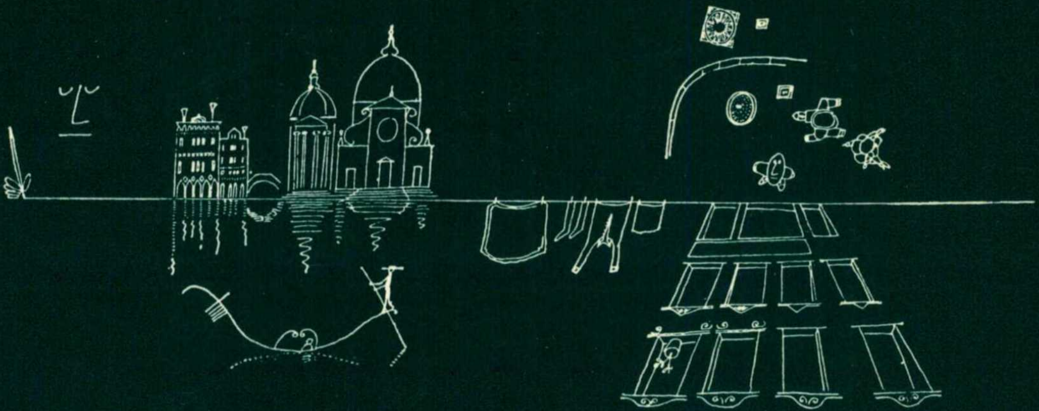
EFFICIENT METAMATHEMATICS



Laurina Christina Verbrugge

EFFICIENT METAMATHEMATICS

ACADEMISCH PROEFSCHRIFT

Ter verkrijging van de graad van Doctor aan de
Universiteit van Amsterdam,
op Gezag van de Rector Magnificus
Prof. Dr. P.W.M. de Meijer
in het openbaar te verdedigen in de
Aula der Universiteit
(Oude Lutherse Kerk, ingang Singel 411, hoek Spui)
op dinsdag 14 september 1993 te 15.00 uur.

door

Laurina Christina Verbrugge

geboren te Amsterdam

Promotor:	Prof. dr A.S. Troelstra
Co-promotores:	Dr D.H.J. de Jongh
	Dr A. Visser

Faculteit Wiskunde en Informatica
Universiteit van Amsterdam
Plantage Muidergracht 24

ILLC Dissertation Series, 1993, no. 3
ISBN 90-800769-8-8

Cover illustration by Saul Steinberg

Druk: Haveka B.V., Alblasserdam

Ter nagedachtenis aan
Laurina Verbrugge-de Graag
(1901-1993)

Contents

Dankwoord

Om te beginnen wil ik mijn oprechte dank betuigen aan het triumviraat van promotor en co-promotores. Mijn promotor, professor A.S. Troelstra, dank ik voor zijn commentaar en voor zijn relativerende opmerkingen. Dick de Jongh deed nuttige en doordachte aanbevelingen naar aanleiding van vele eerdere versies van dit proefschrift. Albert Visser tenslotte wil ik bedanken voor zijn aanstekelijke enthousiasme en de flexibiliteit waarmee hij op een moeilijk moment onmiddellijk een werkplek in Utrecht voor me verzorgde.

Tijdens de eerste paar jaar van mijn promotie-onderzoek vormden de discussies met Edith Spaan een bron van inspiratie, waarvoor ik haar zeer erkentelijk ben.

I have had the good fortune to be supported by NWO (project 611-308-004) and the Tempus Project, which enabled me to visit Prague, Siena and Florence several times. I gratefully thank my colleagues Petr Hájek, Vítězslav Švejdar and Franco Montagna for their hospitality and stimulating discussions. A special thanks goes to Alessandro Berarducci, who has been a wonderful co-author and mentor. In fact, I learnt most of my knowledge about weak arithmetics from his "scatola magica" of index cards crammed with elegant proofs. I also want to thank him and Silvia Cantelli for their generosity and for their fascinating stories, told in a mixture of Italian, German and English.

De laatste tijd heb ik op allerlei manieren veel steun van mijn familie, vriendinnen, vrienden en collega's gekregen. Ik denk bijvoorbeeld aan mijn kamergenoot Domenico Zambella, die me voorzag van krentenbollen en levendige gesprekken over begrensde rekenkunde. Volodya Shavrukov las blijkens zijn rake commentaar het manuscript van dit proefschrift op nauwgezettere wijze dan wie ook, en maakte bovendien een mooie macro.

Mijn twee paranimfen, Marjanne de Jong en Andreja Prijatelj, wil ik bedanken voor de humor en creativiteit waarmee ze me voortdurend hebben gemotiveerd. Joke Hendriks heeft me tijdens onze wekelijkse bijeenkomsten moed ingesproken en is tegelijkertijd een streng extern geweten voor me geweest - die combinatie werkte perfect. Maarten de Rijke wil ik bedanken voor zijn effectieve hulp bij LaTeX- en printproblemen, en Marjet Denijs voor het vormgeven van de voorkant van dit proefschrift.

I would like to thank all my colleagues in logic and theoretical computer science for providing fruitful discussions and a pleasant atmosphere; specifically I thank Lev Beklemishev, Harry Buhrman, Marc Jumelet, Marianne Kalsbeek, Karen Kwast, Alexander Razborov, Dirk Roorda, Víctor Sánchez, Martijn Spaan, and Yde Venema.

Tot slot dank ik mijn ouders, die mijn keuzen soms met opgetrokken wenkbrauwen begroetten, maar altijd achter me stonden.

Rineke Verbrugge
Amsterdam, juli 1993

Part I

Introduction and background

Chapter 1

Introduction

Verde embeloso de la vida humana,
loca esperanza, frenesí dorado,
sueño de los despiertos intrincado,
como de sueños, de tesoros vana;

alma del mundo, senectud lozana,
decrépito verdor imaginado;
el hoy de los dichosos esperado
y de los desdichados el mañana:

sigan tu sombra en busca de tu día
los que, con verdes vidrios por anteojos,
todo lo ven pintado a su deseo;

que yo, más cuerda en la fortuna mía,
tengo en entrambas manos ambos ojos
y solamente lo que toco veo.

(Sor Juana Inés de la Cruz,
seventeenth century, Mexico)

Green allurement of our human life,
mad Hope, wild frenzy gold-encrusted,
sleep of the waking full of twists and turns
for neither dreams nor treasures to be trusted;

soul of the world, new burgeoning of the old,
fantasy of blighted greenery,
day awaited by the happy few,
morrow which the hapless long to see:

let those pursue your shadow's beckoning
who put green lenses in their spectacles
and see the world in colors that appeal.

Myself, I'll act more wisely toward the world:
I'll place my eyes right at my fingertips
and only see what my two hands can feel.

(translation Alan S. Trueblood)

1.1 What is efficiency?

For more than a century now, mathematicians have been concerned with ways to transfer proofs in which the existence of objects satisfying certain properties is claimed in such a way as to enable us to find particular objects satisfying those properties. Stated in modern terms, they wanted to find witnesses for provable existential statements in a constructive way. They invented new constructive logics to accomodate their concern. Thus, for example, if some statement of the form $\forall x \exists y A(x, y)$ ("for all x there exists a y such that $A(x, y)$ holds") is provable in intuitionistic Heyting Arithmetic, then its proof provides a recursive algorithm to find for every x a witness y such that $A(x, y)$ holds. Moreover, we can assign a numerical code e to the recursive algorithm, for which Heyting Arithmetic proves $\forall x \exists y T_1(e, x, y)$. Here T_1 is Kleene's primitive recursive T-predicate for recursive functions of one variable; $T_1(e, x, y)$ stands for "y is the numerical code of a computation of the value of the function with associated code e on input x".

If all quantifiers of A are bounded, then truth of $\forall x \exists y A(x, y)$ already provides a recursive algorithm e for finding witnesses. As in the case of Heyting Arithmetic, we have that if $\forall x \exists y A(x, y)$ is not only true but also provable in Peano Arithmetic, then Peano Arithmetic proves $\forall x \exists y T_1(e, x, y)$ (see [Kr 51, Kr 52, Kr 58]).

Parsons and Mints independently proved a similar result for $I\Sigma_1$, the subsystem of Peano Arithmetic in which induction is only allowed for formulas of the form $\exists z B(z)$ where all quantifiers in B are bounded. If $I\Sigma_1$ proves $\forall x \exists y A(x, y)$ for bounded A, then there is even a primitive recursive function that provides the witnesses (see [Pa 72, Mi 71]).

Since the rise of computer science the desire for constructivity has been growing more and more stringent. It is no longer sufficient that the constructions to find the witnesses are given by recursive, or even primitive recursive algorithms. The algorithms have to be *efficient*. For example algorithms that need a computation time exponential in the length of the input are ruled out.

Unfortunately, the precise mathematical meaning of the noun "efficiency" is hard to pinpoint. The dictionary definition does not offer much guidance:

> **efficiency** [F., from L., *efficiens, -ntem*, pres. p. of *efficere*, to EFFECT], *n.* Adequate fitness; power to produce a desired result; (*Eng*) the ratio of the output of energy to the input of energy [HS 84].

If we want to classify algorithms as to their efficiency in a mathematically fruitful way, we should abstract as much as possible from highly specific factors like the programming language and the size, kind and operating speed of the computer on which the program is run. Cobham gave just such a classification in [Co 64]. He classified algorithms in terms of the number of steps taken by a Turing Machine to complete the computation as a function of the length of the input. For example, an algorithm runs in *polynomial time* if there are fixed integers c and k such that for all n, the computation on inputs of length n is completed in at most $c \cdot n^k$ steps.

Efficiency is thus a relative measure, not an absolute one. When I do abuse language by using the adjective "efficient" in an absolute sense in this dissertation, it refers to algorithms which run in (deterministic or sometimes non-deterministic) polynomial time.

In the literature, the word *"feasible"* is used a little more precisely than "efficient"; a feasible algorithm should run in time polynomial in the input. However, even for this relatively young word non-standard uses abound (see e.g. de volume [BS 90]).

In 1986, Buss introduced systems of arithmetic that cater to the need for efficient algorithms [Bu 86]. For example, if the best-known of his systems proves a statement of the form $\forall x \exists y A(x, y)$ where A defines a predicate in NP (i.e. computable by a non-deterministic polynomial time Turing machine), then there is a polynomial time algorithm f that computes for every x a witness $f(x)$ such that $A(x, f(x))$ holds. For syntactic reasons Buss called his hierarchy of systems Bounded Arithmetic. Induction is not allowed for all first order formulas, as in the standard system of arithmetic, Peano Arithmetic. Instead, every system allows induction only for a specific class of bounded formulas, in which all quantifiers are bounded by a term in the language of Bounded Arithmetic. Because of the concern for efficiency behind his systems we can see them as an example of efficient mathematics, and his most well-known system S^1_2 even as feasible mathematics.

1.2 What is metamathematics?

In 1931, Gödel proved his First and Second Incompleteness Theorems (see [Gö 31, Ho 79]). They are theorems *about* systems of mathematics. For example, the Second Incompleteness Theorem says that every consistent theory that can, by some coding mechanism, express enough information about its syntax, cannot prove its own consistency. Theories like Peano Arithmetic (PA) prove a *formalization* of the Second Incompleteness Theorem, namely the arithmetical sentence that expresses "If PA is consistent, then PA does not prove that PA is consistent". In general, *metamathematics* is the study of mathematical theories by mathematical methods like formalization.

Löb [Lö 55] listed the properties of the formalized provability predicate that are sufficient for proving the formalized version of Gödel's Second Incompleteness Theorem. In the seventies, the conditions listed by Löb came into their own as the modal logic of provability, where $\Box A$ is read as "A is provable" (see e.g. [MS 73]). This provability logic, which we call L after Löb, has proved to be very useful in the study of provability predicates for theories like Peano Arithmetic. The outstanding result in this area was attained by Solovay, who proved in 1976 that L exactly captures the modal properties of the provability predicate of Peano Arithmetic. More precisely, he showed that for any modal formula A, L proves A if and only if Peano Arithmetic proves all translations of A in the language of arithmetic that interpret $\Box$ as the formalized provability predicate [So 76].

1.3 Interpretability and its logics

As early as the 19th century, mathematicians sought a relative consistency proof in order to show that the consistency of Euclidean geometry implies the consistency of Hyperbolic geometry. A positive answer would show that the parallel postulate is not provable from the other Euclidean axioms. One of the earliest proofs, by Poincaré, made use of an *interpretation*, even though that concept was not formally defined at the time. We do not give a precise definition here. Intuitively, the theory U *interprets* the theory V if there exists a translation of the language of V into the language of U that enables us to "see" a model of V inside every model of U. Interpretations have been used for various purposes, at first mainly for relative consistency proofs as in Poincaré's example (see also [Hi 1899]), and later for proving theories to be undecidable (see e.g. [TMR 53]).

The study of interpretability picked up momentum in the seventies with papers by Hájek and Solovay (see e.g. [Há 71, Há 72, So 76b]). Their results, which include the Orey–Hájek characterization for interpretability over theories like Peano Arithmetic, gave the study of interpretability its proper place as a part of metamathematics.

The eighties saw the development of modal logics for interpretability, first introduced in [Šv 83] and [Vi 89]. Here $\rhd$ is a binary modal operator corresponding to interpretability over a base theory T. In contrast to the case of provability logic, there are several interpretability logics around, capturing the principles that govern interpretability over various kinds of base theories (see e.g. [JV 90, Vi 90a]). For example, Berarducci and Shavrukov independently proved by modified Solovay constructions that the interpretability logic ILM is the logic of interpretability over Peano Arithmetic (see [Ber 90, Sh 88]).

1.4 Metamathematics of efficient mathematics

If we want to prove an analog of the formalized version of Gödel's Second Incompleteness Theorem for systems of efficient mathematics like Bounded Arithmetic, we are forced to make our metamathematics efficient, too.

First of all, we need efficient numerals. This is because $\underbrace{S \dots S}_{k \text{ times}}0$ has length exponential in the length of k written in binary; so we cannot prove in Bounded Arithmetic the totality of the function that sends natural numbers k to the code of $\underbrace{S \dots S}_{k \text{ times}}0$. Thus we use numerals that are based on the binary expansion of k and are of length linear in the logarithm of k.

Once we have made this move, it is not difficult to prove that Löb's logic L is sound with respect to Bounded Arithmetic (see subsection 2.3.3). Thus Gödel's Second Incompleteness Theorem holds, and its formalized version is provable.

Rosser strengthened Gödel's First Incompleteness Theorem by constructing an arithmetical sentence R that is independent of Peano Arithmetic, provided that Peano Arithmetic is consistent [Ros 36]. In chapter 3, we prove the formalized version of Rosser's Theorem in Bounded Arithmetic. In more technical terms, we prove the arithmetical sentence expressing "if Bounded Arithmetic is consistent, then neither R nor $\neg R$ are provable from it". Here R is constructed by Gödel's method of diagonalization; informally, R says "there is a proof of my negation which is smaller than any proof of myself". The proof of the formalized version of Rosser's Theorem does not come cheaply. We use almost the whole chapter in order to prove a "small reflection principle" on which the proof is based.

However, the real trouble only starts when we want to prove Solovay's Completeness Theorem for Bounded Arithmetic. As we mentioned, it is easy to prove that L is sound with respect to Bounded Arithmetic, but as far as we know the provability logic of Bounded Arithmetic might well be a proper extension of L. The problems we encounter when we want to adapt Solovay's construction to Bounded Arithmetic are discussed in chapter 3 and chapter 4. They are related to open problems in complexity theory.

In chapter 5, we show that formulas having models on suitably simple Kripke trees can be translated into arithmetical sentences that are consistent with Bounded Arithmetic. We also show that the provability logic of Bounded Arithmetic cannot be the modal theory of a class of Kripke trees. The general question 'What is the provability logic of Bounded Arithmetic' is still left open for future research.

Because interpretability logics always include a provability logic ($\neg A \rhd \bot$ being equivalent to $\Box A$), we do not yet have enough material to find the interpretability logic of Bounded Arithmetic.

1.5 Efficient metamathematics of inefficient mathematics

What happens when the techniques from efficient metamathematics of efficient mathematics are turned loose on normal mathematics? We may study interpretability between extensions of inefficient theories like Peano Arithmetic and Zermelo Fraenkel set theory. For example, it has been known for a long time that in every model of Zermelo Fraenkel set theory, we can see a model of Peano Arithmetic. Now we might ask: can we see it in an efficient way? In other words, is there a translation from the language of arithmetic into the language of set theory such that set theory proves all axioms of Peano arithmetic by proofs that are easy to compute given the original axioms? The answer turns out to be yes.

In chapter 6 we ask a more general question: are all interpretations that we know feasible? The answer is: yes and no! Yes, because well-known interpretations like that of ZF plus the negation of the continuum hypothesis into ZF are feasible. No, because we can use tricks like diagonalization to make some theories U, V such that U interprets V, but not by any feasible interpretation.

We then restict our attention to feasible interpretability between finite extensions of Peano Arithmetic. What is the logic of feasible interpretability over such theories? It turns out to be ILM, the same interpretability logic that is arithmetically sound and complete with respect to normal interpretability over Peano Arithmetic.

Finally, in chapter 7, we establish the intrinsic complexity of the formula "$PA + A$ feasibly interprets $PA+B$". Feasible interpretability over PA turns out to be Σ^0_2-complete, contrasting with the fact that standard interpretability over PA is Π^0_2-complete. We also prove that the formula "PA interprets $PA + A$ but not by any feasible interpretation" is Π^0_2-complete.

1.6 What to expect from the rest of the dissertation?

The remainder of part I contains some preliminaries needed to read parts II and III of the dissertation, as well as some material not covered in those chapters but interesting in its own right.

Part II

Chapter 3 *A small reflection principle for bounded arithmetic.* This is based on the paper [VV] written jointly with Albert Visser; which in its turn is based on [Ve 88] and [Ve 89].

Chapter 4 *Provable completeness for Σ_1-sentences implies something funny, even if it fails to smash the polynomial hierarchy.* This is based on unpublished work with Alexander Razborov. It is reproduced here with his permission.

Chapter 5 *On the provability logic of bounded arithmetic.* This is based on the paper [BV 91]. Preliminary results can be found in [Ve 88].

Part III

Chapter 6 *Feasible interpretability.* This is based on [Ve 93].

Chapter 7 *The complexity of feasible interpretability.* A previous version of this chapter has been submitted to the book *Feasible Mathematics II*, edited by J. Remmel et al., to appear with Birkhauser.

Chapter 2

Background

2.1 Theories of arithmetic

Definition 2.1.1 The language of arithmetic contains 0, S, $+$, $\cdot$, $=$ and $\leq$.

