delta(u) \land u < K g(0) \to \exists y < 0(u = g(y)))$.

Induction step Work in $U$ and suppose $\forall u (\delta(u) \land u < K g(x) \to \exists y < x(u = g(y)))$ (induction hypothesis). Moreover, suppose $\delta(u) \land u < K g(x + 1)$. Then, by 3, $u < K g(x) \lor u = g(x)$. So by the induction hypothesis $\exists y < x(u = g(y)) \lor u = g(x)$, i.e. $\exists y < x + 1(u = g(y))$.

QED

Remark 3.9 Let $I(x)$ be the formula $\exists y (x = g(y))$. Note that $U$ does not prove that $I$ is closed under successor, so $I$ does not define a cut; but by the previous lemma we do have $U \vdash \forall x \forall y (\delta(u) \land I(x) \land u < K x \to I(u))$. 

13
Lemma 3.10 For all formulas $\varphi \in \Delta_1^0$, $U$ proves the following by proofs of length polynomial in $\lceil \varphi(a_1, \ldots, a_n) \rceil$:

$$\varphi(a_1, \ldots, a_n) \leftrightarrow (\varphi^K)(g(a_1), \ldots, g(a_n)).$$

Proof. By induction on the construction of $\varphi$. We will see below that the proofs for the atomic formulae $\psi$ are obviously of length linear in $\lceil \varphi \rceil$, and that all induction steps follow a given proof scheme in which the particular formulas at hand can be plugged in. So, because every $\varphi$ has at most $\lceil \varphi \rceil$ subformulas, there is a polynomial $R$ such that for all $\varphi$, the $U$-proof of $\varphi(a_1, \ldots, a_n) \leftrightarrow (\varphi^K)(g(a_1), \ldots, g(a_n))$ is of length $\leq R(\lceil \varphi \rceil)$.

We will do the atomic step and the $\forall x \leq t$-step of the proof, and leave the others to the reader.

**Atomic step** By lemma 3.7, we have for all terms $t$ by proofs of length polynomial in $\lceil \varphi \rceil$:

$$U \vdash g(t(a_1, \ldots, a_n)) = (t^K)(g(a_1), \ldots, g(a_n)).$$

So suppose $\varphi$ is the formula $t_1(a_1, \ldots, a_n) = t_2(a_1, \ldots, a_n)$ where $a_1, \ldots, a_n$ include all variables appearing in $t_1$ and $t_2$. Then, because $U$ proves that $g$ is an injective function,

$$U \vdash t_1(a_1, \ldots, a_n) = t_2(a_1, \ldots, a_n) \leftrightarrow g(t_1(a_1, \ldots, a_n)) = g(t_2(a_1, \ldots, a_n)) \leftrightarrow (t_1^K)(g(a_1), \ldots, g(a_n)) = (t_2^K)(g(a_1), \ldots, g(a_n)) \leftrightarrow ((t_1 = t_2)^K)(g(a_1), \ldots, g(a_n))$$

$\forall x \leq t$-step Suppose that $\varphi(a_1, \ldots, a_n) = \forall x \leq t(a_1, \ldots, a_n)\psi(x, a_1, \ldots, a_n)$, and that $U \vdash \psi(x, a_1, \ldots, a_n) \leftrightarrow (\psi^K)(g(x), g(a_1), \ldots, g(a_n))$ (induction hypothesis). We will use the fact that, because of lemmas 3.6, 3.7 and 3.8,

$$U \vdash \forall u, a_1, \ldots, a_n \ (\exists x [x \leq t(a_1, \ldots, a_n) \land u = g(x)] \leftrightarrow \delta(u) \land u \leq K t^K(g(a_1), \ldots, g(a_n))).$$

Thus, we have the following equivalences:

$$U \vdash \varphi(a_1, \ldots, a_n) \leftrightarrow \forall x \leq t(a_1, \ldots, a_n)\psi(x, a_1, \ldots, a_n) \leftrightarrow \forall x \leq t(a_1, \ldots, a_n)(\psi^K)(g(x), g(a_1), \ldots, g(a_n)) \ (\text{by ind. hyp.}) \leftrightarrow \forall u(\delta(u) \land u \leq K t^K(g(a_1), \ldots, g(a_n)) \rightarrow \psi^K(u, g(a_1), \ldots, g(a_n))) \leftrightarrow (\forall x \leq t \psi)^K(g(a_1), \ldots, g(a_n)) \ (\text{by def. of } K) \leftrightarrow (\varphi^K)(g(a_1), \ldots, g(a_n))$$

QED

Now we can finish the proof of the uniform version of Montagna’s property $M$ for feasible interpretability.
Theorem 3.11 Suppose

- $U$ satisfies full induction,
- $V$ extends $PA$ and
- $U \vdash_f V$ by interpretation $K$ (preserving $=$) and polynomial $P$.

Then there is a polynomial $Q$ such that for every $\Sigma_1^0$-sentence $\sigma$ there is a $U + \sigma$-proof $p$ of $\sigma^K$ with $|\bar{r}p| \leq Q(|\bar{r}\sigma|)$.

Thus, $U + S \vdash_f V + S$ where $S$ is a finite or infinite set of $\Sigma_1^0$-sentences.

Proof. Suppose $\sigma \in S$ is the $\Sigma_1^0$-sentence $\exists x \varphi(x)$, where $\varphi \in \Delta_1^0$. By lemma 3.10, there is a polynomial $R$ such that we can prove the following by a proof of length $\leq R(|\bar{r}\sigma|)$:

$$
U \vdash \exists x \varphi(x) \rightarrow \exists x \varphi^K(g(x)) \\
\rightarrow \exists y (\delta(y) \land \varphi^K(y)) \\
\rightarrow \exists x \varphi(x) \rightarrow K.
$$

Now we have $U + S \vdash_f V + S$ by the interpretation $K$ and polynomial $Q := P + R$. QED

All results of this section also hold if we add the function symbol $\exp$ to the language of $U$ and $V$, which we need in theorem 4.1. Let $g$ be as defined in lemma 3.5. We will only give the result which needs some adaptation. The following preservation lemma corresponds to lemma 3.7:

Lemma 3.12 Suppose $\exp \in L_U$. Then $U$ proves that $g$ preserves $0, S, +, \cdot, \leq$, and $\exp$.

Proof. We already have a preservation proof for $\cdot$ by lemma 3.7. Preservation of $\exp$ then follows in the same way as preservation of $+$ was proved from preservation of $S$ in lemma 3.7. QED

4 Interpretablity does not imply feasible interpretability

Theorem 4.1 There is a set $S$ of $\Delta_2^0(\exp)$-sentences such that $\omega \models PA \vdash PA + S$, but $\omega \not\models PA \vdash_f PA + S$.

Proof. Define by Gödel’s diagonalization theorem (or rather by the free variable version as formulated by Montague) a $\Delta_2^0(\exp)$-formula $\varphi(y)$ such that

$$
PA \vdash \varphi(y) \leftrightarrow \forall x \leq \exp(y) \neg \Prf(x, \bar{r}\varphi(\bar{y})).
$$

It is easy to see that if we diagonalize directly, there is a polynomial $O$ such that for each $n$, $|n| < |\bar{r}\varphi(\bar{n})| \leq O(|n|)$. Moreover, if $\varphi(\bar{n})$ were false, then by definition we would have a proof of the $\Delta_2^0(\exp)$-sentence $\varphi(\bar{n})$; so $\varphi(\bar{n})$ must be true. But then, since $\varphi(\bar{n})$ is $\Delta_2^0(\exp)$, we have the following:

1. $PA$ proves $\varphi(\bar{n})$, though
2. because $\varphi(\hat{n})$ is true, $PA$ does not prove $\varphi(\hat{n})$ by any proof whose G"odel number is of length $\leq n$.

Define $S := \{\varphi(\hat{n}) | n \in \omega\}$. Then, by the identity interpretation, $\omega \models PA \triangleright PA + S$. Actually, as in [JM 88, section 6], we even have $PA \vdash \forall y \text{Prov}(\neg \varphi(\hat{y}))$, so $PA \vdash PA \triangleright PA + S$.

Now suppose, in order to derive a contradiction, that $\omega \models PA \triangleright_f PA + S$ by interpretation $K$ and polynomial $P$. Thus, for all $n$,

$$PA \vdash \varphi(\hat{n})^K \text{ by a proof of length } \leq P(\|\varphi(\hat{n})\|).$$

We also know by lemma 3.10 (with $U = V = PA$) that there is a polynomial $R$ such that for every $n$,

$$PA \vdash \varphi(\hat{n}) \leftrightarrow \varphi(\hat{n})^K \text{ by a proof of length } \leq R(\|\varphi(\hat{n})\|).$$

Now can construct from $R$ and $P$ a polynomial $Q$ such that for all $n$,

$$PA \vdash \varphi(\hat{n}) \text{ by a proof of length } \leq Q(\|\varphi(\hat{n})\|).$$

However, there will be $n$ such that $n > Q(O(|n|)) \geq Q(\|\varphi(\hat{n})\|)$, and we have a contradiction with 2. QED

A salient feature of the counterexample above is the trivial identity interpretation by which $PA$ interprets $PA + S$. To prove that interpretability does not imply feasible interpretability, it is not essential that the set of formulas added to $PA$ be infinite like $S$ above. We will show a counterexample where one sentence can be normally but not feasibly interpreted over $PA$. Of course in this case the normal interpretation cannot be the identity. The counterexample also shows that in general we cannot feasibly merge two compatible feasible interpretations; i.e. it is not true that if $U \triangleright_f V$, $U \triangleright_f B$ and $U \triangleright V + B$, then $U \triangleright_f V + B$ (take $U = V = PA, B = A(\hat{n})$ or $B = E^*$ as below).

**Theorem 4.2** There is a sentence $A(\hat{n})$ such that $\omega \models PA \triangleright PA + A(\hat{n})$, but $\omega \not\models PA \triangleright_f PA + A(\hat{n})$.

Proof. Let $P(x)$ be some $\Pi^0_2$-complete formula, say $P(x) = \forall y S(x, y)$, with $S \in \Sigma^0_1$. Define the formulas $R$ and $A$ by diagonalization such that

$$PA \vdash R(x, y) \leftrightarrow S(x, y) \leq \Box_{PA} R(x, y)$$

and

$$PA \vdash A(x) \leftrightarrow \Box^* A(x) \leq \exists y \neg R(x, y),$$

where $\Box^*$ is as defined in section 6.

Carrying out the proof of theorem 6.3 of the appendix section 6 in True Arithmetic, and taking the theory $U$ mentioned there to be $PA$, we find the following result: if $PA$ is consistent (as we believe it to be), then

$$\omega \models \forall x (PA \triangleright PA + A(x) \leftrightarrow P(x)).$$

16
Now suppose, to derive a contradiction, that
\[ \omega \models \forall x[(PA \models PA + A(x)) \leftrightarrow (PA \models_f PA + A(x))]. \]

Then
\[ \omega \models \forall x[(PA \models_f PA + A(x)) \leftrightarrow P(x)]. \]

However, it is easy to see that \( PA \models_f PA + A(x) \) is a \( \Sigma^0_2 \)-predicate, contradicting the \( \Pi^0_2 \)-completeness of \( P \). Therefore, there is an \( n \in \omega \) such that
- \( \omega \models PA \models PA + A(n) \) but
- \( \omega \not\models PA \models_f PA + A(n) \).