Definition 2.1.2 Robinson's Arithmetic Q is a theory in the language of arithmetic given by the following axioms:

Q1 $\forall x(Sx \neq 0)$;

Q2 $\forall x, y(Sx = Sy \rightarrow x = y)$;

Q3 $\forall x(x \neq 0 \rightarrow \exists y\, x = Sy)$;

Q4 $\forall x(x + 0 = x)$;

Q5 $\forall x, y(x + Sy = S(x + y))$;

Q6 $\forall x(x \cdot 0 = 0)$;

Q7 $\forall x, y(x \cdot Sy = (x \cdot y) + x)$;

Q8 $\forall x, y(x \leq y \leftrightarrow \exists z(z + x = y))$;

Definition 2.1.3 Peano Arithmetic PA contains the theory Q plus the induction scheme

$$\varphi(0) \wedge \forall x(\varphi(x) \rightarrow \varphi(Sx)) \rightarrow \forall x \varphi(x),$$

where φ may be any first-order formula in the language of arithmetic.

Remark 2.1.4 We presuppose familiarity with the arithmetical hierarchy (see a textbook on recursion theory, e.g. [So 87]). We remind the reader that a formula in the language of arithmetic is $\Delta_0^0 = \Sigma_0^0 = \Pi_0^0$ if all its quantifiers are *bounded*, which means that they are of the form $\forall x \leq t$ where t is any term not involving x. Σ_{n+1}^0-formulas have the form $\exists x \varphi$ for some φ in Π_n^0. Dually, Π_{n+1}^0-formulas have the form $\forall x \varphi$ for some φ in Σ_n^0.

Δ_0^0-formulas define primitive recursive relations of natural numbers, but not all primitive recursive relations are Δ_0^0 definable: only those from the so-called linear time hierarchy (see [HP 93, Definition V.2.10, Theorem V.2.16]). Σ_1^0-formulas define recursively enumerable (r.e) relations, whereas Π_1^0-formulas define co-r.e. relations.

We usually use Δ_0 for Δ_0^0, Σ_n for Σ_n^0 and Π_n for Π_n^0.

Definition 2.1.5 Let Γ be a class of formulas in the language of arithmetic. Then $I\Gamma$ contains the theory Q and additionally the induction scheme

$$\varphi(0) \land \forall x(\varphi(x) \to \varphi(Sx)) \to \forall x \varphi(x)$$

for $\varphi \in \Gamma$.

Remark 2.1.6 Note that the induction axioms for Δ_0-formulas can be written in Π_1-form, namely as

$$\forall y(\varphi(0) \land \forall x \leq y(\varphi(x) \to \varphi(Sx)) \to \forall x \leq y \varphi(x))$$

for $\varphi \in \Delta_0$.

Definition 2.1.7

- $\omega_1(0) = 0$;
- $\omega_1(x) = x^{|x|}$ for $x > 0$; here $|x| = \lceil \log_2(Sx) \rceil$.

Definition 2.1.8

- $2^x_0 = x$;
- $2^x_{k+1} = 2^{2^x_k}$.

We sometimes write $exp(x)$ instead of 2^x.

Many mathematicians, most notably Bennett, Paris, and Pudlák [Ben 62, Di 80, Pu 83] have constructed Δ_0-formulas corresponding to the relation $y = 2^x$. We refer the reader to [HP 93, Section V.3] for a clear description of the construction of such a formula.

Similarly, there exist Δ_0-formulas defining the relation $y = 2^x_z$, $y = \omega_1(x)$ and even to the graphs of most functions that we introduce in section 2.3. For those functions we use "$y = f(\vec{x})$" as shorthand for the appropriate Δ_0-formula. In $I\Delta_0$ we can prove the recursive clauses for these functions, as well as some other useful facts, for example that 2^x grows faster than ω_1.

Definition 2.1.9

- $EXP := \forall x \exists y(\text{"}y = 2^x\text{"})$;
- $SUPEXP := \forall x \exists y(\text{"}y = 2^x_x\text{"})$.

We have the following hierarchy of theories:

$$Q \subseteq I\Delta_0 \subseteq I\Delta_0 + EXP \subseteq I\Delta_0 + SUPEXP \subseteq I\Sigma_1 \subseteq I\Sigma_2 \subseteq \ldots \subseteq PA.$$

2.2 Provability logic

Definition 2.2.1 The language of modal logic contains a countable set of propositional variables, the propositional constant $\bot$, boolean connectives $\neg$, $\wedge$, $\rightarrow$, and the unary modal operator $\Box$. The modal provability logic L is axiomatized by all formulas having the form of propositional tautologies (including those containing the $\Box$-operator) plus the following axiom schemes:

1. $\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B)$
2. $\Box(\Box A \rightarrow A) \rightarrow \Box A$
3. $\Box A \rightarrow \Box\Box A$

The rules of inference are:

1. If $\vdash A \rightarrow B$ and $\vdash A$, then $\vdash B$ (modus ponens)
2. If $\vdash A$, then $\vdash \Box A$ (necessitation)

We suppose that the reader is familiar with Kripke frames and forcing relations on them. For L, finite Kripke trees give a good semantics. To make this precise, we state a definition and a theorem.

Definition 2.2.2 A *Kripke tree* is a frame $(K, \prec)$ in which

- $\prec$ is a strict partial ordering, i.e. it is transitive and asymmetric;
- for every element of K, the set of its predecessors is finite and linearly ordered by $\prec$;
- there is one root which precedes all other elements.

Löb's logic L is modally complete with respect to finite Kripke trees:

Theorem 2.2.3 *For every modal sentence A, the following are equivalent:*

1. $L \vdash A$;
2. *for all finite trees $(K, \prec)$ and points $k \in K$ and for all forcing relations $\Vdash$ on $(K, \prec)$, we have $k \Vdash A$;*
3. *for all finite trees $(K, \prec)$ and for all forcing relations $\Vdash$ on $(K, \prec)$, we have $k_0 \Vdash A$ where k_0 is the root of $(K, \prec)$.*

Proof. See [Sm 85, Theorem 2.2.3] QED

Definition 2.2.4 Let T be a theory in the language of arithmetic. A *T-interpretation* * is a function which assigns to each modal formula A a sentence A^* in the language of T, and which satisfies the following requirements:

1. $\bot^*$ is the sentence $0 = 1$.
2. * distributes over the boolean connectives, i.e. $(A \rightarrow B)^* = A^* \rightarrow B^*$, etc.

3. $(\Box A)^* = Prov_T(\ulcorner A^* \urcorner)$.

Here $Prov_T(\ulcorner A^* \urcorner)$ is the formalization of "A^* is provable from T" (see subsection 2.3.3).

Clearly * is uniquely determined by its restriction to the propositional variables. The presence in the modal language of the propositional constant $\bot$ allows us to consider closed modal formulas, i.e. modal formulas containing no propositional variables. If A is closed, then A^* does not depend on *, e.g. $(\Box\bot)^*$ is the arithmetical sentence $Prov_T(\ulcorner 0 = 1 \urcorner)$.

For arithmetical recursively enumerable theories T with $I\Delta_0 + EXP \subseteq T \subseteq PA$ we have for all modal formulas A:

$$L \vdash A \Longleftrightarrow \text{ for all } T\text{-interpretations } {}^*, T \vdash A^*.$$

The soundness direction ($\Rightarrow$) is not difficult to prove. It hinges on the observation that for $T \supseteq I\Delta_0 + EXP$ the formalized version of Löb's Theorem is provable:

$$T \vdash Prov_T(\ulcorner Prov_T(\ulcorner \varphi \urcorner) \rightarrow \varphi \urcorner) \rightarrow Prov_T(\ulcorner \varphi \urcorner).$$

Incidentally, the formalized version of Löb's Theorem immediately implies the formalized version of Gödel's Second Incompleteness Theorem by taking $\bot$ for φ:

$$T \vdash Prov_T(\ulcorner Prov_T(\ulcorner \bot \urcorner) \rightarrow \bot \urcorner) \rightarrow Prov_T(\ulcorner \bot \urcorner).$$

Solovay proved the completeness direction ($\Leftarrow$) for PA in his landmark paper [So 76]. It was pushed down to $I\Delta_0 + EXP$ in [JMM 91].

2.3 Bounded arithmetic

In this section we introduce classical arithmetical theories that are strictly weaker than $I\Delta_0 + EXP$. It turns out that there are two salient theories of this kind: Paris and Wilkie's $I\Delta_0 + \Omega_1$ [WP 87] and Buss' S^1_2 [Bu 86], both of them satisfying Löb's logic.

We will not state all the interesting results that appeared in the standard references to the area of weak arithmetics (see [Bu 86, WP 87] and Chapter V of [HP 93]). Instead we quickly review those concepts that we need in the sequel.

The principal feature distinguishing various theories of Bounded Arithmetic from Peano Arithmetic is that in the former induction is restricted to bounded formulas.

2.3.1 $I\Delta_0 + \Omega_1$

Definition 2.3.1 The language of $I\Delta_0 + \Omega_1$ as introduced in [WP 87] contains 0, S, $+$, $\cdot$, $=$ and $\leq$, and additionally the logical symbols $\neg, \rightarrow$ and $\forall$, and variables $v_1, v_2, \ldots$. With regard to logical axioms, we use a Hilbert-type system as in [WP 87], but other choices are reasonable too. For example, a Gentzen style sequent calculus with cut rule or natural deduction would do. However, we do not use a logic in which only direct proofs (i.e tableau proofs or cut-free proofs) are allowed.

As non-logical axioms we consider a set containing the following:

- a finite number of universal formulas defining the basic properties of the function and predicate symbols of the language:

  1. $0 \leq 0 \wedge \neg(S0 \leq 0)$;

2. $\forall x(x + 0 = x \land x \cdot 0 = 0 \land x \cdot S0 = x)$;
3. $\forall x \forall y(Sx = Sy \to x = y)$;
4. $\forall x \forall y(x \leq Sy \leftrightarrow (x \leq y \lor x = Sy))$;
5. $\forall x \forall y(x + Sy = S(x + y))$;
6. $\forall x \forall y(x \cdot Sy = (x \cdot y) + x)$;

- a formula $\forall x \exists y \varphi(x, y)$, where φ is the Δ_0-formula defining the relation $y = \omega_1(x)$ ($= x^{|x|}$; see definition 2.1.7);
- the scheme of induction for Δ_0-formulas.

2.3.2 Buss' systems of bounded arithmetic and the polynomial hierarchy

Definition 2.3.2 The language of Buss' bounded arithmetic consists of 0, S, $+$, $\cdot$, $=$, $\leq$, $|x|$ ($= \lceil \log_2(x + 1) \rceil$, the length of the binary representation of x), $\lfloor \frac{1}{2}x \rfloor$, and $x \# y$ ($= 2^{|x| \cdot |y|}$, the smash function).

Remark 2.3.3 Note that the smash function # allows us to express terms approximately equal to $2^{P(|x|)}$ for any polynomial P. More precisely, for every $n, x \geq 2$ the following holds:

$$2^{|x|^n} \leq \underbrace{x \# \ldots \# x}_{n \text{ times}} \leq 2^{2 \cdot |x|^n - 2},$$

as is easily proved by induction. This property of # is useful when we want to define polynomial time functions.

Definition 2.3.4 The *hierarchy of bounded arithmetic formulas* is defined as follows:

1. $\Sigma_0^b = \Pi_0^b = \Delta_0^b$ is the set of formulas with only sharply bounded quantifiers $\forall x \leq |t|$, $\exists x \leq |t|$ (where t is any term not involving x)

2. Σ_{k+1}^b is defined inductively by:

   - $\Sigma_{k+1}^b \supseteq \Pi_k^b$, and is closed under $\land$, $\exists x \leq t$ and $\forall x \leq |t|$;
   - if $B \in \Pi_{k+1}^b$, then $\neg B \in \Sigma_{k+1}^b$.

3. Π_{k+1}^b is defined inductively by:

   - $\Pi_{k+1}^b \supseteq \Sigma_k^b$, and is closed under $\land$, $\forall x \leq t$ and $\exists x \leq |t|$;
   - if $B \in \Sigma_{k+1}^b$, then $\neg B \in \Pi_{k+1}^b$.

4. Σ_{k+1}^b and Π_{k+1}^b are the smallest sets which satisfy 2,3.

Definition 2.3.5 If R is a theory and A a formula, we say that A is Δ_{k+1}^b with respect to R iff there are formulas $B \in \Sigma_{k+1}^b$ and $C \in \Pi_{k+1}^b$ such that $R \vdash A \leftrightarrow B$ and $R \vdash A \leftrightarrow C$.

We never leave out the superscripts b from the levels Σ_n^b and Π_n^b of Buss' bounded arithmetical hierarchy, so our use of Σ_n for Σ_n^0 and Π_n for Π_n^0 should not give rise to confusion.

The hierarchy of bounded arithmetic formulas is constructed in such a way that all levels Π_i^b and Σ_i^b except Σ_0^b correspond to levels of the polynomial hierarchy, which is well-known from structural complexity theory. Without defining all the basic notions of complexity theory, for which the reader may turn to [BDG 87], we give one of the standard definitions.

Definition 2.3.6 The *polynomial hierarchy* is defined as follows:

1. $P = \Delta_1^p$ is the set of predicates on the natural numbers which are recognized by a deterministic polynomial time Turing machine;

2. $NP = \Sigma_1^p$ is the set of predicates on the natural numbers which are recognized by a nondeterministic polynomial time Turing machine;

3. Σ_i^p is the set of predicates Q such that there is an $R \in \Delta_i^p$ and a polynomial P, such that for all $\vec{x}$, $Q(\vec{x}) \iff \exists y \leq 2^{P(|\vec{x}|)} R(\vec{x}, y)$.

4. Π_i^p is the set of predicates Q such that there is an $R \in \Sigma_i^p$, so that for all $\vec{x}$, $Q(\vec{x}) \iff \neg R(\vec{x})$.

5. Δ_{i+1}^p is the set of predicates which are recognized by a deterministic polynomial time Turing machine with some oracle from Σ_i^p.

As usual we use the name *co-NP* for Π_1^p. There are many open questions about the polynomial hierarchy. The most important one is: is there a k such that $\Sigma_k^p = \Sigma_{k+1}^p$, in which case the hierarchy collapses? More particularly, does $NP = co\text{-}NP$? Or even $P = NP$? It is also unknown whether for any k, $\Delta_k^p = \Sigma_k^p \cap \Pi_k^p$, and in particular whether $P = NP \cap co\text{-}NP$.

Definition 2.3.7 A is *polynomially reducible* to B if there is a polynomial time computable function f such that $\forall x(x \in A \leftrightarrow f(x) \in B)$.

Note that polynomial reducibility is analogous to many-one reducibility from ordinary recursion theory.

Definition 2.3.8 B is NP-complete if all $A \in NP$ are polynomially reducible to B. Similarly, B is co-NP-complete if all $A \in$ co-NP are polynomially reducible to B.

Remark 2.3.9 It is easy to see that for every NP-complete set B, the following holds:

- If $B \in$ co-NP, then $NP = co\text{-}NP$;

- If $B \in P$, then $P = NP$.

Remark 2.3.10 From results of Stockmeyer, Wrathall, and Kent and Hodgson [St 76, Wr 76, KH 82] it follows that the bounded arithmetical hierarchy is related to the polynomial hierarchy in the following way: Σ_{k+1}^p is the class of predicates which are defined by formulas in Σ_{k+1}^b. In particular, *NP* is the class of predicates which are defined by Σ_1^b-formulas; similarly *co-NP* is the class of predicates defined by Π_1^b-formulas. We refer the reader to [Bu 86, Chapter 1] for proofs of these correspondences.

Definition 2.3.11 The theory S_2^i consists of BASIC, a finite list of axioms defining the basic properties of symbols in the language of bounded arithmetic, plus the following induction scheme PIND(Σ_i^b):

$$A(0) \wedge \forall x(A(\lfloor \frac{1}{2}x \rfloor) \rightarrow A(x)) \rightarrow \forall x A(x)$$

for $A \in \Sigma_i^b$.

Definition 2.3.12 $S_2 := \bigcup_i S_2^i$.

Definition 2.3.13 The theory T_2^i consists of BASIC plus the following induction scheme:

$$A(0) \wedge \forall x(A(x) \rightarrow A(Sx)) \rightarrow \forall x A(x)$$

for $A \in \Sigma_i^b$.

Definition 2.3.14 $T_2 := \bigcup_i T_2^i$.

Buss proves that for each i, $S_2^{i+1} \vdash T_2^i$ (see [Bu 86, Corollary 2.21]). It is clear that also for each $i \geq 1$, $T_2^i \vdash S_2^i$. Thus, $T_2 = S_2$.

One of the most important theorems about bounded arithmetic is Parikh's Theorem. It implies that every Δ_0-definable provably total function of S_2 can increase the length of its input only polynomially.

Parikh originally proved his theorem for $I\Delta_0$, for which the Δ_0-definable provably total functions are even more severely limited than for S_2: they can increase the length of the input only linearly.

We state a version of Parikh's Theorem for Buss' theories S_2^i.

Theorem 2.3.15 (Parikh's Theorem) *Let $i \geq 0$. Suppose that φ is a bounded formula and that $S_2^i \vdash \forall x \exists y \varphi(x,y)$. Then there is a term $t(x)$ such that $S_2^i \vdash \forall x \exists y \leq t(x) \varphi(x,y)$.*

Proof. Buss gives a proof-theoretic proof (see [Bu 86, Theorem 4.11]). However we prefer to give a model-theoretic proof, because it is easier to understand and much shorter. So suppose that there is a bounded formula φ such that $S_2^i \vdash \forall x \exists y \varphi(x,y)$, but for every term $t(x)$, $S_2^i \nvdash \forall x \exists y \leq t(x)\varphi(x,y)$. Now if c is a fresh constant, the set of formulas

$$S_2^i + \{\forall y \leq \underbrace{c\#\ldots\#c}_{k\ times} \neg\varphi(c,y) \mid k \in \omega\}$$

is finitely satisfiable. Thus by the Compactness Theorem there is a model

$$\mathcal{M} \models S_2^i + \{\forall y \leq \underbrace{c\#\ldots\#c}_{k\ times} \neg\varphi(c,y) \mid k \in \omega\}.$$

Suppose that a is the interpretation of c in this model. Next, take the submodel $\mathcal{M}^*$ of $\mathcal{M}$ defined by:

$$\mathcal{M}^* := \{b \in \mathcal{M} \mid \exists n \in \omega(b \leq \underbrace{c\#\ldots\#c}_{n\ times})\}.$$

It is easy to check that $\mathcal{M}^*$ is closed under 0, S, $+$, $\cdot$, $\leq$, $|x|$, $\lfloor \frac{1}{2}x \rfloor$, and $\#$. Moreover the induction axioms of S_2^i can be written in Π_1-form, so they still hold in $\mathcal{M}$'s initial segment $\mathcal{M}^*$. Therefore $\mathcal{M}^* \models S_2^i$, but $\mathcal{M}^* \not\models \exists y \varphi(a,y)$, contradicting our first assumption that $S_2^i \vdash \forall x \exists y \varphi(x,y)$. QED

Definition 2.3.16 $\Box_1^p$-functions are those computable by a polynomial time Turing machine. For $i > 1$, $\Box_i^p$ contains those functions computable by a polynomial time Turing machine from finitely many oracles in Σ_{i-1}^p.