By this method we do not immediately find the value of a particular \( n \) that works, however.

A. Visser pointed out that we can make a specific counterexample in a more direct way using the Lindström method. Because \( \{ e \mid e \text{ is the Gödel number of a sentence } E \text{ such that } \neg (PA \models_f PA + E) \} \in \Pi^0_2 \), we can construct a formula \( A \) as in the appendix for which the following holds: for all sentences \( E \),
\[ \omega \models \neg (PA \models_f PA + E) \leftrightarrow PA \models PA + A(\neg E). \]

Now let \( E^* \) be the sentence constructed by the fixed point theorem such that
\[ PA \models E^* \leftrightarrow A(\neg E^*). \]

Then
\[ \omega \models \neg (PA \models_f PA + E^*) \leftrightarrow PA \models PA + E^*. \]

Therefore,
\[ \omega \models PA \models PA + E^* \text{ and } \omega \not\models PA \models_f PA + E^*. \]

QED

5 ILM is the interpretability logic of feasible interpretability over PA

In this section, we will show that Berarducci’s proof of the arithmetic completeness of ILM with respect to interpretability over \( PA \) can be adapted to prove that ILM is also arithmetically complete with respect to feasible interpretability over \( PA \).

We have already proved in chapter 3 that for all modal formulas in the language of ILM we have:

if \( ILM \models \varphi \), then for all feasibility interpretations \( ^* \), \( PA \models \varphi^* \).

Therefore, we will only need to show the converse:

if \( ILM \not\models \varphi \), then there is a feasibility interpretation \( ^* \) such that \( PA \not\models \varphi^* \).
We suppose that the reader has a copy of [Be 90] at hand in order to follow the original proofs. For the lemmas 5.5 up to 5.7, knowledge of [Pu 86], [Pu 87] or [Ve 89] will be helpful to the reader. As in [Pu 87], we take the logical complexity of a formula to be its quantifier depth. We can then adapt the results obtained in [Pu 87] to find for every standard $n$ a formula $Sat_n$, a satisfaction predicate for formulas of logical complexity $\leq n$, such that $Sat_n$ is of length linear in $n$. Subsequently, we can find proofs of length quadratic in $n$ of the Tarski conditions and of the truth lemma for these satisfaction predicates $Sat_n$. Moreover, all these results can be formalized in PA. In the formalized case, we read $Sat_n$ and $True_n$ as G"odel numbers found as function value in $n$. We will not go into the details here but refer the reader to the papers by Pudlák and Verbrugge.

First, we define some of the concepts that we use in the subsequent lemmas.

**Definition 5.1** Formally, we define the following concepts:

- $Sent(a)$ for “$a$ is the G"odel number of a sentence”;
- $Fmla(a)$ for “$a$ is the G"odel number of a formula”;
- $Fmla_n(a)$ for “$a$ is the G"odel number of a formula of logical complexity $\leq n$”; 
- $Cl(a)$ for “the G"odel number of the universal closure of the formula with G"odel number $a$”; note that $Cl$ denotes a function;
- $Indax_n(b)$ for “$b$ is the G"odel number of an induction axiom of logical complexity $\leq n$”, i.e.
  \[
  Indax_n(b) \iff Fmla_n(b) \land \exists y [Fmla(y) \land \bigwedge b = Sub(y, \gamma \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \neg \n

We need to discriminate between a few different kinds of restricted provability, as defined below. In this section, provability means provability in PA, unless we explicitly state otherwise.

**Definition 5.2** We formally define the following:

- $BP_{Prf_n}(x, y)$ for “$x$ codes a proof of the formula coded by $y$, where only formulas of logical complexity $\leq n$ appear in the proof”;
- $P-Prf_n(x)$ for “$x$ codes a formula that is provable by a proof of length $\leq P(n)$” where $P$ is a polynomial;
- $Polpr_n(x)$ for “there is a polynomial $P$ such that $\forall n \exists p (|p| \leq P(n) \land Prf(p, x)$; 
- $Polpr_{W, n}(x)$ for “there is a polynomial $P$ such that $\forall n \exists p (|p| \leq P(n) \land Prf_{W}(p, x)$
- $Prov_n(x)$ for “$x$ codes a formula that is provable by a proof which only uses those axioms of $PA$ with G"odel number $\leq n$”; abbreviation $\Box_n \varphi$ for $Prov_n(\varphi)$;
- $Prov_{W, n}(x)$ for “$x$ codes a formula that is provable by a proof which only uses those axioms of $W$ with G"odel number $\leq n$”; abbreviation $\Box_{W, n} \varphi$ for $Prov_{W, n}(\varphi)$.
In the context of satisfaction predicates $\text{Sat}_n(x, w)$, we need a few more concepts.

**Definition 5.3** We formally define the following:

- $Evalseq(w, x)$ for "$w$ encodes an evaluation sequence for the formula or term with Gödel number $x$; i.e. the length of the sequence $w$ exceeds any $i$ for which a variable $v_i$ occurs in the formula or term coded by $x$";

- $s^*(i, x, w)$ for "the sequence which is identical to $w$, except that $x$ appears in the $i$-th place"; note that $s^*$ denotes a function;

- $\text{True}_n(x)$ for $\forall w (Evalseq(w, x) \rightarrow \text{Sat}_n(x, w))$;

**Remark 5.4** When we prove formalized results, we read $\text{True}_n$ as a Gödel number just as $\text{Sat}_n$. So in that case the appropriate definition is as follows:

$$\text{True}_n(x) \vdash \forall w (Evalseq(w, x) \rightarrow \neg \text{Sat}_n(x, w)) \rightarrow \neg \forall w (Evalseq(w, x) \rightarrow \neg \text{Sat}_n(x, w)) \rightarrow \neg \forall w (Evalseq(w, x) \rightarrow \neg \text{Sat}_n(x, w)).$$

**Lemma 5.5 (feasible subformula property)** There is a polynomial $P$ such that

$$PA \vdash \forall k \forall a (\text{Fmla}(a) \rightarrow P \cdot \text{Polprov}_{|k|+|a|} [\text{Prov}_k(a) \rightarrow \exists q (B \text{Prf}_{|k|+|a|} (q, a))])$$

Proof. In [Ta 75], Takeuti gives a proof of the free cut-elimination theorem for $PA$, where $PA$ is formulated as a Gentzen system. Free cut-elimination works in such a way that all principal formulas of induction inferences in the new freely cut-free proof are substitution instances of principal formulas of induction inferences in the old proof. From this result Takeuti derives a proof of the corresponding subformula property for $PA$.

The proof of the subformula property can be adapted to the natural deduction formulation of $PA$, and can subsequently be formalized in $PA$. Thus, we can substitute any $k$ bounding the Gödel numbers of axioms used, and any Gödel number $a$ of a formula into the proof of the subformula property. Therefore, there is a polynomial $P$ such that $PA$ proves the following by proofs of length $\leq P(|k| + |a|)$:

$$PA \vdash \text{Prov}_k(a) \rightarrow \exists q (B \text{Prf}_{|k|+|a|} (q, a)).$$

Now this statement can again be formalized, so that we find

$$PA \vdash \forall k \forall a (\text{Fmla}(a) \rightarrow P \cdot \text{Polprov}_{|k|+|a|} [\text{Prov}_k(a) \rightarrow \exists q (B \text{Prf}_{|k|+|a|} (q, a))]),$$

as desired. QED

**Lemma 5.6** There is a polynomial $P$ such that

$$PA \vdash \forall k \forall a (\text{Fmla}(a) \rightarrow P \cdot \text{Polprov}_{|k|+|a|} [\exists q (B \text{Prf}_{|k|+|a|} (q, a) \rightarrow \text{True}_{|k|+|a|}(a))].$$
Proof. First, we work informally by induction on the construction of \( q \). We work in \( PA \), and we take any \( k \) and an \( a \) such that \( a \) is the Gödel number of a formula. We have to prove by polynomial length proofs (where the polynomial is fixed from outside) that \( \text{True}_{k+|a|} \) preserves the axioms and rules as applied to formulas of logical complexity \( \leq |k| + |a| \).

As an example, we show how this works for the induction schema. We take \( v_i \) as the induction variable in all our instances of the induction axioms. So suppose \( b \) codes an induction axiom of logical complexity \( \leq |k| + |a| \), e.g. \( b = (\text{Sub}(y, \check{v} \check{r}, \check{0}) \land \forall v_1 (\check{y} \check{r} \rightarrow \neg \text{Sub}(y, \check{v} \check{r}, \check{S} \check{v} \check{r})) \rightarrow \forall v_1 \check{y}) \). We have to prove the following by a proof of length polynomial in \( n := |k| + |a| \):

\[ \text{True}_n (\text{Sub}(y, \check{r} \check{v} \check{r}, \check{0}) \land \forall v_1 (\check{y} \check{r} \rightarrow \neg \text{Sub}(y, \check{r} \check{v} \check{r}, \check{S} \check{v} \check{r})) \rightarrow \forall v_1 \check{y}) . \] (4)

By a proof of length quadratic in \( n \) of the Tarski properties for \( Sat_n \) and a proof of length quadratic in \( n \) of a call by name / call by value lemma for \( Sat_n \) (cf. the proofs of lemmas 3.12 and 3.16 in [Ve 89]), (4) is equivalent to the following:

\[ \forall w (Sat_n(y, s \check{x}(1, 0, w)) \land \forall x (Sat_n(y, s \check{x}(1, x, w)) \rightarrow Sat_n(y, s \check{x}(1, Sx, w))) \rightarrow \forall x (Sat_n(y, s \check{x}(1, x, w))) . \] (5)

The formulas (5) are themselves instances of induction of length linear in \( n \), so they are provable by proofs of length linear in \( n \). A polynomial of the form \( P(n) = K \cdot n^3 \) should now suffice to carry out the proofs of (4).

Again, we can formalize the argument to derive the following:

\[ PA \vdash \forall k \forall a (Fmla(a) \rightarrow P\text{-Polprov}_{k+|a|} (\forall b (\text{Indax}_{k+|a|} (b) \rightarrow \text{True}_{k+|a|} (b)))) . \]

Similarly, we can show by polynomially short proofs that the other axioms of logical complexity \( \leq |k| + |a| \) are true, and that the rules preserve truth. We leave these proofs and their formalizations to the reader. QED

**Lemma 5.7** There is a polynomial \( P \) such that

\[ PA \vdash \forall k \forall a (Fmla(a) \rightarrow P\text{-Polprov}_{k+|a|} (\text{True}_{k+|a|} (a) \rightarrow \neg \text{Cl}(a))) \]

Proof. By a formalized Tarski’s showing lemma; cf. lemma 3.10 of [Ve 89]. QED

The following theorem corresponds to the reflection theorem 1.6 in [Be 90].