Buss proved that the provably total Σ_1^b-definable functions of S_2^1 are exactly the functions computable by a polynomial time Turing machine. More precisely, and at the same time more generally, we have the following two theorems:

Theorem 2.3.17 *Let $i \geq 1$. Let g be an m-ary $\Box_i^p$-function. Let $t(\vec{x})$ be a term so that for all $\vec{x} \in \omega^m$, $g(\vec{x}) \leq t(\vec{x})$. Then there is a Σ_i^b-formula A such that:*

1. *$S_2^i \vdash \forall \vec{x} \exists y \leq t A(\vec{x}, y)$;*
2. *$S_2^i \vdash \forall \vec{x}, y, z(A(\vec{x}, y) \wedge A(\vec{x}, z) \rightarrow y = z)$;*
3. *For all $\vec{x} \in \omega^m$, $A(\vec{x}, g(\vec{x}))$ is true.*

Proof. See [Bu 86, Theorem 3.1] QED

Theorem 2.3.18 (Buss' Main Theorem) *Let $i \geq 1$. Suppose $S_2^i \vdash \forall \vec{x} \exists y A(x, y)$ where $A(x, y)$ is a Σ_i^b-formula with only $\vec{x}, y$ free. Then there is a term $t(\vec{x})$, a Σ_i^b-formula B and a function g in $\Box_i^p$ such that*

1. *$S_2^i \vdash \forall \vec{x} \forall y(B(\vec{x}, y) \rightarrow A(\vec{x}, y))$;*
2. *$S_2^i \vdash \forall \vec{x} \forall y, z(B(\vec{x}, y) \wedge B(\vec{x}, z) \rightarrow y = z)$;*
3. *$S_2^i \vdash \forall \vec{x} \exists y \leq t B(\vec{x}, y)$;*
4. *For all $\vec{n}$, $\omega \models B(\vec{n}, g(\vec{n}))$.*

Proof. See [Bu 86]. Buss uses methods well-known from proof theory. We give a short sketch. Suppose $S_2^i \vdash \forall \vec{x} \exists y A(x, y)$, by a proof p. Then we can apply cut elimination to obtain a term t and an S_2^i proof p' of $\forall \vec{x} \exists y \leq t A(x, y)$ that only cuts Σ_i^b-formulas. Thus, p' contains only Σ_i^b and Π_i^b-formulas. Next we can directly extract from p' a $\Box_i^p$-algorithm for computing a function g such that for all $\vec{n}$, $\omega \models A(\vec{n}, g(\vec{n}))$.

For an elegant model-theoretical argument, which is inspired by Visser's unpublished proof of Parson's and Mints' theorem [Pa 72, Mi 71] that the primitive recursive functions are exactly the provably total functions of $I\Sigma_1$, see [Za 93]. QED

Remark 2.3.19 Note that, if for some Σ_i^b-formula A and some term $t(\vec{x})$, $\forall \vec{x} \exists y \leq t A(\vec{x}, y)$ is *true* but not necessarily provable in S_2^i, then we know only that there is a witnessing function in $\Box_{i+1}^p$.

Corollary 2.3.20 *Let $A(\vec{a})$ be a formula such that S_2^1 proves that A is equivalent to a Σ_1^b- and to a Π_1^b-formula. Then A is a polynomial time predicate.*

In other words, if S_2^1 *proves* that some predicate is in $NP \cap co\text{-}NP$, then it is already in P. We remind the reader that it is an open question whether $NP \cap co\text{-}NP = P$.

Even though we do not need it in the sequel of the dissertation, we cannot resist the temptation to end this introduction to relations between complexity theory and bounded arithmetic with the statement of a beautiful result by Krajíček, Pudlák and Takeuti.

Theorem 2.3.21 *For $i \geq 0$, if $T_2^i \vdash S_2^{i+1}$, then $\Sigma_{i+2}^p = \Pi_{i+2}^p$.*

Proof. For the very ingenious argument, see [KPT 89]. QED

2.3.3 Metamathematics for bounded arithmetic

In order to prove Gödel's Incompleteness Theorems for bounded arithmetic, Buss arithmetized the usual notions of metamathematics (see [Bu 86, Chapter 7]). It turns out that most predicates needed can be Δ_1^b-defined (or sometimes $\exists\Delta_1^b$-defined) in S_2^1. Moreover, these definitions are intensionally correct in the sense of [Fe 60], which means that the usual connections between them can be proved in S_2^1.

Here follows a list of predicates used in the sequel.

- $\mathit{Seq}(w)$ for "w encodes a sequence";
- $\mathit{Len}(w) = a$ for "if w encodes a sequence, then the length of that sequence is a; otherwise $a = 0$";
- $\mathit{Term}(v)$ for "v is the Gödel number of a term";
- $\mathit{Fmla}(v)$ for "v is the Gödel number of a formula";
- $\mathit{Prf}_T(u, v)$ for $\mathit{Fmla}(v) \wedge$ "u is the Gödel number of a proof in T of the formula with Gödel number v"; when T is clear from the context, we drop the subscript.
- $\mathit{Prov}_T(v) := \exists u \mathit{Prf}_T(u, v)$; we sometimes abbreviate $\mathit{Prov}(\ulcorner\varphi\urcorner)$ as $\Box\varphi$.

The predicates Seq, Len, Term, and Fmla are Δ_1^b-definable in S_2^1, and so is Prf_α where the formula α is Δ_1^b with respect to S_2^1. The condition on α is not a severe restriction. To any recursively enumerable set one can associate a polynomial time function having that set as its range, therefore one can suitably axiomatize any theory T which has a recursively enumerable set of axioms including BASIC.

Notation 2.3.22 Instead of the usual *numerals* S^k0 of Peano Arithmetic, we use canonical numerals $\bar{k}$ defined inductively by:

- $\bar{0} = 0$;
- $\overline{2k+1} = \overline{2k} + (S0)$;
- $\overline{2k+2} = (SS0) \cdot (\overline{k+1})$.

Note that the length of the term $\bar{k}$ is linear in the length of the binary representation of k, a property that the S^k0 obviously do not satisfy. The shortness of canonical terms plays a crucial rôle in many proofs, for example in Buss' proof that S_2^1 enjoys provable Σ_1^b-completeness (see [Bu 86, Theorem 7.4]).

S_2^1 can Σ_1^b-define a function $\mathit{Num}(x)$ such that $\mathit{Num}(x)$ stands for the Gödel number of the term $\bar{x}$. For ease of reading, we will however abuse notation; thus if $A(x)$ is a formula with free variable x we write $\ulcorner A(\bar{a})\urcorner$ instead of $\mathit{Sub}(\ulcorner A\urcorner, \ulcorner x\urcorner, \mathit{Num}(a))$. Sometimes we are even more sloppy and leave out the numeral dashes altogether. In those cases the context should provide enough material for the reader to know what is meant.

Lemma 2.3.23 (cf. Lemma 7.5 of [Bu 86])
Let t be any term with free variables $a_1, \ldots, a_k$. Then

$$S_2^1 \vdash \forall a_1, \ldots, a_k \mathit{Prov}(\ulcorner t(\overline{a_1}, \ldots, \overline{a_k}) = \overline{t(a_1, \ldots, a_k)}\urcorner).$$

Proof. We use induction on the complexity of t.

- If t is 0, then the equality axioms immediately give us $S^1_2 \vdash Prov(\ulcorner \bar{0} = \bar{0} \urcorner)$.
- If t is a variable symbol x, then the equality axioms give us $S^1_2 \vdash \forall x Prov(\ulcorner \bar{x} = \bar{x} \urcorner)$.
- If $t = Sr$, then we have by induction hypothesis

  $$S^1_2 \vdash \forall a_1, \ldots, a_k Prov(\ulcorner r(\overline{a_1}, \ldots, \overline{a_k}) = \overline{r(a_1, \ldots, a_k)} \urcorner).$$

  Therefore it suffices to show that $S^1_2 \vdash \forall b Prov(\ulcorner S\bar{b} = \overline{Sb} \urcorner)$. In order to be able to apply the formalized version of $PIND(\Sigma^b_1)$ in S^1_2, we need to find some fixed polynomial P such that the proofs of $S\bar{b} = \overline{Sb}$ are of length $\leq P(|b|)$, because then there is a term $t(x)$ for which we can prove $S^1_2 \vdash \forall b \exists y \leq t(b) Prf(y, \ulcorner S\bar{b} = \overline{Sb} \urcorner)$. The polynomial that we need will be quadratic. We leave the exact computation to the reader. Informally, the reasoning inside S^1_2 is as follows:

  - We clearly have $Prov(\ulcorner S\bar{0} = \overline{S0} \urcorner)$.
  - Suppose that $b > 0$ is even. By the definition of efficient numerals and the BASIC axioms we immediately have proofs of length linear in $|b|$ of $S\bar{b} = \bar{b} + S0 = \overline{Sb}$.
  - Suppose that $b > 0$ is odd. Then we have a proof from BASIC of length linear in $|b|$ of $S\bar{b} = 2 \cdot \overline{\lfloor \frac{1}{2} b \rfloor} + 2 = 2 \cdot S\overline{\lfloor \frac{1}{2} b \rfloor}$.
    By the induction hypothesis and the BASIC axioms we have a proof of length quadratic in $||\lfloor \frac{1}{2} b \rfloor|| = |b| - 1$ of $S\overline{\lfloor \frac{1}{2} b \rfloor} = \overline{S\lfloor \frac{1}{2} b \rfloor}$. Combined, these two give a proof of length quadratic in $|b|$ of $Sb = S\bar{b}$.

- We leave the cases for $+, \cdot, \#, \lceil \frac{1}{2} x \rceil$, and $|\ |$ to the reader.

QED

Theorem 2.3.24 (Provable Σ^b_1-completeness, Buss) *Let A be any Σ^b_1-formula. Let $a_1, \ldots, a_k$ be all the free variables of A. Then there is a term $t(a_1, \ldots, a_k)$ such that*

$$S^1_2 \vdash \forall a_1, \ldots, a_k (A(a_1, \ldots, a_k) \rightarrow \exists w \leq t Prf(w, \ulcorner A(\overline{a_1}, \ldots, \overline{a_k}) \urcorner)).$$

Proof. We give only a small hint: the reader may look up the full proof in [Bu 86, Theorem 7.4]. We use induction on the complexity of A. The most difficult step is the one for the bounded universal quantifier. So suppose that A is $\forall x \leq |s| B(a_1, \ldots, a_k, x)$, and that for all $b \leq |s|$ we have proofs of length polynomial in $\max(|a_1|, \ldots, |a_k|, |b|)$ of $B(a_1, \ldots, a_k, b)$. Then we can combine these $|s| + 1$ short proofs in order to construct a proof of length polynomial in $\max(|a_1|, \ldots, |a_k|, |s(a_1, \ldots, a_k)|)$ of the formula
$\forall x \leq |s| B(a_1, \ldots, a_k, x)$. QED

Using theorem 2.3.24, it is easy to see that Löb's logic (see definition 2.2.1) is arithmetically sound with respect to S^1_2. In particular this means that we can, in the standard way, prove Gödel's Second Incompleteness Theorem and its formalized version for S^1_2.

Theorem 2.3.25 (cf. Theorem 7.10 of [Bu 86])
$S^1_2 \nvdash \neg Prov(\ulcorner \bot \urcorner)$ and $S^1_2 \vdash Prov(\ulcorner Prov(\ulcorner \bot \urcorner) \rightarrow \bot \urcorner) \rightarrow Prov(\ulcorner \bot \urcorner)$.

Proof. We leave the well-known proofs to the reader. QED

We refer the reader to section 2.6 for the proof of a much stronger result: $I\Delta_0 + EXP \nvdash Con(Q)$.

By the way, Rosser's strengthening of Gödel's Second Incompleteness Theorem is also provable for theories like S^1_2 and $I\Delta_0 + \Omega_1$. However, we use the "small reflection theorem" to prove it. Thus the reader will have to wait for theorem 3.3.24 of Chapter 3 in order to find Rosser's Theorem as a corollary.

Sometimes, we will use the name $I\Delta_0 + \Omega_1$ for Buss' theory S_2 (see Definition 2.3.12), in which induction for formulas from the hierarchy of bounded arithmetic formulas in a language containing # and $|\ |$ is allowed. Because S_2 is a conservative extension of $I\Delta_0 + \Omega_1$, the name change has no repercussions on results that do not hinge on the details of formalization. More formally, we have the following:

Lemma 2.3.26 *There is a Δ_0-formula $\psi(x, y, z)$ such that $\omega \models x\#y = z \leftrightarrow \psi(x, y, z)$, $I\Delta_0 + \Omega_1$ proves the BASIC properties of # for ψ, and*

1. $I\Delta_0 + \Omega_1 \vdash \forall x, y \exists z \psi(x, y, z)$ *and*

2. $I\Delta_0 + \Omega_1 \vdash \forall x, y, z_1, z_2(\psi(x, y, z_1) \wedge \psi(x, y, z_2) \rightarrow z_1 = z_2)$.

Similarly there is a Δ_0-formula $\chi(x, y)$ defining $|\ |$ in $I\Delta_0 + \Omega_1$, and there is a Δ_0-formula $\xi(x, y)$ defining $\lfloor \frac{1}{2}x \rfloor$ in $I\Delta_0 + \Omega_1$.

Definition 2.3.27 A $\Delta_0(f_1, \ldots, f_n)$-formula is a bounded formula in the language of arithmetic to which the function symbols $f_1, \ldots, f_n$ are added.

Lemma 2.3.28

$$I\Delta_0 + \Omega_1 + \forall xyz((x\#y = z \leftrightarrow \psi(x, y, z)) \wedge (|x| = y \leftrightarrow \chi(x, y)) \wedge (\lfloor \tfrac{1}{2}x \rfloor = y \leftrightarrow \xi(x, y))$$
$$\vdash I\Delta_0(\#, |\ |, \lfloor \tfrac{1}{2}x \rfloor) \wedge \text{"}BASIC\text{"}.$$

Proof. As [Bu 86, Theorem 2.2 and Corollary 2.3]. See also [PD 82] for a more general lemma. QED

Definition 2.3.29 (see [Pu 85]). A theory T is *sequential* if it is a theory with equality, there is a distinguished provably non-empty domain $N(x)$ that satisfies Robinson's Arithmetic Q, and there exists a formula $\beta(t, w, z)$ ("t is the w-th element of the sequence coded by z") such that:

$$\begin{aligned} T \vdash \forall x, y, v \exists z \forall t, w \quad & [N(v) \wedge w \leq v \\ & \rightarrow (\beta(t, w, z) \leftrightarrow ((\beta(t, w, x) \wedge w < v) \vee (t = y \wedge w = v)))]. \end{aligned}$$

Examples of sequential theories are $I\Delta_0$, S^i_2 for $i \geq 1$, PA, ZF, and GB.

2.4 Interpretations

Tarski introduced the formal notion of interpretability in [TMR 53]. We give a variant of his definition here.

Let U, V be two Σ_1^b-axiomatized theories in languages containing finitely many non-logical symbols. Let the axioms of V be given by the Σ_1^b-formula α_V. An interpretation K of V into U is given by:

- a formula $\delta(x)$ of L_U defining the universe, such that $U \vdash \exists x \delta(x)$;
- a function from the relation symbols of L_V to formulas of L_U, respecting the original arities;
- a function from the function symbols f of L_V to formulas ψ_f of L_U, such that if f is k-ary, then ψ_f has $k+1$ free variables and

$$U \vdash \delta(x_1) \wedge \ldots \wedge \delta(x_k) \rightarrow \exists! y(\delta(y) \wedge \psi_f(x_1, \ldots, x_k, y)).$$

We warn the reader that the image of $=$ need not be $=$.

We can extend K in the obvious way to map all formulas φ of L_V into formulas φ^K of L_U. To do this we relativize all quantifiers to δ while we respect the Boolean connectives. In fact we can, in an intensionally correct way, Δ_1^b-define in $\mathrm{I}\Delta_0 + \Omega_1$ a function K corresponding to this mapping. For ease of reading, we will write a^K even if a is a Gödel number. Thus $U \rhd V$ can be defined in as follows:

$$U \rhd V \;:\leftrightarrow\; \exists K(\text{“}K \textit{is an interpretation}\text{”} \wedge \forall a(\alpha_V(a) \rightarrow \exists p Prf_U(p, a^K))). \tag{2.1}$$

(By abuse of notation we denote by $\rhd$ both the arithmetization of the interpretability predicate and the corresponding modal operator that we will introduce in section 2.5.)

It would be nice to be able to prove in the theories that we are interested in such as $\mathrm{I}\Delta_0 + \Omega_1$ that interpretability gives rise to relative consistency. However it seems that one cannot do this straight-away, but one needs a collection principle:

Definition 2.4.1 $\mathrm{B}\Sigma_1 \;:=\; \mathrm{I}\Delta_0$ plus the scheme

$$\forall u(\forall x < u \exists y \varphi(x, y) \rightarrow \exists v \forall x < u \exists y < v \varphi(x, y))$$

for $\varphi \in \Sigma_1$.

Indeed we have

$$\mathrm{B}\Sigma_1 + \Omega_1 \vdash U \rhd V \rightarrow (Con(U) \rightarrow Con(V)).$$

We leave the proof to the reader.

However, we prefer to make a definitional move. For the remainder of the dissertation, we define $U \rhd V$ as *smooth* interpretability (discussed in [Vi 91a]):

$$U \rhd V \;:\leftrightarrow\; \exists K(\text{“}K \textit{is an interpr.}\text{”} \wedge \forall u \exists v \forall a < u \exists p < v(\alpha_V(a) \rightarrow Prf_U(p, a^K))). \tag{2.2}$$

Now we do have

$$\mathrm{I}\Delta_0 + \Omega_1 \vdash U \rhd V \rightarrow (Con(U) \rightarrow Con(V)).$$

For theories containing $B\Sigma_1 + \Omega_1$ the definitions of standard interpretability (2.1) and smooth interpretability (2.2) collapse.

In Part III of this dissertation, we are concerned mostly with extensions of $PA \supseteq B\Sigma_1 + \Omega_1$, thus we may freely use the standard definition. Moreover, for extensions of PA one can, and we will, without loss of generality assume the image of = to be =. This restriction makes life easier, although it is not essential for most results.