**Theorem 5.8 (feasible reflection theorem)** There is a polynomial \( P \) such that

\[ PA \vdash \forall k \forall a (\text{Sent}(a) \rightarrow P\text{-Polprov}_{k+|a|} (\neg \text{Prov}_{k} (a) \rightarrow \check{a})) \]

Proof. Combine lemmas 5.5, 5.6 and 5.7. QED

In the following lemmas and theorems, \( \exists K \) abbreviates \( \exists K ("K \text{ codes an interpretation} \land \ldots") \).

The next lemma was proved by Albert Visser [Vi 89, Chapter 6, Claim 3] in the course of a formalized Henkin construction in \( 1 \Delta_0 + \Omega_1 \).
Lemma 5.9  Suppose \( \beta \) axiomatizes some subset of a \( \Sigma_1^1 \)-language \( L \). Then there is an \( r \) such that

\[
I \Delta_0 + \Omega_1 \vdash \Box_U \text{Con}(\beta) \rightarrow \exists K \forall a \in \text{Sent}(L) \exists p < \omega^L_1(a) \text{Prf}_U(p, \Gamma_{\Box a} \rightarrow \Gamma a^K).
\]

Proof. See [Vi 89]. QED

Because of the correspondence between the values of \( \omega_1 \)-terms in \( a \) and \( \text{exp}(a) \) (the values of polynomials in \( |a| \)), lemma 5.9 implies the following lemma:

Lemma 5.10  Suppose \( \beta \) axiomatizes some subset of a \( \Sigma_1^1 \)-language \( L \). Then there is a polynomial \( P \) such that

\[
I \Delta_0 + \Omega_1 \vdash \Box_U \text{Con}(\beta) \rightarrow \exists K \forall a \in \text{Sent}(L) P(\text{Polprov}_{U, |a|}^{(\Gamma \Box a \rightarrow \Gamma a^K)}).
\]

The following theorem corresponds to Orey's theorem; see for example [Be 90, Theorem 2.9]

Theorem 5.11 (feasible Orey's theorem)  Suppose that \( U \supseteq PA \) and \( W \) is axiomatized by \( \alpha \), where \( \alpha \) is a \( \Sigma_1^1 \)-formula. Then

\[
PA \vdash \forall x \text{Polprov}_{U, |x|}^{(\Gamma \Diamond a_x \rightarrow \top)} \rightarrow U \triangleright W.
\]

Proof. Work in \( PA \) and suppose

\[
\forall x \text{Polprov}_{U, |x|}^{(\neg \Box a_x \rightarrow \top)}.
\]

In \( U \), we will do a Henkin construction for the Feferman proof predicate for \( W \). First define:

\[
\beta(x) := \alpha(x) \land \Diamond a_{x+1} \rightarrow \top.
\]

As in Feferman's original proof, we can prove that

\[
\Box_U \text{Con}(\beta).
\]

(For, reason in \( U \) and suppose \( \text{Prf}_\beta(x, \bot) \), then for the axiom of \( \beta \) coded by the biggest Gödel number \( y \) to appear in \( x \) we have \( \alpha(y) \land \neg \Diamond a_{y+1} \rightarrow \top \), thus \( \neg \beta(y) \): a contradiction.)

On the other hand, by provable \( \Sigma_1^1 \)-completeness for \( \alpha(a) \) and by the assumption

\[
\forall x \text{Polprov}_{U, |x|}^{(\neg \Diamond a_x \rightarrow \top)}
\]

we have:

\[
\forall a(\alpha(a) \rightarrow \text{Polprov}_{U, |a|}^{(\neg \alpha(a) \land \Diamond a_{x+1} \rightarrow \top)}).
\]

So, by definition of \( \beta \), we have the following:

\[
\forall a(\alpha(a) \rightarrow \text{Polprov}_{U, |a|}^{(\neg \beta(a) \rightarrow \top)}).
\]

But, using \( \Box_U \text{Con}(\beta) \) we can apply lemma 5.10 to first derive

\[
\exists K \forall a \in \text{Sent}(L) \text{Polprov}_{U, |a|}^{(\neg \Box a \rightarrow \Gamma a^K)},
\]

21
and thus
\[ \exists K \forall a \in \text{Sent}(L) \text{Prov}_{U,a}((\forall \beta(a) \rightarrow \neg a^K)). \]
(7)

Finally we can combine 6 and 7 to get the desired conclusion
\[ \exists K \forall a (a(a) \rightarrow \text{Prov}_{U,a}(a^K)), \]
i.e. \( U \triangleright_f W \). QED

Now we can start the proof of the arithmetical completeness of ILM with respect to feasible interpretations (cf. definition 3.2) over \( PA \).

Theorem 5.12 If \( ILM \upharpoonright B \), then there is a feasibility interpretation * such that \( PA \upharpoonright B^* \).

The proof will in most places be identical to the one in [Be 90]. First we will sketch the outline of the proof, then we will prove the propositions that we need in the feasible case but differ essentially from those used in [Be 90].