We can view interpretability in a semantic way. An interpretation K of V into U determines, in every model $\mathcal{M}$ of U, a new model $\mathcal{M}^K$ with underlying set $\{a \in \mathcal{M} \mid \mathcal{M} \models \delta(a)\}$. The reader may check that for every $a_1, \ldots, a_k \in \mathcal{M}^K$, we have the following:

$$\mathcal{M}^K \models \varphi(a_1, \ldots, a_k) \iff \mathcal{M} \models \varphi(a_1, \ldots, a_k)^K.$$

For finitely axiomatized theories V, Montague in [Mo 65] was the first to explicitly relate the syntactic and semantic definitions of $U \rhd V$.

It is interesting to note that many famous relative consistency proofs in the mathematical literature arise from interpretations. Thus we have both $ZF \rhd ZF + CH$ and $ZFC \rhd ZFC + \neg CH$.

For arithmetical theories extending PA, Hájek [Há 71] gave an elegant characterization of interpretability. Because the characterization was implicit already in Orey's [Or 61], many authors refer to it as the Orey-Hájek characterization. In order to describe it we first need two definitions and a lemma.

Definition 2.4.2 $Prov_{k,T}(\ulcorner A \urcorner)$ stands for "there is a proof of A from T in which only axioms with Gödel number $\leq k$ are used".

$$Con_k(T) := \neg Prov_{k,T}(\ulcorner \bot \urcorner).$$

Definition 2.4.3 A theory T with $T \supseteq I\Sigma_1$ in the language of arithmetic is *essentially reflexive* if for all sentences A and for all k, $T \vdash Prov_{k,T}(\ulcorner A \urcorner) \to A$.

Remark 2.4.4 In the literature, different definitions of essential reflexivity abound. For example, [HP 93, Definition III-2.33] is as follows: A theory T with $T \supseteq I\Sigma_1$ in the language of arithmetic is essentially reflexive if for each theory $T' \supseteq T$ and for all k, $T' \vdash Con_k(T')$.

Lemma 2.4.5 *If T is an extension of PA by a primitive recursive set of axioms in the language of PA, then T is essentially reflexive.*

Proof. See [Ber 90, Theorem 2.6]. A feasible version can be found as our lemma 7.2.5. QED

Theorem 2.4.6 (Orey-Hájek, [Or 61], [Há 71]) *Let U and V be primitive recursive extensions of PA in the language of PA. Then the following holds:*

$$PA \vdash U \rhd V \leftrightarrow \forall k Prov_U(\ulcorner Con_k(V) \urcorner).$$

Proof. See [Ber 90, Theorem 2.9 and Remark 2.10]. For a feasible version, see lemma 7.3.1 of this dissertation. QED

Definition 2.4.7 A theory V is Π_1-*conservative over U* if for all Π_1-sentences π, $V \vdash \pi \Rightarrow U \vdash \pi$.

We abbreviate the formalization of "V is Π_1-conservative over U" as $U \rhd_{\Pi_1} V$.

Theorem 2.4.8 *Let U and V be primitive recursive extensions of PA in the language of PA. Then the following holds:*

$$PA \vdash U \rhd_{\Pi_1} V \leftrightarrow \forall k Prov_U(\ulcorner Con_k(V)\urcorner).$$

Proof. See [HP 93, Theorem III-2.40]. A feasible version appears as lemma 7.3.2. QED

Theorem 2.4.9 *Let U and V be primitive recursive extensions of PA in the same language. Then the following holds:*

$$PA \vdash U \rhd V \leftrightarrow U \rhd_{\Pi_1} V.$$

Proof. Immediately from theorem 2.4.6 and theorem 2.4.8. See corollary 7.3.3 for a feasible version. QED

Although the modal principle M (see section 2.5) corresponding to the following theorem was baptized 'Montagna's Principle' in the eighties, the unformalized version of the underlying theorem was proved by Lindström already in the seventies.

Theorem 2.4.10 (Lindström, [Li 79]) *Let U and V be primitive recursive extensions of PA in the language of PA. Let $\mathcal{S}$ be a primitive recursive set of Σ^0_1-sentences. Then the following holds:*

$$PA \vdash U \rhd V \rightarrow U + \mathcal{S} \rhd V + \mathcal{S}.$$

Proof. We remind the reader that for interpretations between theories extending PA, we take the image of $=$ to be $=$.

A precise formal proof can be gleaned from the proof of theorem 6.3.10 in chapter 6. Here we give a sketch with a more model-theoretic flavor. It is easy to see that the following fact implies our theorem:

> Let $\mathcal{M}$ be a model of U, and let K be an interpretation of V into the theory of $\mathcal{M}$; we call the interpreted structure $\mathcal{M}^K$. Then $\mathcal{M}$ can be embedded as an initial segment of $\mathcal{M}^K$.

In order to prove this fact, we define *pism(s)* for "s is a partial isomorphism" and the relation $G(x, y)$ as follows:

$$\begin{array}{lcl} pism(s) & := & seq(s) \wedge (s)_0 = 0^K \wedge \forall i < lh(s) - 1((s)_{i+1} = S^K(s)_i) \\ G(j,y) & := & \exists s(pism(s) \wedge lh(s) = j + 1 \wedge (s)_j = y) \end{array}$$

By induction it follows that for every $x \in \mathcal{M}$ there is a unique $y \in \mathcal{M}^K$ such that $\mathcal{M} \models G(x, y)$. Therefore, there is a function g corresponding to G. It is easy to see that g is an embedding into $\mathcal{M}^K$ and that it preserves $0, S, +$ and $\cdot$.

Now we need only show that the image of $\mathcal{M}$ is an initial segment of $\mathcal{M}^K$. But because $V \supseteq PA$, and $\mathcal{M} \models \forall x(g(Sx) = S^K(g(x)))$, this is not difficult: we have $\mathcal{M} \models \forall x \forall u(\delta(u) \wedge u <^K g(Sx) \rightarrow u <^K g(x) \vee u = g(x))$. Now by induction on $x \in \mathcal{M}$ we find that for every $u \in \mathcal{M}^K$ such that $\mathcal{M} \models u <^K g(x)$, there is a $y \in \mathcal{M}$ such that $\mathcal{M} \models y < x$ and $\mathcal{M} \models u = g(y)$. QED

For finitely axiomatized theories the situation is different. As reflexivity for such theories would be in contradiction with Gödel's Second Incompleteness Theorem, the Orey-Hájek characterization is not applicable (see however section 2.7).

Instead, we have Friedman's characterization. Readers unfamiliar with tableau provability may find a description in definition 2.8.2. We define

$$Tabprov_T(\ulcorner\varphi\urcorner) := \neg Tabcon(T + \neg\varphi).$$

Theorem 2.4.11 (Friedman's characterization) *Suppose U and V are sequential theories and V is finitely axiomatized. Then*

$$\mathrm{I}\Delta_0 + EXP \vdash U \rhd V \leftrightarrow Tabprov_{\mathrm{I}\Delta_0+EXP}(\ulcorner Tabcon(U) \to Tabcon(V)\urcorner).$$

Proof. See [Vi 90a]. QED

It is also clear that for finitely axiomatized theories U, V, $U \rhd V$ is $\exists\Sigma_1^b$, so

$$\mathrm{I}\Delta_0 + \Omega_1 \vdash U \rhd V \to \Box_Q(U \rhd V).$$

Examples of finitely axiomatizable sequential theories are $\mathrm{I}\Delta_0 + EXP$ (even verifiably in $\mathrm{I}\Delta_0 + \Omega_1$; see lemma 2.8.22), $\mathrm{I}\Delta_0 + SUPEXP$, and S_2^i and $I\Sigma_i$ for $i \geq 1$. At the moment of writing, as far as we heard nobody knows whether $\mathrm{I}\Delta_0$ and $\mathrm{I}\Delta_0 + \Omega_1$ are finitely axiomatizable.

2.5 Interpretability logic

Interpretability logic extends provability logic. The modal formulas $A \rhd B$ correspond to arithmetical formulas $T + A^* \rhd T + B^*$, where T is an arithmetical theory.

Definition 2.5.1 IL contains the provability logic L (see definition 2.2.1) plus the following five axiom schemes:

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

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

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

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

J5 $\Diamond A \rhd A$.

Definition 2.5.2 $ILM = IL + M$, where M is the axiom $(A \rhd B) \to (A \wedge \Box C \rhd B \wedge \Box C)$.

Definition 2.5.3 $ILP = IL + P$, where P is the axiom $(A \rhd B) \to \Box(A \rhd B)$.

Definition 2.5.4 An IL-frame is a frame (W, R), where R is a transitive conversely well-founded relation on W, with additional relations S_w for each $w \in W$, having the following properties:

- S_w is a relation on $\{w' \in W \mid wRw'\}$;
- S_w is reflexive and transitive;

- if wRw', wRw'' and $w'Rw''$, then $w'S_w w''$.

Definition 2.5.5 An ILM-frame is an IL-frame satisfying the following extra condition:

- if $uS_w vRz$, then uRz.

Definition 2.5.6 A *simplified* ILM-frame is a frame (W, R), where R is a transitive conversely well-founded relation on W, with root b say, with one additional binary relation S such that

- S is reflexive and transitive and $R \subseteq S$;
- if $uSvRz$, then uRz.

Definition 2.5.7 An IL-model is given by an IL-frame $(W, R, \{S_w \mid w \in W\})$ combined with a forcing relation satisfying the following clauses:

- $u \Vdash \Box A$ if and only if $\forall v(uRv \Rightarrow v \Vdash A)$;
- $u \Vdash A \rhd B$ if and only if $\forall v(uRv$ and $v \Vdash A \Rightarrow \exists w(vS_u w$ and $w \Vdash B))$.

In [JV 90], de Jongh and Veltman prove that IL is modally sound and complete with respect to IL-models, and that ILM is modally sound and complete with respect to IL-models on ILM-frames.

Visser showed that ILM is already complete with respect to models on *simplified* ILM-frames; for a proof see [Ber 90].

Definition 2.5.8 A T-interpretation is a map * which assigns to every propositional variable p a sentence p^* of the language of T, and which is extended to all modal formulas as follows:

1. $(A \rhd B)^* = T + A^* \rhd T + B^*$
2. $(\Box A)^* = Prov_T(A^*)$
3. * commutes with the propositional connectives.

Here $\rhd$ abbreviates the formalization of (smooth) interpretability.

IL is arithmetically sound with respect to sequential theories extending $I\Delta_0 + \Omega_1$.

Smoryński and Visser proved that ILP is arithmetically sound and complete with respect to the finitely axiomatized theories GB and ACA_0. Next Visser generalized the result and proved arithmetical soundness of ILP with respect to finitely axiomatized sequential theories extending $I\Delta_0 + SUPEXP$ (see [Vi 90a]). This means that for such theories T and for all modal formulas A,

$$ILP \vdash A \iff \text{for all } T\text{-interpretations } {}^*, T \vdash A^*.$$

Berarducci [Ber 90] and Shavrukov [Sh 88] independently proved that ILM is arithmetically complete with respect to interpretability over PA. It is also arithmetically sound (see [Vi 90a]).

2.6 Definable cuts

Because PA proves induction for all first order formulas, no proper cuts of models of PA can be defined by formulas. In the context of weaker theories where induction is restricted to a proper subset of all formulas, on the contrary, definable cuts have proved to be highly useful tools.

For example, in weak theories like $\mathrm{I}\Delta_0$, fast growing functions such as *exp* are not provably total. For results that one normally derives using the totality of such functions, one can find analogs in weak theories by constructing small cuts of numbers for which some of the fast growing functions in question do have the necessary values. The reader will find a formalization of this intuition in lemma 2.6.9, lemma 2.6.11 and theorem 2.6.12.

Moreover definable cuts provide very natural interpretations in which the domain is restricted, but the original operations are left intact. We give examples of interpretations by definable cuts in lemma 2.6.14 and theorem 2.6.16. Such interpretations in turn give rise to relative consistency results which are provable in theories as weak as $\mathrm{I}\Delta_0 + \Omega_1$.

It is time to give some formal definitions.

Definition 2.6.1 Let $T \supseteq Q$ be a Σ_1^b-axiomatized theory. A T-cut is a formula I such that:

1. $T \vdash I(0)$,

2. $T \vdash \forall x \forall y (I(y) \wedge x \leq y \rightarrow I(x))$,

3. $T \vdash \forall x (I(x) \rightarrow I(Sx))$.

Definition 2.6.2 Let $T \supseteq Q$ be a Σ_1^b-axiomatized theory. A T-initial segment is a formula J such that:

1. $T \vdash J(0)$,

2. $T \vdash \forall x \forall y (J(y) \wedge x \leq y \rightarrow J(x))$,

3. $T \vdash \forall x \forall y (J(x) \wedge J(y) \rightarrow (J(Sx) \wedge J(x+y) \wedge J(x \cdot y)))$.

Remark 2.6.3 Note that if J is a T-initial segment in an arithmetic language, it determines an initial substructure in every model of T. Because J is T-provably closed under $0, S, +$ and $\cdot$ and because the induction axioms of $\mathrm{I}\Delta_0$ can be written in Π_1^0-form (see remark 2.1.6), these substructures will themselves be models of $\mathrm{I}\Delta_0$. Thus T-initial segments provide interpretations of $\mathrm{I}\Delta_0$ into T.

Remark 2.6.4 For cuts I, we frequently write $x \in I$ instead of $I(x)$.

The word cut is not used uniformly in the literature. For example, $\mathrm{I}\Delta_0 + \Omega_1$-cut often refers to a $\mathrm{I}\Delta_0 + \Omega_1$-initial segment which is even provably closed under ω_1 (see e.g. [Vi 90a]). The reason that in many applications such confusion is not harmful is provided by lemma 2.6.6 and lemma 2.6.10.

Lemma 2.6.5 *Suppose that $T \supseteq \mathrm{I}\Delta_0$ and let I be a T-cut. Then there is a formula J such that*

1. $T \vdash \forall x (J(x) \rightarrow I(x))$;

2. *J is a T-cut;*

3. *$T \vdash \forall x \forall y(J(x) \wedge J(y) \rightarrow J(x+y))$, i.e. J is closed under $+$.*

Proof. Take

$$J(x) \;:\leftrightarrow\; I(x) \wedge \forall y(I(y) \rightarrow I(x+y)).$$

It is easy to see that $T \vdash \forall x(J(x) \rightarrow I(x))$ and that J is a T-cut.

For closure under $+$, reason in $\mathrm{I}\Delta_0$ and suppose that $x_1, x_2 \in J$ and that $y \in I$. Then by definition of J we have, first, $x_1 + x_2 \in I$. Also $y + x_1 \in I$, thus $y + (x_1 + x_2) = (y + x_1) + x_2 \in I$. We may conclude that $x_1 + x_2 \in J$. QED

Lemma 2.6.6 (Solovay's shortening lemma, [So 76b]) *Suppose that $T \supseteq \mathrm{I}\Delta_0$ and let I be a T-cut. Then there is a formula K such that*

1. *$T \vdash \forall x(K(x) \rightarrow I(x))$;*

2. *K is a T-initial segment;*

Proof. First construct J from I as in lemma 2.6.5. Next, define

$$K(x) \;:\leftrightarrow\; J(x) \wedge \forall y(J(y) \rightarrow J(x \cdot y)).$$

We leave it to the reader to prove that K is indeed the desired T-initial segment. QED

The following lemma 2.6.8 is used in almost all applications of cuts. Note that it is essential that we use the efficient numerals $\bar{x}$ which are based on the binary expansion of x. First we introduce a notational convention.

Notation 2.6.7 As in Pudlák's papers [Pu 86], we use the following notation:

$$T \overset{n}{\vdash} \varphi$$

to denote that there exists a proof of φ in T whose length (to which the length of proof lines contributes) is $\leq n$. Furthermore we use

$$T \overset{|n|}{\underset{*}{\vdash}} \varphi(n)$$

to denote that for some polynomial P we have for all n, $T \overset{P(|n|)}{\vdash} \varphi(n)$. Par abus de langage, we also use these abbreviations in formalized contexts whenever we think that their use will not confuse the reader.

Lemma 2.6.8 (Pudlák) *Suppose J is a T-initial segment. Then $T \overset{|n|}{\underset{*}{\vdash}} J(\bar{n})$. Also we have $\mathrm{I}\Delta_0 + \Omega_1 \vdash \forall x Prov_T(\ulcorner J(\bar{x})\urcorner)$.*

Proof. We give only a sketch, and leave the formal details to the reader. Essentially, in the proof of $J(\bar{x})$, we follow the $|x|$ steps it takes to build $\bar{x}$ from $\bar{0}$. At every step we instantiate either the proof of $\forall y(J(y) \rightarrow J(Sy))$ or the proof of $\forall y(J(y) \rightarrow J(SS0 \cdot y))$ with the appropriate efficient numeral. By using Modus Ponens a total of $|x|$ times, we finally derive $J(\bar{x})$. The length of the proof can evidently be bounded by a polynomial in $|x|$.

By inspection of the proof we see that it can be formalized to get $\mathrm{I}\Delta_0 + \Omega_1 \vdash \forall x Prov_T(\ulcorner J(\bar{x})\urcorner)$. Also it is useful to remark that in the proofs of $J(\bar{x})$, only formulas of a fixed complexity depending only on J are used. QED

Lemma 2.6.9 *Suppose that $T \supseteq \mathrm{I}\Delta_0$. For every k and every T-cut I, there exists an T-initial segment J such that $T \vdash \forall x(J(x) \to I(2^x_k))$. (We use functional notation for brevity, but we remind the reader that there are appropriate equivalents using the Δ_0-formulas that correspond to $2^x_k = y$.)*

Proof. We define $I_0, \dots, I_k$ and $J_0, \dots, J_k$ by recursion.

- $I_0(x) \;:\leftrightarrow I(x)$;
- for every $i \le k$, J_i is constructed from I_i by lemma 2.6.5;
- $I_{i+1}(x) \;:\leftrightarrow J_i(2^x)$.

We prove by induction on i that every I_i is a T-cut such that $T \vdash \forall x(I_i(x) \to I(2^x_i))$. For $i = 0$ this is clear. Suppose as induction hypothesis that it holds for i, and reason in T. First we show that $\forall x(I_{i+1}(x) \to I(2^x_{i+1}))$. Suppose $x \in I_{i+1}$, then by definition $2^x \in J_i$, so because $J_i \in I_i$ we have $2^{2^x}_i = 2^x_{i+1} \in I$. Next we show that I_{i+1} is a T-cut. Again, suppose that $x \in I_{i+1}$, thus $2^x \in J_i$. Since J_i is closed under $+$, we also have $2^x + 2^x = 2^{x+1} \in J_i$, thus $x + 1 \in I_{i+1}$.

To find the desired J, simply close J_k under $\cdot$ by lemma 2.6.6.