Proof sketch. Suppose \( ILM \upharpoonright B \), and take, by modal completeness of \( ILM \) with respect to simplified models, a provably primitive recursive \( ILM \)-Kripke model \( V = \langle V, R, S, b, \vdash \rangle \), with \( b = 1 \) and \( 1 \vdash B \). Extend \( V \) with a new root 0 with 0Rx for all \( x \in V \), as in definition 5.1 of [Be 90]. Adapting definition 5.2 of [Be 90], we define a feasibility interpretation * such that for all propositional variables \( p \),
\[ p^* := "\exists x \in V \cup \{0\} : L = x \land x \vdash p", \]
where \( L \) is defined as the limit of the Solovay function \( F \), which is in turn defined in definition 5.7 of [Be 90]. We want to prove the following:

whenever \( 1 \vdash A \), then \( PA \upharpoonright A^* \),
(8)

Then we will be done, as we have chosen \( V \) such that \( 1 \vdash B \). To prove (8), we need to prove in \( PA \) a few properties of \( F \) and its limit \( L \). Subsequently we need to prove by induction on the construction of the formula that for all formulas \( A \), the feasibility interpretation * is faithful on \( A \), i.e.
\[ PA \vdash \forall x \in V (x \vdash A \land L = x \rightarrow A^*) \]
and
\[ PA \vdash \forall x \in V (x \vdash \neg A \land L = x \rightarrow \neg A^*). \]

It is clear from the definition of \( F \) that * is faithful on atomic formulas. Moreover, the induction steps for the propositional connectives and \( \square \) immediately follow from the proofs in [Be 90]. Even the "negative" induction step for \( \triangleright \) has a straightforward proof:

Work in \( PA \) and suppose \( x \in V \), \( x \vdash \neg (A \triangleright B) \), and \( L = x \); then by part 2 in the proof of lemma 5.6 of [Be 90] and by the induction hypothesis, \( \neg (A^* \triangleright_f B^*) \). But then surely \( \neg (A^* \triangleright_f B^*) \), thus, as * is a feasibility interpretation, \( \neg (A \triangleright B)^* \).
For the "positive" direction, we need two extra lemmas. First we will prove in PA that F satisfies a feasible adaptation of Berarducci’s property S, which we then use to finish the induction step for \( \triangleright \).

For \( x \in V \), let \( \text{rank}(x,n) \), the rank of \( x \) at stage \( n \), be defined as in definition 5.7 of [Be 90]. The following proposition is an analogue of proposition 5.14 in [Be 90].

**Proposition 5.13 (F has feasible property S)** PA proves the following:

\[
\text{PA} \vdash \forall x \in V \cup \{0\}[L = x \rightarrow \text{Polprov}_{[k]}(\forall y, z \in V \cup \{0\}(L = y \land xRz \land ySz \rightarrow \Diamond_k L = z))]\]

Proof. We will prove the proposition by combining a few facts that are easy to check. For brevity’s sake, we will leave out “\( \in V \cup \{0\} \)” after quantifiers \( \forall x, \forall y, \forall z \). Likewise, capital P, with or without subscript, refers to formalized polynomials.

**Fact 1** PA \( \vdash \text{Polprov}_{[k]}(\forall y(L = y \rightarrow \Diamond_k L = y)) \)

Proof. Immediately from the reflection theorem 5.8. The formula \( L = y \) has a fixed length, so the polynomial found in the proof of the reflection theorem in this case depends only on \( |k| \). QED

**Fact 2** PA \( \vdash \text{Polprov}_{[k]}(\forall y(L = y \rightarrow \forall n(k < \text{rank}(y,n)))) \)

Proof. Immediately from fact 1 and the definition of rank. The appearance of \( k \) as an efficient numeral keeps the length of the proof polynomial in \( [k] \). (This is also the case in the other facts below) QED

**Fact 3** PA \( \vdash \text{Polprov}_{[k]}(\forall z(\Box_k L \neq z \rightarrow \exists m \forall n \geq m(\text{rank}(z,n) \leq k))) \)

Proof. Immediately from the definition of rank. QED

**Fact 4** PA \( \vdash \text{Polprov}_{[k]}(\forall y \forall z(L = y \land \Box_k L \neq z \rightarrow \exists n(\text{F(n)} = y \land n \text{ codes } y \land \text{rank}(z,n) \leq \hat{k} \land \text{rank}(z,n) < \text{rank}(y,n)))) \)

Proof. From the definition of limit and fact 3: just take \( n \) big enough. We can take care that \( n \) codes \( y \) because we have an infinitely repetitive primitive recursive coding of the elements of \( V \cup \{0\} \). Finally, to prove \( \text{rank}(z,n) < \text{rank}(y,n) \), we use fact 2. QED

**Fact 5** PA \( \vdash \forall x(L = x \rightarrow \text{Polprov}_{[k]}(\exists j(J \geq \hat{k} \land F(j) = x)) \)

Proof. Immediate from the definition of the limit \( L \) of \( F \). QED

**Fact 6** We have the following:

\[
\text{PA} \vdash \forall x(L = x \rightarrow \text{Polprov}_{[k]}(\forall y \forall z(L = y \land \Box_k L = z \land xRz \land ySz \rightarrow \exists n \exists j(F(n) = y \land n \text{ codes } y \land \text{rank}(z,n) < \text{rank}(y,n) \land \text{rank}(z,n) \leq k \leq j \land F(j) = x \land F(\text{rank}(z,n))SxRz \land F(\text{rank}(z,n))Rz))) \)
\]
Proof. For the part up to \( F(j) = x \), we combine facts 5 and 4. For the last two conjuncts, we use the monotonicity of \( F \) and the property corresponding to \( M \) of Veltman \( ILM \)-frames. QED

**Fact 7** We have the following:

\[
PA \vdash \forall x (L = x \rightarrow \text{Polprov}_{[\Gamma]}(\forall y \forall z (L = y \land \square_{y} L = z \land x R z \land y S z \rightarrow \exists n (F(n) = y \land F(n + 1) = z))) \rightarrow
\exists n (F(n) = y \land F(n + 1) = z))
\]

Proof. Immediate from fact 6 and the definition of the function \( F \), clause 2. QED

Now we can wrap up the proof: we see that \( \exists n (F(n) = y \land F(n + 1) = z) \) is inconsistent with \( L = y \), so in fact we have what we were looking for:

\[
PA \vdash \forall x (L = x \rightarrow \text{Polprov}_{[\Gamma]}(\forall y \forall z (L = y \land x R z \land y S z \rightarrow \Diamond_{x} L = z)) \rightarrow \exists n (F(n) = y \land F(n + 1) = z))
\]

QED

The following proposition corresponds to part 1 of Lemma 5.6 of [Be 90].

**Proposition 5.14 (positive induction step for \( \triangleright \))** Let \( * \) be the feasibility interpretation defined in the proof sketch of theorem 5.12. Suppose as induction hypothesis that

\[
PA \vdash \forall y (L = y \rightarrow (y \vdash A \leftarrow A^{*})) \text{ and }
\]

\[
PA \vdash \forall z (L = z \rightarrow (z \vdash B \leftarrow B^{*})).
\]

Then

\[
PA \vdash \forall x (L = x \land x \vdash A \triangleright B \rightarrow (A \triangleright B)^{*}).
\]

Proof. Let \( b \) be such that

\[
PA \vdash \forall y (L = y \rightarrow (y \vdash A \leftarrow A^{*})) \text{ and }
\]

\[
PA \vdash \forall z (L = z \rightarrow (z \vdash B \leftarrow B^{*})),
\]

both by proofs that use axioms of Gödel number up to \( b \). Moreover suppose \( c \) is such that

\[
PA \vdash \forall z (z \vdash B \rightarrow \Box_{C}(z \vdash B)));
\]

for this, any \( c \geq \) the Gödel number of the biggest axiom of Robinson's arithmetic \( Q \) will do. Define \( d := \text{max}(b,c) \). By theorem 5.11, the feasible version of Orey's theorem, it is sufficient to prove the following:

\[
PA \vdash \forall x (L = x \land x \vdash A \triangleright B \rightarrow \forall k \geq d \text{ Polprov}_{[\Gamma]}(A^{*} \rightarrow \Diamond_{k} B^{**})).
\]

Again, we will state a list of easily provable facts from which the result immediately follows.
Fact 1 $PA \vdash \forall x(L = x \land x \vdash A \vdash B \rightarrow \Box[A^* \rightarrow \exists y(L = y \land xRy \land y \vdash A \land x \vdash A \vdash B)])$  
Proof. $L = x \rightarrow \Box[y(L = y \land xRy)]$ by property ($\neg R$), $\Box(A^* \land L = y \rightarrow y \vdash A)$ by the induction hypothesis, and $\Box(z \vdash A \vdash B)$ by provable $\Sigma_0^0$-completeness. QED

Fact 2 $PA \vdash \forall x(L = x \land x \vdash A \vdash B \rightarrow \Box[A^* \rightarrow \exists y \exists z(L = y \land xRy \land y \vdash A \land x \vdash A \vdash B \land xRz \land ySz \land z \vdash B)])$  
Proof. From fact 1 and the definition of $x \vdash A \vdash B$. QED

Fact 3 $PA \vdash \forall x \forall k \geq d Polyprov_{|k|}(z \vdash B \rightarrow \Box[z \vdash B])$  
Proof. From the assumption, and the definition of $k$ appears only as efficient numeral. QED

Fact 4 $PA \vdash \forall x(L = x \land x \vdash A \vdash B \rightarrow \forall k \geq d Polyprov_{|k|}(A^* \rightarrow \exists y \exists z(L = y \land xRy \land xRz \land ySz \land z \vdash B))$  
Proof. From fact 2 for $A^* \rightarrow \exists y \exists z(L = y \land xRy \land xRz \land ySz \land z \vdash B)$; fact 3 for a proof of length polynomial in $k$ of $z \vdash B \rightarrow \Box[z \vdash B]$, and proposition 5.13 for a proof of length polynomial in $|k|$ of $L = y \land xRy \land xRz \land ySz \rightarrow \Box[z \vdash B]$. QED

Fact 5 $PA \vdash \forall x(L = x \land x \vdash A \vdash B \rightarrow \forall k \geq d Polyprov_{|k|}(A^* \rightarrow \exists z \land \Box[z \vdash B] \land (L = z \land z \vdash B))$  
Proof. If $k$ is big enough (and $k \geq d$ will do), then by an easily formalized property of modus ponens, we have the following by proofs of length polynomial in $|k|$: $PA \vdash \forall z([\Box[z \vdash B \rightarrow L \neq z] \land \Box[kz \vdash B] \rightarrow \Box[kL \neq z])$, and thus $PA \vdash \forall z(\Box[kL = L = z \land z \vdash B])$. This argument can be formalized and combined with fact 4 to derive fact 5. QED

Fact 6 $PA \vdash \forall x(L = x \land x \vdash A \vdash B \rightarrow \forall k \geq d Polyprov_{|k|}(A^* \rightarrow \Box[kB^*])$  
Proof. From fact 5 and the induction hypothesis; the fact that $k \geq d$ is used at this place. QED

From fact 5 and the feasible version of Orey's theorem, we may indeed derive  
$PA \vdash \forall x(x \vdash A \vdash B \land L = x \rightarrow (A \vdash B)^*)$,  
as desired.  
QED

Proof sketch of theorem 5.12, continued. Concluding by induction that $^*$ is faithful on all formulas $A$, we have proved that $ILM \not \vdash B^*$. Therefore, $ILM$ is arithmetically complete with respect to feasible interpretability over $PA$.  
QED
References


[Há 79] P. Hájek, On partially conservative extensions of arithmetic, Logic Colloquium 78 (M. Boffa et al., editors), North Holland, Amsterdam, 1979, pp. 225-234.