(Note that we do *not* have $\mathrm{I}\Delta_0 \vdash \forall x(J(x) \to J(2^x_k))$.) QED

We remind the reader that ω_1 is defined in definition 2.1.7.

Lemma 2.6.10 *Suppose that $T \supseteq \mathrm{I}\Delta_0$ and let I be a T-cut. Then there is a formula J such that*

1. *$T \vdash \forall x(K(x) \to I(x))$;*
2. *J is a T-initial segment;*
3. *$T \vdash \forall x(J(x) \to J(\omega_1(x)))$.*

Proof. First take J_2 as defined in the proof of lemma 2.6.9, and close it off under $\cdot$ by lemma 2.6.6 to get a T-initial segment K. Next define

$$J(x) \;:\leftrightarrow \exists y(K(y) \wedge x \le 2^{2^y}).$$

We leave it to the reader to show that J is a T-initial segment such that $T \vdash \forall x(J(x) \to K(x))$.

For closure under ω_1, we reason in T and we use the fact, provable by induction, that for $n > 1$, $\omega_1(2^{2^n}) \le 2^{2^{4\cdot n}}$. Now suppose that $x \in J$. Then for some $y \in K$ (where we may take $y > 1$ without loss of generality), $\omega_1(x) \le \omega_1(2^{2^y}) \le 2^{2^{4\cdot y}}$. But because K is closed under $+$, we have $4 \cdot y \in K$, so by definition $\omega_1(x) \in J$. QED

Lemma 2.6.11 *Let $\varphi \in \Delta_0(exp)$. Suppose $\mathrm{I}\Delta_0 + EXP \vdash \forall x\varphi(x)$. Then there is a k such that $\mathrm{I}\Delta_0 \vdash \forall x(2^x_k \downarrow\to \varphi(x))$.*

Proof. The proof is reminiscent of Parikh's Theorem (see theorem 2.3.15). Suppose, in order to derive a contradiction, that there is no k such that $\mathrm{I}\Delta_0 \vdash \forall x(2^x_k \downarrow\rightarrow \varphi(x))$. Then, for a fresh constant c and for all k, $\mathrm{I}\Delta_0 \nvdash 2^c_k \downarrow\rightarrow \varphi(c)$. By the compactness theorem, there is a model $\mathcal{M} \models \mathrm{I}\Delta_0 + \{2^c_k \downarrow \mid k \in \omega\} + \neg\varphi(c)$. Next, we define $\mathcal{M}^* := \{b \in \mathcal{M} \mid \exists k(\mathcal{M} \models b < 2^c_k)\}$. Now $\mathcal{M}^* \models \mathrm{I}\Delta_0 + EXP$, so $\mathcal{M}^* \models \forall x\varphi(x)$, in particular $\mathcal{M}^* \models \varphi(c)$. But $\mathcal{M}^* \subseteq_e \mathcal{M}$, so $\mathcal{M} \models \varphi(c)$. There we have our contradiction. QED

Wilkie and Paris proved that the Π_1-consequences of $\mathrm{I}\Delta_0 + EXP$ can be characterized using $\mathrm{I}\Delta_0$-initial segments.

Theorem 2.6.12 (Wilkie and Paris [WP 87], Corollary 8.8) *Let $\varphi \in \Delta_0(exp)$. Then the following two statements are equivalent:*

1. $\mathrm{I}\Delta_0 + EXP \vdash \forall x\varphi(x)$
2. *There is an $\mathrm{I}\Delta_0$-initial segment J such that* $\mathrm{I}\Delta_0 \vdash \forall x(J(x) \rightarrow \varphi(x))$.

Proof.

$1 \rightarrow 2$ Suppose $\mathrm{I}\Delta_0 + EXP \vdash \forall x\varphi(x)$. By lemma 2.6.11, there is a k such that $\mathrm{I}\Delta_0 \vdash \forall x(2^x_k \downarrow\rightarrow \varphi(x))$. By lemma 2.6.9, there is an $\mathrm{I}\Delta_0$-initial segment such that $\mathrm{I}\Delta_0 \vdash \forall x(J(x) \rightarrow 2^x_k \downarrow)$. Combining these two facts, we derive the desired conclusion $\mathrm{I}\Delta_0 \vdash \forall x(J(x) \rightarrow \varphi(x))$.

$2 \rightarrow 1$ For the other direction, we need lemma 2.8.10. Thus we refer the reader to Corollary 2.8.11.

QED

Lemma 2.6.13 *Let $\varphi \in \Delta_0(exp)$. If $\mathrm{I}\Delta_0 + EXP \vdash \forall x\varphi(x)$, then $\mathrm{I}\Delta_0 \vdash^{|n|}_{*} \varphi(\overline{n})$.*

Proof. Suppose $\mathrm{I}\Delta_0 + EXP \vdash \forall x\varphi(x)$. By theorem 2.6.12, there is an $\mathrm{I}\Delta_0$-initial segment such that $\mathrm{I}\Delta_0 \vdash \forall x(J(x) \rightarrow \varphi(x))$. Moreover by lemma 2.6.8, we have $\mathrm{I}\Delta_0 \vdash^{|n|}_{*} J(\overline{n})$. Thus $\mathrm{I}\Delta_0 \vdash^{|n|}_{*} \varphi(\overline{n})$. QED

We give the most famous examples of interpretations provided by initial segments. One of them has an almost trivial proof, while the second one, on account of the weakness of Robinson's Arithmetic Q, needs a complicated argument.

Lemma 2.6.14 $\mathrm{I}\Delta_0 \rhd \mathrm{I}\Delta_0 + \Omega_1$ *on an $\mathrm{I}\Delta_0$-initial segment.*

Proof. This is a particularly easy application of cuts. By lemma 2.6.10, we simply construct an $\mathrm{I}\Delta_0$-initial segment closed under ω_1. Remember that $\mathrm{I}\Delta_0$ is a Π_1 theory, so $\mathrm{I}\Delta_0 \vdash (\mathrm{I}\Delta_0 + \Omega_1)^J$ (see remark 2.6.3). QED

In order to prove that $Q \rhd \mathrm{I}\Delta_0$ on a Q-initial segment, we need a Δ_0 "truth definition" (with one extra parameter) for Δ_0-formulas. Such a definition was provided by Paris and Dimitracopoulos.

Lemma 2.6.15 (Paris and Dimitracopoulos) *There is a Δ_0-formula $\Gamma(x, z, u)$ ("x is satisfied by the sequence of numbers z, with bound u) and a constant k such that:*

$$\begin{aligned} \mathrm{I}\Delta_0 \vdash \;\; & |u| \geq (\max(z) + 2)^{k^{|x|}} \rightarrow \\ & \text{“}\,\Gamma(x, z, u) \textit{ satisfies Tarski's conditions for } x \in \Delta_0\text{”} \end{aligned}$$

Proof. See [PD 82] or [HP 93, Theorem V.5.4] QED

Theorem 2.6.16 (Wilkie) *$Q \rhd \mathrm{I}\Delta_0$ on a Q-initial segment.*

Proof. See [HP 93, Theorem V.5.7]. We give only the skeleton of the proof. There are three steps:

1. Let Q' be Q with three additional axioms: associativity of $+$ and $\cdot$, and left-distributivity (i.e. $x \cdot (y + z) = x \cdot y + x \cdot z$). Nelson has shown that $Q \rhd Q'$ on a Q-initial segment [Ne 86].

2. Every finite fragment of $\mathrm{I}\Delta_0$ can be interpreted in Q' via a Q'-initial segment.

3. We take a finite fragment T of $\mathrm{I}\Delta_0$ which is so strong that:

   - T proves the properties of the exponentiation relation;
   - T proves Tarski's conditions for the satisfaction formula Γ for Δ_0-formulas;
   - T proves the least number principle for $\Gamma(x, (y, p), u)$ with y as induction parameter.

   Now to be able to prove the least number principle for all Δ_0-formulas on an initial segment, it is sufficient to construct by lemma 2.6.9 a T-initial segment J which is so short that $T \vdash \forall x(J(x) \rightarrow 2^{2^x}\downarrow)$, so that we can replace the scheme by a single axiom.

Finally, we combine all three interpretations. QED

2.7 Cuts may help to characterize interpretability

In this dissertation many theories that we consider are not extensions of PA. On the other hand, they are almost all sequential.

For such sequential theories U and V that extend Robinson's arithmetic Q, we can still prove analogs of theorem 2.4.6 and theorem 2.4.9 using definable initial segments.

First we need a definition.

Definition 2.7.1 For V a theory the axioms of which are defined by the Σ_1^b-formula $\alpha_V(y)$, let $V[x]$ be the theory axiomatized by the formula $\alpha_V(y) \wedge y \leq x$. Now we can define local interpretability as follows:

$$U \rhd_{loc} V :\Longleftrightarrow \forall x \exists K[U \rhd V[x] \textit{ by interpretation } K].$$

Of course, if V is finitely axiomatized, $U \rhd V$ and $U \rhd_{loc} V$ are synonimic.

For the sake of legibility we will use quasi-modal abbreviations $\Box_U$ and $\Diamond_U$ for $Prov_{\alpha_U}$ and Con_{α_U}, where α_U is the Σ_1^b-formula defining the axioms of U.

The following lemma has a proof analogous to the proof of theorem 2.4.10. Let Q^+ be Q plus the axioms expressing that $\leq$ is a linear order.

Lemma 2.7.2 (Pudlák's Theorem on cuts, see [Pu 85]) *Suppose U is sequential, V extends Q^+, and $U \rhd V$ by interpretation K. Then for every V-initial segment I there is a U-initial segment J such that there is a definable initial embedding from J into I^K.*

Proof. See [Vi 90a] and [Vi 93]. There it is also stated that the theorem is verifiable in $I\Delta_0 + \Omega_1$. QED

In the sequel of this section we use some abbreviations to improve ease of reading. $\exists J \in U$-cuts stands for the formalization of "there is a U-initial segment J such that". By $\Diamond^J_{x,U}\top$ we abbreviate the formalization of "J does not contain a proof of $\bot$ using only U-axioms with Gödel number $< x$".

Lemma 2.7.3 *Let U be any sequential theory. Then we have*

$$I\Delta_0 + EXP \vdash \forall x \exists J \in U\text{-}cuts \Box_U \Diamond^J_{x,U}\top.$$

Proof. See [Vi 93]. Partial truth predicates for formulas of limited complexity (see section 3.3 and [Pu 86, Pu 87]) play a crucial rôle in the argument. QED

Theorem 2.7.4 (Visser, [Vi 93])
Let U and V be sequential theories extending Q^+. Then we have

$$I\Delta_0 + EXP \vdash U \rhd_{loc} V \leftrightarrow \forall x \exists J \in U\text{-}cuts\ \Box_U \Diamond^J_{x,V}\top.$$

Proof.

→ Work inside $I\Delta_0 + EXP$ and suppose $U \rhd_{loc} V$. By lemma 2.7.3, we know that $\forall x \exists I \in V$-cuts $\Box_V \Diamond^I_{x,V}\top$. For every x, we can apply theorem 2.7.2 to find a U-initial segment J such that $\Box_U \Diamond^J_{x,V}\top$. Thus we derive $\forall x \exists J \in U$-cuts $\Box_U \Diamond^J_{x,V}\top$, as desired.

← Work inside $I\Delta_0 + EXP$ and suppose $\forall x \exists J \in U$-cuts $\Box_U \Diamond^J_{x,V}\top$. For every x we can now carry out a Henkin-construction giving us the desired local interpretation.

QED

Theorem 2.7.5 (Visser, [Vi 93]) *Let U and V be sequential theories extending Q^+. Then we have*

$$I\Delta_0 + EXP \vdash U \rhd_{loc} V \leftrightarrow \forall P \in \Pi_1(\exists I \in V\text{-}cuts\ \Box_V P^I \rightarrow \exists J \in U\text{-}cuts\ \Box_U P^J).$$

Proof.

→ This follows by theorem 2.7.2 formalized in $I\Delta_0 + EXP$.

$\leftarrow$ This follows immediately from the $\leftarrow$-direction of theorem 2.7.4, when we note that $\Diamond_{x,V}\top$ is a Π_1-formula.

QED

In the next theorem, the quantifier $\exists K$ stands for "there is an interpretation K".

Theorem 2.7.6 *Suppose U and V are sequential theories extending Q^+. We have the following scheme of relationships (provable in $\mathrm{I}\Delta_0 + EXP$) between the various definitions of interpretability. Both arrows pointing down are strict.*

$$\begin{tikzcd}
\exists J \in U\text{-}cuts\ \forall x \Box_U \Diamond^J_{x,V}\top \arrow[r, Leftrightarrow] \arrow[d, Rightarrow] & \exists K \forall x \Box_U \Diamond^K_{x,V}\top & \\
U \rhd V \arrow[d, Rightarrow] & & \\
U \rhd_{loc} V \arrow[r, Leftrightarrow] & \forall x \exists J \in U\text{-}cuts\ \Box_U \Diamond^J_{x,V}\top \arrow[r, Leftrightarrow] & \forall x \exists K \Box_U \Diamond^K_{x,V}\top
\end{tikzcd}$$

Proof.

- The $\Rightarrow$-direction of $\exists J \in U$-cuts $\forall x \Box_U \Diamond^J_{x,V}\top \iff \exists K \forall x \Box_U \Diamond^K_{x,V}\top$ is clear, because every U-initial segment provides an interpretation; the $\Leftarrow$-direction follows from theorem 2.7.2 as formalized in $\mathrm{I}\Delta_0 + EXP$.

- To prove $\exists J \in U$-cuts $\forall x \Box_U \Diamond^J_{x,V}\top \Rightarrow U \rhd V$, one uses a formalized Henkin construction for a Feferman proof predicate; the argument is analogous to the proof of theorem 6.5.11.

- By definition we have $U \rhd V \Rightarrow U \rhd_{loc} V$.

- $U \rhd_{loc} V \iff \forall x \exists J \in U$-cuts $\Box_U \Diamond^J_{x,V}\top$ is just theorem 2.7.4.

- $\forall x \exists J \in U$-cuts $\Box_U \Diamond^J_{x,V}\top \iff \forall x \exists K \Box_U \Diamond^K_{x,V}\top$ is proved again by the fact that initial segments provide interpretations and by theorem 2.7.2.

- The arrow $\exists J \in U$-cuts $\forall x \Box_U \Diamond^J_{x,V}\top \Rightarrow U \rhd V$ is strict. For take $U = V = \mathrm{I}\Delta_0 + \Omega_1$ and let J be any $\mathrm{I}\Delta_0 + \Omega_1$-initial segment. Then we have of course $\mathrm{I}\Delta_0 + \Omega_1 \rhd \mathrm{I}\Delta_0 + \Omega_1$, but for big enough k (with respect to $\mathrm{I}\Delta_0 + \Omega_1$ and J) we have $\Box_{\mathrm{I}\Delta_0+\Omega_1} \Diamond^J_{k,\mathrm{I}\Delta_0+\Omega_1}\top \rightarrow \Box_{\mathrm{I}\Delta_0+\Omega_1}\bot$, by Visser's adaptation of Löb's Theorem (see [Vi 93, Corollary 4.4]).

- The arrow $U \rhd V \Rightarrow U \rhd_{loc} V$ is strict. For take $U = \mathrm{I}\Delta_0 + \Omega_1$ and $V = \mathrm{I}\Delta_0 + \Omega_1 + \{\Diamond_{n,U}\top \mid n \in \omega\}$. Then $U \rhd_{loc} V$, but by [Vi 93, Corollary 4.5], U does not interpret V.

QED

2.8 Between $I\Delta_0$ and $I\Delta_0 + EXP$

The results of this section are not needed in the subsequent chapters of the dissertation. Instead they appear here to give the flavor of the model-theoretic methods used by Wilkie and Paris in [WP 87]. Also we hope that the reader will gain some understanding of the difference in strength between the theories $I\Delta_0$ and $I\Delta_0 + EXP$.

For example, whereas $I\Delta_0$ interprets $I\Delta_0 + \Omega_1$, it does not interpret $I\Delta_0 + EXP$. Thus $I\Delta_0 + EXP$ is much stronger that $I\Delta_0$ and $I\Delta_0 + \Omega_1$. Another advertisement for the strength of $I\Delta_0 + EXP$ is the fact that it proves tableau- consistency of $I\Delta_0 + \Omega_1$.

On the other hand, perhaps surprisingly, $I\Delta_0 + EXP$ does not even prove consistency of the extremely weak theory Q. This is caused essentially by the failure of $I\Delta_0 + EXP$ to prove a formalization of Gentzen's cut-elimination theorem.

Another interesting question is the following: when is EXP necessary to prove some true Π_1-statement? A partial answer is given in subsection 2.8.2: in $I\Delta_0 + EXP$-proofs of sufficiently simple Π_1-sentences, namely those of $\forall\Pi_1^b$-form, one can get by without EXP and replace it by restricted consistency statements.

2.8.1 $I\Delta_0 + EXP$ proves restricted consistency statements

Definition 2.8.1 A k-formula is a formula with $\leq k$ logical connectives. (Note that a k-formula may be arbitrarily long due to the presence of non-standard terms.)

A k-proof is a proof in which only k-formulas appear. $Prov_T(\varphi, k)$ means that there is a k-proof of φ from T. Similarly, $Con(T, k)$ means that there is no k-proof of a contradiction from T.

Definition 2.8.2 Let T be a set of sentences. We say that a sequence of sets of sets of formulas $\Gamma_0, \dots, \Gamma_s$ is a tableau proof of an inconsistency from T if the following conditions hold:

- For each $X \in \Gamma_s$, there is an atomic θ such that $\theta \in X$ and $\neg\theta \in X$.
- $X \in \Gamma_0$ implies $X \subseteq T \cup$ the set of logical equality axioms.
- For each $X \in \Gamma_i$ with $i < s$ one of the following holds:
  1. $X \in \Gamma_{i+1}$,
  2. $X \cup \{\theta\} \in \Gamma_{i+1}$ for some $\neg\neg\theta \in X$,
  3. $X \cup \{\neg\theta_1\}, X \cup \{\theta_2\} \in \Gamma_{i+1}$ for some $(\theta_1 \rightarrow \theta_2) \in X$,
  4. $X \cup \{\theta_1, \neg\theta_2\} \in \Gamma_{i+1}$ for some $\neg(\theta_1 \rightarrow \theta_2) \in X$,
  5. $X \cup \{\theta(t)\} \in \Gamma_{i+1}$ for some $\forall x\theta(x) \in X$ and some term t which is free for x in $\theta(x)$,
  6. $X \cup \{\neg\theta(y)\} \in \Gamma_{i+1}$ for some $\neg\forall x\theta(x) \in X$ and some variable y which does not occur in any formula in X.
- For each $Y \in \Gamma_{i+1}$ with $i < s$ there is an $X \in \Gamma_i$ such that Y is obtained from X by one of the rules 1-6.

We write $Tabcon(T)$ if there is no tableau proof of an inconsistency from T.

Definition 2.8.3 Let L^* be the language of arithmetic where successor, addition and multiplication are relation symbols. The only terms of L^* are variables and $\bar{0}$. $I\Delta_0^*$ is the reformulation of $I\Delta_0$ in L^*, with extra axioms expressing the totality of successor, addition and multiplication. Similarly any formula φ has an obvious L^*-translation φ^* with the same unbounded quantifier complexity.

Lemma 2.8.4 *Suppose $\mathcal{M} \models I\Delta_0 + EXP$. There is a $\Delta_0(exp)$ formula $Tr(a, y)$ such that for every sentence $\varphi \in L^*$ and every $a \in \mathcal{M}$,*

$$a \models \varphi \Longleftrightarrow \mathcal{M} \models Tr(a, \ulcorner\varphi\urcorner).$$

(Here we have identified a with the substructure of $\mathcal{M}$ that has universe $\{x \in \mathcal{M} \mid \mathcal{M} \models x \leq a\}$. This presents no problem when we work in a relational language.)

Proof. See [PD 82], and cf. lemma 2.6.15. QED

Lemma 2.8.5

$$I\Delta_0 + EXP \vdash \forall a \forall i \exists b \, [b = \omega_1^{(i)}(a)].$$

(Here $\omega_1^{(i)}(a)$ is defined informally as ω_1 applied i times to a.)

Proof. Remember that there is a Δ_0-formula $\varphi(a, i, b)$ which expresses $b = \omega_1^{(i)}(a)$. The lemma then follows easily by $\Delta_0(exp)$ induction, using an appropriate bound on $\omega_1^{(i)}(a)$. QED

Theorem 2.8.6 (Wilkie and Paris [WP 87], Lemma 8.10)

1. $I\Delta_0 + EXP \vdash Tabcon(I\Delta_0^* + \Omega_1^*)$.

2. *If $\sigma \in \Sigma_2$, then*

$$I\Delta_0 + EXP + \sigma \vdash Tabcon(I\Delta_0^* + \Omega_1^* + \sigma^*).$$

Proof. We prove the second, more general, part of the theorem. Suppose that $\sigma = \exists x \forall y \delta(x, y)$, where δ is a Δ_0-formula, and reason in $\mathcal{M} \models I\Delta_0 + EXP + \sigma$. Let a be such that $\forall y \delta(a, y)$. Suppose that $\Gamma_0, \ldots, \Gamma_s$ is a tableau proof in $\mathcal{M}$ of a contradiction from $I\Delta_0^* + \Omega_1^* + \sigma^*$. Take $b = \omega_1^{(s+1)}(a + 2)$, as is justified by lemma 2.8.5.

Let $\Psi_i := \Gamma_i \cap 2^{\Pi_1 \cup \Sigma_1}$.

Define $P(i) \leftrightarrow \exists f : Var(\Gamma_i) \longmapsto \{u \mid u < \omega_1^{(i+1)}(a + 2)\}$ such that

$$b \models \bigwedge_{X \in \Psi_i} \bigvee_{\varphi \in X} \varphi^f.$$

Using the appropriate truth definition, $P(i)$ can be expressed by a $\Delta_0(exp)$ formula. We will use $\Delta_0(exp)$ induction (which is available in $I\Delta_0 + EXP$) to prove that $\forall i \leq s P(i)$. This contradicts the fact that Γ_s contains both θ and $\neg\theta$ for some atomic θ.

- The base step relies on the observation that if β is a Π_1 axiom of $\mathrm{I}\Delta_0$ then $b \models \beta$. Suppose for example that β is the induction axiom

  $$\forall x, z(\psi(x,0) \wedge \forall y \leq z(\psi(x,y) \rightarrow \psi(x,Sy)) \rightarrow \forall y \leq z\ \psi(x,y)).$$

  Let $\beta^*(x_0)$ be obtained from β by replacing ψ with '$x_0 \models \psi(x,y)$' (using the truth definition from Lemma 2.8.4). Then $\mathcal{M} \models \forall x_0 \beta^*(x_0)$, so $\mathcal{M} \models \beta^*(b)$, hence $b \models \beta$.

- The induction step from $P(i)$ to $P(i+1)$ hinges on the fact that the only time an unbounded quantifier $\exists$ (i.e. $\neg\forall\neg$) is to be eliminated in the tableau proof is on a subformula beginning with $\exists$ of $\mathrm{I}\Delta_0^* + \Omega_1^* + \sigma^*$; but for the formula $\exists x \forall y \delta(x,y)$ we already know that $b \models \forall y \delta(a,y)$; also the formulas $\exists y(y = x+1), \exists y(y = x_1 + x_2), \exists y(y = x_1 \cdot x_2)$, and $\exists y(y = \omega_1(x))$ present no problem, because by induction hypothesis their free variables can be instantiated by parameters $< \omega_1^{(i+1)}(a+2)$.

QED

Domenico Zambella found the following generalization of Theorem 8.2 of [WP 87].

Theorem 2.8.7 *Suppose τ is a sentence, $i \geq 1$ and $\mathcal{M}$ is a countable model satisfying*

1. $\mathcal{M} \models \mathrm{I}\Delta_0 + \Omega_1$, *and*

2. *for all k and for all Π_i^b formulas φ with parameters $a_1, \ldots, a_n$ from $\mathcal{M}$,*

   $$\mathcal{M} \models \forall a_1, \ldots, a_n(Prov_{\mathrm{I}\Delta_0+\tau}(\ulcorner \varphi(\overline{a_1}, \ldots, \overline{a_n})\urcorner, k) \rightarrow \varphi(a_1, \ldots, a_n)).$$

Then there is a model $\mathcal{M}^* \models \mathrm{I}\Delta_0 + \tau$ such that $\mathcal{M} \prec_{\Sigma_i^b} \mathcal{M}^*$.

Proof. It is sufficient to find a model $\mathcal{M}^*$ such that

$$\mathcal{M}^* \models Diag_{\Sigma_i^b}(\mathcal{M}) + \mathrm{I}\Delta_0 + \tau,$$

where in $Diag_{\Sigma_i^b}$ new constants c_a are used for elements $a \in \mathcal{M}$.

So, in order to derive a contradiction, suppose that $Diag_{\Sigma_i^b}(\mathcal{M}) + \mathrm{I}\Delta_0 + \tau$ is inconsistent. Then there is a Σ_i^b-formula φ and there are $a_1, \ldots, a_n \in \mathcal{M}$ such that on the one hand $\mathcal{M} \models \varphi(a_1, \ldots, a_n)$, but on the other hand $\mathrm{I}\Delta_0 + \tau \vdash \neg\varphi(c_{a_1}, \ldots, c_{a_n})$ by a proof in which all formulas have complexity k, for some standard k. This contradicts assumption (2). QED

Corollary 2.8.8 (Wilkie and Paris [WP 87], Theorem 8.2)
Suppose τ is a sentence and $\mathcal{M}$ is a countable model satisfying

1. $\mathcal{M} \models \mathrm{I}\Delta_0 + \Omega_1$, *and*

2. *for all k, $\mathcal{M} \models Con(\mathrm{I}\Delta_0 + \tau, k)$.*

Then there is a model $\mathcal{M}^* \models \mathrm{I}\Delta_0 + \tau$ such that $\mathcal{M} \prec_{\Sigma_1^b} \mathcal{M}^*$.

Proof. In order to be able to apply theorem 2.8.7, it is sufficient to prove that assumption 2 of theorem 2.8.7 is implied by the assumption that for all k, $\mathcal{M} \models Con(\mathrm{I}\Delta_0 + \tau, k)$.

So suppose in order to derive a contradiction that for some k_1 and for some Π_i^b formula φ with parameters $a_1, \ldots, a_n$ from $\mathcal{M}$,

$$\mathcal{M} \models Prov_{\mathrm{I}\Delta_0+\tau}(\ulcorner \varphi(\overline{a_1}, \ldots, \overline{a_n})\urcorner, k_1), \tag{2.3}$$

but

$$\mathcal{M} \models \neg\varphi(a_1, \ldots, a_n). \tag{2.4}$$

Clearly $\neg\varphi(a_1, \ldots, a_n)$ is a Σ_1^b-formula. By inspection of the proof of Σ_1^b-completeness in $\mathrm{I}\Delta_0 + \Omega_1$ (cf. theorem 2.3.24), we conclude from (2.4) that there is a k_2 such that $\mathcal{M} \models Prov_{\mathrm{I}\Delta_0+\tau}(\ulcorner \neg\varphi(\overline{a_1}, \ldots, \overline{a_n})\urcorner, k_2)$. Together with (2.3), this implies the existence of a k_3 such that $\mathcal{M} \models \neg Con(\mathrm{I}\mathit{\Delta}_0 + \tau, k_3)$, contradicting the assumption that for all k, $\mathcal{M} \models Con(\mathrm{I}\mathit{\Delta}_0 + \tau, k)$. QED

Remark 2.8.9 In assumption (2) of theorem 2.8.7, we may replace the formula $Prov_{\mathrm{I}\Delta_0+\tau}(\ulcorner \varphi(\overline{a_1}, \ldots, \overline{a_n})\urcorner, k)$ by $Tabprov_{\mathrm{I}\Delta_0+\tau}(\ulcorner \varphi(\overline{a_1}, \ldots, \overline{a_n})\urcorner)$. Similarly we may replace $Con(\mathrm{I}\mathit{\Delta}_0, k)$ in assumption (2) of corollary 2.8.8 by $Tabcon(\mathrm{I}\Delta_0)$. (However, for the use of corollary 2.8.8 and theorem 2.8.7 in theorem 2.8.13 and remark 2.8.14, we need the original formulation.)

Also, the assumption that $\mathcal{M} \models \mathrm{I}\Delta_0 + \Omega_1$ is not needed for theorem 2.8.7, although it is essential for its corollary 2.8.8.

Lemma 2.8.10 *If σ is a Σ_2-sentence, then*

$$\forall k\ (\mathrm{I}\Delta_0 + EXP + \sigma \vdash Con(\mathrm{I}\mathit{\Delta}_0 + \mathit{\Omega}_1 + \sigma, k)).$$

Proof. By theorem 2.8.6, we know that

$$\mathrm{I}\Delta_0 + EXP + \sigma \vdash Tabcon(\mathrm{I}\Delta_0 + \Omega_1 + \sigma).$$

We remind the reader that cut-free proofs can easily be converted into tableau proofs by an algorithm that increases the length of the proofs only polynomially.

Now take some $k \in \omega$. From the formalization of the cut-elimination theorem given in the appendix to [Vi 92], it follows that

$$\mathrm{I}\Delta_0 + EXP \vdash Tabcon(\mathrm{I}\Delta_0 + \Omega_1 + \sigma) \rightarrow Con(\mathrm{I}\mathit{\Delta}_0 + \mathit{\Omega}_1 + \sigma, k).$$

(Indeed every k-proof with code p can be converted into a cut-free proof, and thus also into a tableau proof, whose Gödel number is bounded by 2_k^p; see definition 2.1.8 for an inductive definition of 2_k^p.)

We may conclude $\mathrm{I}\Delta_0 + EXP + \sigma \vdash Con(\mathrm{I}\mathit{\Delta}_0 + \mathit{\Omega}_1 + \sigma, k)$. QED

The above lemma is a strengthening of [WP 87, Proposition 8.5]. There it was proved that if σ is a Π_1-sentence, then

$$\forall k\ (\mathrm{I}\Delta_0 + EXP + \sigma \vdash Con(\mathrm{I}\mathit{\Delta}_0 + \sigma, k)).$$

Remark 2.8.11 At this point we can provide the postponed proof of one direction of theorem 2.6.12. Let $\varphi \in \Delta_0(exp)$, and suppose that there is an $\mathrm{I}\Delta_0$-initial segment J such that

$$\mathrm{I}\Delta_0 \vdash \forall x(J(x) \to \varphi(x)). \tag{2.5}$$

We want to prove $\mathrm{I}\Delta_0 + EXP \vdash \forall x \varphi(x)$.

When we inspect the usual proof of $\Delta_0(exp)$-completeness in $\mathrm{I}\Delta_0 + EXP$, we note that there is a k_1 such that:

$$\mathrm{I}\Delta_0 + EXP \vdash \forall x(\neg\varphi(x) \to Prov_{\mathrm{I}\Delta_0}(\ulcorner \neg\varphi(\bar{x})\urcorner, k_1)). \tag{2.6}$$

Also, because J is an $\mathrm{I}\Delta_0$-cut, there is a k_2 such that $\mathrm{I}\Delta_0 + EXP \vdash \forall x Prov_{\mathrm{I}\Delta_0}(\ulcorner J(\bar{x})\urcorner, k_2)$; thus by (2.5), there is a k_3 such that

$$\mathrm{I}\Delta_0 + EXP \vdash \forall x Prov_{\mathrm{I}\Delta_0}(\ulcorner \varphi(\bar{x})\urcorner, k_3). \tag{2.7}$$

Next, we combine (2.6) and (2.7) to find a k such that $\mathrm{I}\Delta_0 + EXP \vdash \forall x(\neg\varphi(x) \to \neg Con(\mathrm{I}\Delta_0, k)$. By lemma 2.8.10 we finally conclude that indeed $\mathrm{I}\Delta_0 + EXP \vdash \forall x \varphi(x)$. (Note that we only needed the fact that J is an $\mathrm{I}\Delta_0$-cut, not that it is an $\mathrm{I}\Delta_0$-initial segment.)

2.8.2 Conservativity

Wilkie and Paris characterize the $\forall\Pi_1^b$-consequences of $\mathrm{I}\Delta_0 + EXP$ by providing a basis over $\mathrm{I}\Delta_0 + \Omega_1$: if one adds the restricted consistency statements to $\mathrm{I}\Delta_0 + \Omega_1$, one can already derive all $\forall\Pi_i^b$-consequences of $\mathrm{I}\Delta_0 + EXP$. We first need a definition.

Definition 2.8.12 A U_i-formula is a formula of the form $\forall x \varphi$ where φ is a Π_i^b formula. Note that all consistency statements for Σ_1^b-axiomatized theories can be written in U_1 form.

Theorem 2.8.13 (Wilkie and Paris [WP 87], Theorem 8.6)
Let τ be a Π_1-sentence. Then the following two theories have the same U_1 consequences:

- $T_1 := \mathrm{I}\Delta_0 + EXP + \tau;$
- $T_2 := \mathrm{I}\Delta_0 + \Omega_1 + \{Con(\mathrm{I}\Delta_0 + \tau, k) | k \in \omega\}.$

Proof. $T_1 \vdash T_2$ follows from lemma 2.8.10. For the U_1-conservativity of T_1 over T_2, suppose that

$$\mathcal{M} \models \mathrm{I}\Delta_0 + \Omega_1 + \{Con(\mathrm{I}\Delta_0 + \tau, k) | k \in \omega\} + \neg\forall x \varphi(x),$$

where $\neg\varphi(x) \in \Sigma_1^b$. We want to find $\mathcal{M}'$ with

$$\mathcal{M}' \models \mathrm{I}\Delta_0 + EXP + \tau + \neg\forall x \varphi(x).$$

First we construct, by corollary 2.8.8, a model $\mathcal{M}^* \succ_{\Sigma_1^b} \mathcal{M}$ such that

$$\mathcal{M}^* \models \mathrm{I}\Delta_0 + \tau + \neg\forall x \varphi(x).$$

Then by a trick reminiscent of the proof of Parikh's theorem, we let

$$\mathcal{M}' := \{a \in \mathcal{M}^* \mid \exists k \in \omega \exists b \in \mathcal{M}\ \mathcal{M}^* \models a < 2_k^b\}.$$

Note that 2_k^b is defined in $\mathcal{M}^*$ for all $k \in \omega$ and $b \in \mathcal{M}$. This depends on the following fact which can be proved using Solovay's cuts, more precisely by inspection of the proof of lemma 2.6.8 and by lemma 2.6.9:

For any standard k there is a standard m such that

$$\mathrm{I}\Delta_0 + \Omega_1 \vdash \forall b \Box_{\mathrm{I}\Delta_0, m} (2^b_k \downarrow).$$

Now $\mathcal{M} \subseteq \mathcal{M}'$ and $\mathcal{M}' \models \mathrm{I}\Delta_0 + EXP$. Since $\mathcal{M}'$ is an initial segment of $\mathcal{M}^*$, τ is preserved downwards hence $\mathcal{M}' \models \tau$.

Moreover there is a $d \in \mathcal{M}$ such that $\mathcal{M} \models \neg\varphi(d)$, so by corollary 2.8.8 part 2, since $\mathcal{M} \prec_{\Sigma^b_1} \mathcal{M}^*$, $\mathcal{M}^* \models \neg\varphi(d)$; and since $\mathcal{M}$ is an initial segment of $\mathcal{M}^*$, $\mathcal{M}' \models \neg\varphi(d)$, i.e. $\mathcal{M}' \models \neg\forall x \varphi(x)$. QED

Remark 2.8.14 If we use the full strength of theorem 2.8.7 instead of its corollary 2.8.8, we find that if τ is a Π_1-sentence, the following two theories have the same U_i consequences:

- $T_1 := \mathrm{I}\Delta_0 + EXP + \tau$ and
- $T_2 := \mathrm{I}\Delta_0 + \Omega_1 + \{Prov_{\mathrm{I}\Delta_0+\tau}(\ulcorner \varphi(\overline{x_1}, \ldots, \overline{x_n})\urcorner, k) \rightarrow \varphi(x_1, \ldots, x_n) \mid k \in \omega,$ $\varphi \in \Pi^b_i$ *with variables among* $x_1, \ldots, x_n\}$.

(In fact, by lemma 2.8.10, $T_1 \vdash T_2$.)