6 Appendix

Solovay proved that the set \( \{ A \mid PA \vdash PA + A \} \) is \( \Pi_2^0 \)-complete [So]. This result inspired Hájek to prove that, for every \( n \), the set \( \{ A \mid A \text{ is } \Pi_{n+1}^0 \text{-conservative over } PA \} \) is also \( \Pi_2^0 \)-complete [Há 79].

We have adapted the proof of theorem 6.3 from Visser’s unpublished rendition [Vi 90] of an alternative proof by Lindström of Hájek’s general result.

Definition 6.1 Define \( \Box^x_U B \) for “there is a proof of the formula \( B \) which only uses those axioms of \( U \) with Gödel number \( \leq x \).”

Suppose \( U \) and \( V \) are theories extending \( PA \), such that for all \( B \) \( PA \vdash \forall x \Box_U (\Box^x_U B \rightarrow B) \) (reflection for \( U \)), in particular \( PA \vdash \forall x \Box_U \Box^x_U \top \). Then by the Orey-Hájek theorem, \( PA \vdash U \supset V \iff \forall x \Box_U \Box^x_U \top \). The rest of the proof is taken almost verbatim from [Vi 90].

Let \( P \) be any \( \Pi_2 \)-predicate, say \( P = \forall x S(x) \), with \( S \in \Sigma_1^0 \). Pick \( R \) by diagonalization such that \( PA \vdash R \iff S \leq \Box^x_U R \). Let \( Q := \Box^x_U R \leq S \). (We suppress free variables when convenient).

We first prove a lemma.

Lemma 6.2 \( PA \vdash \Box^x_U R \iff S \vee \Box^x_U \bot \).

Proof. Work inside \( PA \) and suppose \( \Box^x_U R \). Then either \( R \) or \( Q \) holds. In case that \( R \) holds we have \( S \) by definition. In case that \( Q \) holds we have \( \Box^x_U Q \) by \( \Sigma_1^0 \)-completeness, and hence by definition both \( \Box^x_U R \) and \( \Box^x_U \neg R \), thus \( \Box^x_U \bot \).

For the other direction, suppose \( S \). Again we have either \( R \) or \( Q \). From \( R \) we find \( \Box^x_U R \) by \( \Sigma_1^0 \)-completeness. From \( Q \) we immediately derive \( \Box^x_U R \). Finally \( \Box^x_U \bot \) gives \( \Box^x_U R \) as well. QED

Define \( A \) by diagonalization such that \( PA \vdash A \iff \Box^x_U \neg A \leq \exists y \neg R(y) \). Note that by lemma 6.2 we have \( PA \vdash \Box^x_U \top \rightarrow [\forall x \Box_U R(x) \rightarrow P] \) and \( PA \vdash P \rightarrow \forall x \Box_U R(x) \).

Theorem 6.3 \( PA \vdash \Box^x_U \top \rightarrow (U \supset U + A \iff P) \)

Proof. Work in \( PA \) and suppose \( \Box^x_U \top \).
Suppose \( U \rhd U + A \). Then by the Orey-Hájek theorem \( \forall x \Box_U \Diamond_{U,x} A \). We will prove \( \forall x \Box_U R(x) \). Pick any \( x \). We have \( \Box_U [Q(x) \rightarrow \neg R(x)] \); therefore by definition of \( A \),

\[ \Box_U [Q(x) \rightarrow \neg A \lor \Box_{U,x} \neg A] \]

and hence by reflection

\[ \Box_U [Q(x) \rightarrow \neg A] \]

But then there is a \( y \) such that

\[ \Box_{U,y} [Q(x) \rightarrow \neg A] \]

so by \( \Sigma^0_1 \)-completeness

\[ \Box_U \Box_{U,y} [Q(x) \rightarrow \neg A] \]

Also by \( \Sigma^0_1 \)-completeness, there is a \( z \) such that

\[ \Box_U [Q(x) \rightarrow \Box_{U,z} Q(x)] \]

Combining the previous two facts, we find a \( u \) such that

\[ \Box_U [Q(x) \rightarrow \Box_{U,u} \neg A] \]

and thus, by the assumption, \( \Box_U \neg Q(x) \). It follows that

\[ \Box_U [\Box_U R(x) \rightarrow R(x)] \]

hence by Löb's theorem \( \Box_U R(x) \). We may conclude \( \forall x \Box_U R(x) \), thus, because we have \( \Diamond_U T \), we conclude \( P \).

Suppose \( P \). Then \( \forall x \Box_U R(x) \) and thus \( \forall x \Diamond_U (\forall y < x R(y)) \). It follows by definition of \( A \) that

\[ \forall x \Box_U (\Box_{U,x} \neg A \rightarrow A) \]

On the other hand, we have

\[ \forall x \Box_U (\Box_{U,x} \neg A \rightarrow \neg A) \]

by reflection, hence \( \forall x \Box_U (\Diamond_{U,x} A) \). But then by the Orey-Hájek theorem \( U \rhd U + A \).

QED