2.8.3 Non-conservativity and incompleteness

Theorem 2.8.15 (Wilkie and Paris [WP 87], Theorem 8.11)

1. *$\mathrm{I}\Delta_0 + EXP$ is not U_1-conservative over $\mathrm{I}\Delta_0 + \Omega_1$;*
2. *If σ is a Σ_2 sentence consistent with $\mathrm{I}\Delta_0 + EXP$, then $\mathrm{I}\Delta_0 + EXP + \sigma$ is not U_1-conservative over $\mathrm{I}\Delta_0 + \Omega_1 + \sigma$.*

Proof. We prove the second statement. Define $T := \mathrm{I}\Delta_0 + \Omega_1 + \sigma$, and construct ψ by diagonalization such that

$$\mathrm{I}\Delta_0 + \Omega_1 \vdash \psi \leftrightarrow \neg\, Tabcon(T + \psi).$$

It is easy to see that $\neg\psi$ is U_1. Now suppose that $T \vdash \neg\psi$. Then $T \vdash \neg\, Tabcon(T + \psi)$, so by definition $T \vdash \psi$, contradicting the consistency of T. Thus

$$\mathrm{I}\Delta_0 + \Omega_1 + \sigma \nvdash \neg\psi.$$

On the other hand, $\sigma \wedge \psi$ is Σ_2, so by lemma 2.8.6,

$$\mathrm{I}\Delta_0 + EXP + \sigma + \psi \vdash Tabcon(\mathrm{I}\Delta_0 + \Omega_1 + \sigma + \psi).$$

Therefore by definition of ψ,

$$\mathrm{I}\Delta_0 + EXP + \sigma \vdash \neg\psi.$$

QED

Corollary 2.8.16 (Wilkie and Paris [WP 87], Corollary 8.13)

1. $\forall k \exists n\, \mathrm{I}\Delta_0 + \Omega_1 + Con(\mathrm{I}\Delta_0, k) \nvdash Con(\mathrm{I}\Delta_0, n)$;

2. *Suppose π is a Π_1 sentence consistent with* $\mathrm{I}\Delta_0 + EXP$. *Then*

$$\forall k \exists n\, \mathrm{I}\Delta_0 + \Omega_1 + Con(\mathrm{I}\mathit{\Delta_0} + \pi, k) \nvdash Con(\mathrm{I}\mathit{\Delta_0} + \pi, n).$$

Proof. By theorem 2.8.15, there is a U_1 sentence ψ such that

$$\mathrm{I}\Delta_0 + EXP + \pi + Con(\mathrm{I}\mathit{\Delta_0} + \pi, k) \vdash \psi,$$

but

$$\mathrm{I}\Delta_0 + \Omega_1 + \pi + Con(\mathrm{I}\mathit{\Delta_0} + \pi, k) \nvdash \psi.$$

On the other hand $\mathrm{I}\Delta_0 + EXP + \pi \vdash Con(\mathrm{I}\mathit{\Delta_0} + \pi, k)$, so $\mathrm{I}\Delta_0 + EXP + \pi \vdash \psi$.

By theorem 2.8.13, there is an n such that $\mathrm{I}\Delta_0 + \Omega_1 + Con(\mathrm{I}\mathit{\Delta_0} + \pi, n) \vdash \psi$. Therefore

$$\mathrm{I}\Delta_0 + \Omega_1 + Con(\mathrm{I}\mathit{\Delta_0} + \pi, k) \nvdash Con(\mathrm{I}\mathit{\Delta_0} + \pi, n).$$

QED

Corollary 2.8.17 (Wilkie and Paris [WP 87], Corollary 8.14)

1. $\mathrm{I}\Delta_0 + EXP \nvdash Con(\mathrm{I}\mathit{\Delta_0})$;
2. *If π is a Π_1 sentence, then* $\mathrm{I}\Delta_0 + EXP + \pi \nvdash Con(\mathrm{I}\mathit{\Delta_0} + \pi)$

Proof. We prove the second statement. Suppose that $\mathrm{I}\Delta_0 + EXP + \pi \vdash Con(\mathrm{I}\mathit{\Delta_0} + \pi)$, then by theorem 2.8.13, there is a $k \in \omega$ such that $\mathrm{I}\Delta_0 + \Omega_1 + Con(\mathrm{I}\mathit{\Delta_0} + \pi, k) \vdash Con(\mathrm{I}\mathit{\Delta_0} + \pi)$, contradicting theorem 2.8.16. QED

Corollary 2.8.18 $\mathrm{I}\Delta_0 + EXP \nvdash Con(Q)$, *but* $\mathrm{I}\Delta_0 + SUPEXP \vdash Con(Q)$.

Proof. By theorem 2.6.16, $Q \rhd \mathrm{I}\Delta_0$ on an initial segment. This can be verified in $\mathrm{I}\Delta_0 + \Omega_1$ (see [HP 93, Theorem V.5.12]). Therefore $\mathrm{I}\Delta_0 + \Omega_1 \vdash Con(Q) \leftrightarrow Con(\mathrm{I}\mathit{\Delta_0})$, so by corollary 2.8.17, $\mathrm{I}\Delta_0 + EXP \nvdash Con(Q)$.

The fact that $\mathrm{I}\Delta_0 + SUPEXP \vdash Con(Q)$ follows from theorem 2.8.6, the consideration that as far as $\mathrm{I}\Delta_0 + EXP$ is concerned, tableau provability and cut free provability are equivalent, and from the fact that the cut-elimination theorem for the predicate calculus can be proved in $\mathrm{I}\Delta_0 + SUPEXP$. QED

Lemma 2.8.19 *Let* $U \supseteq \mathrm{I}\Delta_0 + \Omega_1$. *Then* $\mathrm{I}\Delta_0 + \Omega_1$ *proves the following: if U is consistent, then U does not interpret* $\mathrm{I}\Delta_0 + \Omega_1 + Con(U)$.

Proof. The argument is similar to the proof of [Vi 90a, Proposition 6.2.2.2] Reason in $\mathrm{I}\Delta_0 + \Omega_1$ and suppose that $U \rhd \mathrm{I}\Delta_0 + \Omega_1 + Con(U)$ by the interpretation M. Define

$$Prov_W(x) \;:\Longleftrightarrow\; Prov_{\mathrm{I}\Delta_0 + \Omega_1 + Con(U)}(x) \wedge Prov_U(x^M).$$

Because $Prov_W$ can be written as an $\exists\Sigma_1^b$-formula, the principles of L can be verified for $\Box_W$. Moreover we have the IL-consequence

$$W \rhd W + \neg Con(W).$$

See section 2.5 for a definition of IL. The reader may enjoy to figure out how to prove $IL \vdash \top \rhd \Box\bot$. Also we have by definition of $Prov_W$:

$$Prov_W(\neg Con(W) \rightarrow \neg Con(U)),$$

so

$$W \rhd W + \neg Con(U). \tag{2.8}$$

But by definition of $Prov_W$, we have $Prov_W(Con(U))$, so (2.8) implies $\neg Con(W)$. Again by definition of $Prov_W$, we conclude that $\neg Con(U)$. QED

Theorem 2.8.20 $\mathrm{I}\Delta_0 + \Omega_1 + Con(\mathrm{I}\Delta_0) \nvdash Con(\mathrm{I}\Delta_0 + EXP)$.

Proof. Assume that $\mathrm{I}\Delta_0 + \Omega_1 + Con(\mathrm{I}\Delta_0) \vdash Con(\mathrm{I}\Delta_0 + EXP)$.

In $\mathrm{I}\Delta_0 + EXP$, one can prove that the set $\{x \mid 2^x_x \downarrow\}$ is closed under successor, so it can be shortened to an $\mathrm{I}\Delta_0 + EXP$-cut J closed under ω_1. Now

$$\mathrm{I}\Delta_0 + EXP \rhd \mathrm{I}\Delta_0 + \Omega_1 + Con(\mathrm{I}\Delta_0),$$

using the interpretation provided by J. Thus by our starting assumption $\mathrm{I}\Delta_0 + EXP \rhd \mathrm{I}\Delta_0 + \Omega_1 + Con(\mathrm{I}\Delta_0 + EXP)$, contradicting lemma 2.8.19. QED

Remark 2.8.21 Wilkie and Paris [WP 87] show in their Theorem 8.19 that the statement $Con(\mathrm{I}\Delta_0)$ is even more hopelessly weak than theorem 2.8.20 suggests, for adding EXP makes no difference. That is,

$$\mathrm{I}\Delta_0 + EXP + Con(\mathrm{I}\Delta_0) \nvdash Con(\mathrm{I}\Delta_0 + EXP).$$

For the next theorem, we need one lemma.

Lemma 2.8.22 *$\mathrm{I}\Delta_0 + EXP$ is finitely axiomatizable.*

Proof. See [HP 93, Theorem V.5.6]. The proof uses a Δ_0 truth definition for Δ_0-formulas (see lemma 2.6.15) and then follows the argument of step 3 from the proof sketch of theorem 2.6.16. QED

Theorem 2.8.23 *Q does not interpret $\mathrm{I}\Delta_0 + EXP$.*

Proof. Let EXP^* be the finitely axiomatized version of $\mathrm{I}\Delta_0 + EXP$, which exists according to lemma 2.8.22. If $Q \rhd \mathrm{I}\Delta_0 + EXP$, then certainly $Q \rhd EXP^*$.

Since both Q and EXP^* are finitely axiomatized, we have $\mathrm{I}\Delta_0 + \Omega_1 \vdash Q \rhd EXP^*$, so $\mathrm{I}\Delta_0 + \Omega_1 \vdash Con(Q) \rightarrow Con(EXP^*)$, and a fortiori $\mathrm{I}\Delta_0 + \Omega_1 \vdash Con(\mathrm{I}\Delta_0) \rightarrow Con(EXP^*)$, contradicting theorem 2.8.20 QED

Part II

Metamathematics for Bounded Arithmetic

Chapter 3

A small reflection principle for bounded arithmetic

"What a curious feeling!" said Alice,
"I must be shutting up like a telescope!"
And so it was indeed: she was now only ten inches high, and her face brightened up at the thought that she was now the right size for going through the little door into that lovely garden.

(Lewis Carroll, *Alice in Wonderland*)

Abstract. We investigate the theory $I\Delta_0 + \Omega_1$, and strengthen [Bu 86, Theorem 8.6] to the following: if *NP* $\neq$ *co-NP*, then Σ-completeness for witness comparison formulas is not provable in bounded arithmetic, i.e.

$$\begin{aligned} I\Delta_0 + \Omega_1 \nvdash \forall b \forall c \quad & (\exists a(\mathit{Prf}(a,c) \wedge \forall z \leq a \neg \mathit{Prf}(z,b)) \\ & \rightarrow \mathit{Prov}(\ulcorner \exists a(\mathit{Prf}(a,\bar{c}) \wedge \forall z \leq a \neg \mathit{Prf}(z,\bar{b}))\urcorner)). \end{aligned}$$

Next, we study a "small reflection principle" in bounded arithmetic. We prove that for all sentences φ,

$$I\Delta_0 + \Omega_1 \vdash \forall x \mathit{Prov}(\ulcorner \forall y \leq \bar{x}(\mathit{Prf}(y, \overline{\ulcorner \varphi \urcorner}) \rightarrow \varphi)\urcorner)$$

The proof hinges on the use of definable cuts and partial satisfaction predicates akin to those introduced by Pudlák in [Pu 86].

Finally we give some applications of the small reflection principle, showing that the principle can sometimes be invoked in order to circumvent the use of provable Σ-completeness for witness comparison formulas.

3.1 Introduction

A striking feature of Solovay's Theorem that *Löb's logic is complete for arithmetical interpretations* is its amazing stability. If one sticks to the unimodal propositional language and standard arithmetical interpretations, the result holds (modulo a trivial variation) for any decently axiomatized extension of $I\Delta_0 + EXP$. Such stability is in some sense a weakness: unimodal propositional logic combined with the standard interpretation cannot serve to classify or give information on specific theories in a broad range. Of course this weakness disappears when we extend the modal language, but that is not our subject here (however see [Vi 90a, Bek 91]; [Bek 89]).

Is there life outside the broad range of arithmetical theories satisfying Solovay's Completeness Theorem? Clearly the question is only sensible if the theories under consideration verify Löb's logic, or perhaps some still interesting weakening of it.

Two directions of research come to mind. The first one is to weaken the logic of the arithmetical theory. Specifically one can study theories like Heyting Arithmetic (**HA**), the constructive version of Peano Arithmetic. It turns out that **HA** verifies the obvious constructive version of Löb's logic plus a wide variety of extra principles (see [Vi 81, Vi 82, Vi 85]). The only definitive information that we have is a characterization of the closed fragment of **HA**. For all we know the provability logic corresponding to **HA** itself could be Π^0_2-complete. Moreover, extensions of **HA** have quite different provability logics. Note by the way that provability logics need not be monotonic in their arithmetical theories.

The second direction of research is simply to look at classical arithmetical theories that are strictly weaker than, or even incompatible with, $I\Delta_0 + EXP$. It turns out that there are two salient theories of this kind: Paris and Wilkie's $I\Delta_0+\Omega_1$ and Buss' S^1_2, both of them satisfying Löb's logic (see [WP 87, Bu 86]). Does Solovay's Theorem still hold for them? At present nobody knows - or to be precise, we haven't heard that anybody knows.

This chapter is a first contribution to an understanding of the difficulties involved in proving or disproving Solovay's Theorem for theories like $I\Delta_0+\Omega_1$ and S^1_2. Solovay's proof involves Rosser methods. The problem for us resides in the instances of Π^b_1-completeness that occur in the proof. Two points are important.

- We do not know whether the instances of Π^b_1-completeness used in Solovay's proof are provable in our target theories. Buss proved that provability of Π^b_1-completeness with parameters in S^1_2 implies $NP = co\text{-}NP$ (see [Bu 86]). In section 3.2 we elaborate on this theme. To be specific, we prove that if $NP \neq co\text{-}NP$, then Σ-completeness for witness comparison formulas is not provable in bounded arithmetic, i.e.

$$\begin{aligned} I\Delta_0 + \Omega_1 \nvdash \forall b \forall c \quad & (\exists a(\mathit{Prf}(a,c) \wedge \forall z \leq a \neg \mathit{Prf}(z,b)) \\ & \rightarrow \mathit{Prov}(\ulcorner \exists a(\mathit{Prf}(a,\bar{c}) \wedge \forall z \leq a \neg \mathit{Prf}(z,\bar{b}))\urcorner)). \end{aligned}$$

- In many cases we can circumvent the use of instances of Π^b_1-completeness. Švejdar discovered the first alternative argument when he surprisingly provided a proof of Rosser's Theorem that genuinely differed from Rosser's own proof (see [Šv 83]). To this end he introduced a principle which we have dubbed Švejdar's principle. In section 3.3 we prove a "small reflection principle" in our target theories from which Švejdar's principle immediately follows. More precisely, we show that for all sentences φ,

$$I\Delta_0 + \Omega_1 \vdash \forall x \mathit{Prov}(\ulcorner \forall y \leq \bar{x}(\mathit{Prf}(y, \overline{\ulcorner\varphi\urcorner}) \rightarrow \varphi)\urcorner).$$

  Švejdar's principle is not sufficient to derive Solovay's Theorem. However, it has been fruitfully exploited in the dogged attempt to use Solovay-like methods to embed larger and larger classes of Kripke models for Löb's logic in our weak arithmetical theories. The state of this dogged art can be found in chapter 5 and [BV 93].

  We end section 3.3 with some other applications of the small reflection principle.

In section 3.4, we use the small reflection principle in order to extend Krajíček and Pudlák's result on the injection of inconsistencies into models of $\mathrm{I}\Delta_0 + EXP$.

Theorem 3.2.7 and theorem 3.3.20, the main results of section 3.2 and section 3.3, were published previously in the technical report [Ve 89], which in turn is based on my master's thesis [Ve 88].

We assume that the reader is familiar with [Bu 86] and [WP 87]. However, most of the definitions we need can be found in section 2.3.

Remark 3.1.1 In cases where confusion seems unlikely, we will sloppily leave out some numeral dashes, in particular deeper nested ones.

3.2 Σ-completeness and the NP=co-NP problem

In this section we will prove that, under the assumption that *NP* ≠ *co-NP*, the following holds:

$$\begin{aligned}\mathrm{I}\Delta_0 + \Omega_1 \nvdash \forall b \forall c \quad & (\exists a(\mathit{Prf}(a,c) \wedge \forall z \leq a \neg \mathit{Prf}(z,b)) \\ & \rightarrow \mathit{Prov}(\ulcorner \exists a(\mathit{Prf}(a,\bar{c}) \wedge \forall z \leq a \neg \mathit{Prf}(z,\bar{b}))\urcorner)).\end{aligned}$$

In the proofs of the lemmas leading up to this result we will frequently, often without mention, make use of the following proposition and its corollary.

Proposition 3.2.1 (**[Bu 86]**) *Suppose A is a closed, bounded formula in the language of S^1_2 and let* **R** *be a consistent theory extending S^1_2. Then* $\mathbf{R} \vdash A$ *iff* $\omega \models A$.

Corollary 3.2.2 (**[Bu 86], Prop. 8.3)** *Suppose $A(\vec{a})$ is a bounded formula in the language of S^1_2,and let* **R** *be a consistent theory extending S^1_2. If* $\mathbf{R} \vdash \forall \vec{x} A(\vec{x})$, *then* $\omega \models \forall \vec{x} A(\vec{x})$.

In this section, we will use the name $\mathrm{I}\Delta_0+\Omega_1$ for Buss' theory S_2 (see Definition 2.3.12), in which induction for formulas from the hierarchy of bounded arithmetic formulas in a language containing $|\ |$, $\lfloor \frac{1}{2}x \rfloor$ and # is allowed. Because S_2 is conservative over $\mathrm{I}\Delta_0 + \Omega_1$, the name change has no repercussions on the results of this section. (In the next section, where we need to construct formalized satisfaction predicates, we will be more careful.)

In order to prove the main theorem of this section, we need to prove a few seemingly far-fetched lemmas. Their proofs borrow heavily from the formalization carried out in [Bu 86]. To make these lemmas understandable, we will give some details of the formalization of the predicate *Prf*. Buss uses a sequent calculus akin to Takeuti's (see [Ta 75]). He considers a proof to be formalized as a tree, of which the root corresponds to the end sequent, and the leaves to the initial sequents of the proof. Every node of the proof tree is labeled by an ordered pair $\langle a, b\rangle$. The second member of this pair codes a sequent, and the first member codes the rule of inference by which this sequent has been derived from the sequents corresponding to the children of the node in question. For leaves, the first member of the corresponding ordered pair codes the axiom of which the initial sequent is an instantiation.

The only extra fact we need here is that logical axioms are all numbered 0; in particular, for all terms t, the tree containing just one node labeled $\langle 0, \ulcorner \longrightarrow t = t \urcorner\rangle$ is a proof of $\longrightarrow t = t$. Because of a peculiarity in the encoding of trees, by which 0 and 1 are reserved as codes for brackets, Buss encodes the proof just mentioned by $\langle 0, \ulcorner \longrightarrow t = t\urcorner\rangle + 2$.

In the sequel, we will sometimes abuse Buss' conventions in order to keep the formulas legible. Thus, we will write $\langle 0, \ulcorner \longrightarrow \bar{d} = \bar{d} \urcorner \rangle$ for Buss' $\langle 0, (0 * \overline{Arrow}) * *(\ulcorner I_d \urcorner * \overline{Equals} * *\ulcorner I_d \urcorner)\rangle + 2$.

Lemma 3.2.3 *Let $\psi(d, b)$ be the formula $\forall z \leq \langle 0, \ulcorner \longrightarrow \bar{d} = \bar{d} \urcorner \rangle \neg \mathit{Prf}(z, b)$. The predicate represented by ψ is co-NP-complete.*

Proof. Straightforwardly, ψ is a Π^b_1-formula, hence it represents a *co-NP* predicate. For the other side, viz. *co-NP*-hardness, begin by supposing $A(a_1, \ldots, a_k) \in$ *co-NP*. We will polynomially reduce A to ψ. (For definitions of the complexity theoretic concepts that we mention, see definition 2.3.7 and definition 2.3.8; and see remark 2.3.10 or [Bu 86, Thm 1.8]).

By provable Σ^b_1-completeness (see theorem 2.3.24), there is a term $r(\vec{a})$ such that

$$\mathrm{I}\Delta_0 + \Omega_1 \vdash \neg A(\vec{a}) \rightarrow \exists z \leq r(\vec{a}) \mathit{Prf}(z, \ulcorner \neg A(\overline{a_1}, \ldots, \overline{a_k}) \urcorner)$$

and thus

$$\omega \models \neg A(\vec{a}) \rightarrow \exists z \leq r(\vec{a}) \mathit{Prf}(z, \ulcorner \neg A(\overline{a_1}, \ldots, \overline{a_k}) \urcorner)$$

Because $r(\vec{a}) \leq \ulcorner \overline{r(\vec{a})} \urcorner \leq \langle 0, \ulcorner \longrightarrow \overline{r(\vec{a})} = \overline{r(\vec{a})} \urcorner \rangle$, we also have

$$\omega \models \neg A(\vec{a}) \rightarrow \exists z \leq \langle 0, \ulcorner \longrightarrow \overline{r(\vec{a})} = \overline{r(\vec{a})} \urcorner \rangle \mathit{Prf}(z, \ulcorner \neg A(\overline{a_1}, \ldots, \overline{a_k}) \urcorner). \tag{3.1}$$

On the other hand, by Proposition 3.2.1 and the consistency of $\mathrm{I}\Delta_0 + \Omega_1$, we have

$$\omega \models \exists z \leq \langle 0, \ulcorner \longrightarrow \overline{r(\vec{a})} = \overline{r(\vec{a})} \urcorner \rangle \mathit{Prf}(z, \ulcorner \neg A(\overline{a_1}, \ldots, \overline{a_k}) \urcorner) \rightarrow \neg A(\vec{a}). \tag{3.2}$$

From (3.1) and (3.2), we conclude that

$$\omega \models A(\vec{a}) \leftrightarrow \forall z \leq \langle 0, \ulcorner \longrightarrow \overline{r(\vec{a})} = \overline{r(\vec{a})} \urcorner \rangle \neg \mathit{Prf}(z, \ulcorner \neg A(\overline{a_1}, \ldots, \overline{a_k}) \urcorner).$$

This means by the definition of ψ that

$$\omega \models A(\vec{a}) \leftrightarrow \psi(r(\vec{a}), \ulcorner \neg A(\overline{a_1}, \ldots, \overline{a_k}) \urcorner).$$

As both $\ulcorner \neg A(\overline{a_1}, \ldots, \overline{a_k}) \urcorner$ and $r(\vec{a})$ can be computed from $\vec{a}$ by polynomial time functions, we have reduced the *co-NP* predicate A to ψ. QED

Lemma 3.2.4 *Let $B(a_1, \ldots, a_k)$ be a Π^b_1-formula representing a co-NP complete predicate. If NP $\neq$ co-NP, then*

$$\mathrm{I}\Delta_0 + \Omega_1 \nvdash \forall \vec{a} (B(\vec{a}) \rightarrow \mathit{Prov}(\ulcorner B(\overline{a_1}, \ldots, \overline{a_k}) \urcorner)).$$

Proof. An application of Parikh's Theorem for $\mathrm{I}\Delta_0 + \Omega_1$ (cf. theorem 2.3.15). We leave the details, which are similar to part of the proof of [Bu 86, Theorem 8.6], to the reader. QED

Lemma 3.2.5 *If NP $\neq$ co-NP, then*

$$\begin{aligned} \mathrm{I}\Delta_0 + \Omega_1 \nvdash \forall b \forall d \quad & (\forall z \leq \langle 0, \ulcorner \longrightarrow \bar{d} = \bar{d} \urcorner \rangle \neg \mathit{Prf}(z, \bar{b}) \\ & \rightarrow \mathit{Prov}(\ulcorner \forall z \leq \langle 0, \ulcorner \longrightarrow \bar{d} = \bar{d} \urcorner \rangle \neg \mathit{Prf}(z, \bar{b}) \urcorner). \end{aligned}$$

Proof. Directly from Lemma 3.2.3 and Lemma 3.2.4. QED

Lemma 3.2.6 $\mathrm{I}\Delta_0 + \Omega_1$ *proves the following:*

$$\begin{aligned} \forall b \forall d \quad & (\mathit{Prov}(\ulcorner \exists a(\mathit{Prf}(a, \ulcorner \longrightarrow \bar{d} = \bar{d}\urcorner) \wedge \forall z \leq a \neg \mathit{Prf}(z, \bar{b}))\urcorner) \\ & \rightarrow \mathit{Prov}(\ulcorner \forall z \leq \langle 0, \ulcorner \longrightarrow \bar{d} = \bar{d}\urcorner\rangle \neg \mathit{Prf}(z, \bar{b})\urcorner)) \end{aligned}$$

Proof. It is not difficult to see that for Buss' formalization of *Prf*, we have the following:

$$\mathrm{I}\Delta_0 + \Omega_1 \vdash \forall d \forall a(\mathit{Prf}(a, \ulcorner \longrightarrow \bar{d} = \bar{d}\urcorner) \rightarrow a \geq \langle 0, \ulcorner \longrightarrow \bar{d} = \bar{d}\urcorner\rangle),$$

and thus

$$\begin{aligned} \mathrm{I}\Delta_0 + \Omega_1 \vdash \forall b \forall d \quad & (\exists a(\mathit{Prf}(a, \ulcorner \longrightarrow \bar{d} = \bar{d}\urcorner) \wedge \forall z \leq a \neg \mathit{Prf}(z, b)) \\ & \rightarrow \forall z \leq \langle 0, \ulcorner \longrightarrow \bar{d} = \bar{d}\urcorner\rangle \neg \mathit{Prf}(z, b)). \end{aligned}$$

This in turn immediately implies our lemma. QED

Theorem 3.2.7 *If NP ≠ co-NP, then*

$$\begin{aligned} \mathrm{I}\Delta_0 + \Omega_1 \nvdash \forall b \forall c \quad & (\exists a(\mathit{Prf}(a, c) \wedge \forall z \leq a \neg \mathit{Prf}(z, b)) \\ & \rightarrow \mathit{Prov}(\ulcorner \exists a(\mathit{Prf}(a, \bar{c}) \wedge \forall z \leq a \neg \mathit{Prf}(z, \bar{b}))\urcorner)). \end{aligned}$$

Proof. Suppose that *NP* ≠ *co-NP*, and suppose, in order to derive a contradiction, that

$$\begin{aligned} \mathrm{I}\Delta_0 + \Omega_1 \vdash \forall b \forall c \quad & (\exists a(\mathit{Prf}(a, c) \wedge \forall z \leq a \neg \mathit{Prf}(z, b)) \\ & \rightarrow \mathit{Prov}(\ulcorner \exists a(\mathit{Prf}(a, \bar{c}) \wedge \forall z \leq a \neg \mathit{Prf}(z, \bar{b}))\urcorner)). \end{aligned}$$

Then, in particular,

$$\begin{aligned} \mathrm{I}\Delta_0 + \Omega_1 \vdash \quad \forall b \forall d \quad & (\mathit{Prf}(\langle 0, \ulcorner \longrightarrow \bar{d} = \bar{d}\urcorner\rangle, \ulcorner \longrightarrow \bar{d} = \bar{d}\urcorner) \\ & \wedge \forall z \leq \langle 0, \ulcorner \longrightarrow \bar{d} = d\urcorner\rangle \neg \mathit{Prf}(z, b) \\ & \rightarrow \mathit{Prov}(\ulcorner \exists a(\mathit{Prf}(a, \ulcorner \longrightarrow \bar{d} = \bar{d}\urcorner) \wedge \forall z \leq a \neg \mathit{Prf}(z, \bar{b}))\urcorner)). \quad (3.3) \end{aligned}$$

We know that

$$\mathrm{I}\Delta_0 + \Omega_1 \vdash \forall d(\mathit{Prf}(\langle 0, \ulcorner \longrightarrow \bar{d} = \bar{d}\urcorner\rangle, \ulcorner \longrightarrow \bar{d} = \bar{d}\urcorner))$$

Combined with (3.3), this implies the following:

$$\begin{aligned} \mathrm{I}\Delta_0 + \Omega_1 \vdash \quad \forall b \forall d \quad & (\forall z \leq \langle 0, \ulcorner \longrightarrow \bar{d} = \bar{d}\urcorner\rangle \neg \mathit{Prf}(z, b) \\ & \rightarrow \mathit{Prov}(\ulcorner \exists a(\mathit{Prf}(a, \ulcorner \longrightarrow \bar{d} = \bar{d}\urcorner) \wedge \forall z \leq a \neg \mathit{Prf}(z, \bar{b}))\urcorner)). \end{aligned}$$

Now we apply Lemma 3.2.6 to derive

$$\begin{aligned} \mathrm{I}\Delta_0 + \Omega_1 \vdash \quad \forall b \forall d \quad & (\forall z \leq \langle 0, \ulcorner \longrightarrow \bar{d} = \bar{d}\urcorner\rangle \neg \mathit{Prf}(z, b) \\ & \rightarrow \mathit{Prov}(\ulcorner \forall z \leq \langle 0, \ulcorner \longrightarrow \bar{d} = \bar{d}\urcorner\rangle \neg \mathit{Prf}(z, \bar{b})\urcorner)), \end{aligned}$$

in contradiction with Lemma 3.2.5. QED

Remark 3.2.8 We can prove that provable Σ_1^0-completeness fails already for a much simpler Π_1^b-formula $\chi(a,b,c)$ defined as follows:

$$\chi(a,b,c) \ := \ \forall x \leq c \forall y \leq c(a \cdot x^2 + b \cdot y \neq c).$$

The fact that Σ_1^0-completeness fails for χ follows immediately from Lemma 3.2.4 and the following lemma, to which A. Wilkie attracted our attention.

Lemma 3.2.9 (Manders and Adleman, see [MA 78]) *The set of equations of the form $(a \cdot x^2 + b \cdot y = c)$, solvable over the natural numbers, with a, b, c positive natural numbers, is NP-complete.*

Note that Lemma 3.2.9 implies that the formula $\exists x \leq c \exists y \leq c(a \cdot x^2 + b \cdot y = c)$ represents an *NP*-complete predicate, and thus that χ as defined above represents a *co-NP* complete predicate.

3.3 The small reflection principle

In this section, we will present a proof of the fact that $I\Delta_0 + \Omega_1$ proves the small reflection principle, i.e. for all φ:

$$I\Delta_0 + \Omega_1 \vdash \forall x \Box(\Box_x \varphi \rightarrow \varphi),$$

where $\Box\varphi$ is an abbreviation for $Prov(\ulcorner\varphi\urcorner)$ and $\Box_x\varphi$ is a formalization of the fact that φ has a proof in $I\Delta_0 + \Omega_1$ of Gödel number $\leq x$. In fact, all arguments that we use can be carried out already in Buss' S_2^1, as the reader may check for him/herself.

In the proof, we will use the existence of partial truth- (or satisfaction-) predicates Sat_n for formulas of length $\leq n$. The intended meaning of $Sat_n(x,w)$ will be "the formula of length $\leq n$ with Gödel number x is satisfied by the assignment sequence coded by w". Pudlák [Pu 86] has constructed partial truth predicates much like the ones we need. (An analogous construction, where Sat_n is related to quantifier depth instead of length, can be found in [Pu 87].)

However, our construction departs from Pudlák's in two ways. Firstly, whereas Pudlák presents his results for theories in relational languages, we allow function symbols.

Secondly and more importantly, $I\Delta_0 + \Omega_1$ is neither finitely nor sparsely axiomatized. Regrettably we cannot even apply to $I\Delta_0 + \Omega_1$ a trick of Pudlák's which turns some non-sparse theories like *PA* and *ZF* into sparse ones (see Theorem 5.5. of [Pu 86]). Therefore we introduce new satisfaction predicates $Sat_{n,\Delta}(x,w)$ with as intended meaning: "the Δ_0-formula of length $\leq n$ with Gödel number x is satisfied by the assignment sequence coded by w". Using these satisfaction predicates, we will be able to prove by short proofs that the Δ_0-induction axioms are true.

In order to start the construction of short satisfaction predicates, we need a few more assumptions and definitions. First of all, when formalizing, we view $I\Delta_0+\Omega_1$ in a restricted way more akin to Paris and Wilkie [WP 87] than to Buss [Bu 86]: see definition 2.3.1.

For this system, we can define the appropriate Δ_1^b-predicates $Term(v)$, $Fmla(v)$, $Sent(v)$, $Prf(u,v)$ in S_2^1.

In Buss' formalization of sequences, $*$ stands for a function which adds a new element to the end of a sequence; $**$ stands for a function which concatenates two sequences; and $\beta(t,w)$ stands for the function giving the value of the t-th place in the sequence coded by w.

In this chapter, we denote concatenation of sequences sloppily by juxtaposition, and we leave out some outer parentheses; thus, for example, $y^{\ulcorner} \rightarrow {}^{\urcorner}z$ stands for Buss's $(0 * \overline{LParen}) * *(y * \overline{Implies}) * *(z * \overline{RParen})$.

Definition 3.3.1 We formally define four concepts that we need in order to construct truth predicates.

- $w =_i w' \;:=\; Len(w) = Len(w') \wedge \forall t(t \leq Len(w) \wedge t \neq i \rightarrow \beta(t, w) = \beta(t, w'))$, i.e. the only possible difference between the sequences coded by w and w' is at the i-th value;
- $Fmla_n(v) \;:=\; Fmla(v) \wedge Len(v) \leq n$, i.e. v is the Gödel number of a formula of length $\leq n$;
- $Fmla_{n,\Delta}(v) \;:=\; Fmla_n(v)$ "and v codes a Δ_0-formula";
- $Evalseq(x, w)$ will mean that the sequence coded by w is long enough to evaluate all variables appearing in x, i.e.

$$\begin{aligned} Evalseq(x, w) \;:=\; & Seq(w) \wedge (Fmla(x) \vee Term(x)) \wedge \forall i(\text{“\textit{the variable} } v_i \\ & \text{\textit{occurs in the term or formula with Gödel number} } x\text{”} \\ & \rightarrow Len(w) \geq i). \end{aligned}$$

Furthermore, we introduce the following two abbreviations:

- $Evalseq_n(x, w) \;:=\; Fmla_n(x) \wedge Evalseq(x, w)$;
- $Evalseq_{n,\Delta}(x, w) \;:=\; Fmla_{n,\Delta}(x) \wedge Evalseq(x, w)$;

Next we define, by Buss' method of p-inductive definitions, a function *Val* such that, if $t(v_{i_1}, \ldots v_{i_n})$ is a term of the (restricted) language of $\mathrm{I}\Delta_0 + \Omega_1$ and w codes a sequence evaluating all variables $v_{i_1}, \ldots v_{i_n}$ appearing in t, then $Val(\ulcorner t \urcorner, w)$ gives the value of $t[\beta(i_1, w), \ldots, \beta(i_n, w)]$.

Definition 3.3.2 Let *Val* satisfy the following conditions:

- $\neg Term(t) \vee \neg Evalseq(t, w) \rightarrow Val(t, w) = 0$;
- the p-inductive condition:

$$\begin{aligned} & Term(t) \wedge Evalseq(t, w) \rightarrow \\ & \quad (t = \ulcorner 0 \urcorner \wedge Val(t, w) = 0) \\ & \quad \vee \exists i < t(t = \ulcorner v_i \urcorner \wedge Val(t, w) = \beta(i, w)) \\ & \quad \vee \exists t_1, t_2 < t(Term(t_1) \wedge Term(t_2) \\ & \quad \wedge ((t = \ulcorner S \urcorner t_1 \wedge Val(t, w) = S(Val(t_1, w))) \\ & \quad \vee (t = t_1{}^{\ulcorner} + {}^{\urcorner}t_2 \wedge Val(t, w) = Val(t_1, w) + Val(t_2, w)) \\ & \quad \vee (t = t_1{}^{\ulcorner} \cdot {}^{\urcorner}t_2 \wedge Val(t, w) = Val(t_1, w) \cdot Val(t_2, w)))) \end{aligned}$$

By induction, we can show that $t\#w$ will be a bound for $Val(t, w)$. Thus, by [Bu 86, Theorem 7.3], *Val* is Δ_1^b-definable (thus provably total) in S_2^1; furthermore, the definition of *Val* in S_2^1 is intensionally correct in that properties of *Val* can be proved in S_2^1 (and thus also in $\mathrm{I}\Delta_0 + \Omega_1$) by the use of induction.

Remark 3.3.3 Note that we cannot construct in $I\Delta_0 + \Omega_1$ a correct valuation function *Val* for a language that contains #. For, to any a we can associate a formalized term $f(a)$ given informally as $1\#2\#\ldots\#2$ where the number of 2's is $|a|$. A correctly defined *Val* should give $Val(f(a), w) = \exp(\exp(|a|+1)-2) \geq \exp(a)$ (cf. [Ta 88]). Therefore by Parikh's Theorem (cf. theorem 2.3.15), *Val* could not be Δ_0-definable and provably total in $I\Delta_0 + \Omega_1$.

In the sequel, we will freely make use of induction for $\Delta_0(Val)$-formulas in $I\Delta_0 + \Omega_1$, as is justified by the $I\Delta_0 + \Omega_1$-analogs of Buss' Theorem 2.2 and Corollary 2.3. We will especially need the following lemma.

Lemma 3.3.4 *There exists a constant c such that for every term t with free variables among $v_{i_1}, \ldots, v_{i_m}$ and for every n with $Len(\ulcorner t \urcorner) \leq n$, we can prove the following by proofs of length $\leq c \cdot n$:*

$$I\Delta_0 + \Omega_1 \vdash Evalseq(\ulcorner t \urcorner, w) \rightarrow Val(\ulcorner t \urcorner, w) = t[\beta(i_1, w), \ldots, \beta(i_m, w)].$$

Proof. Straightforward by induction on the construction of t. QED

For the definition of satisfaction predicates, we need one more definition.

Definition 3.3.5 We formally define the following:

$$s(i, x, w) \ := \ (Subseq(w, 1, i) * x) * *Subseq(w, i+1, Len(w)+1).$$

Thus, if w is a sequence of length $\geq i$, $s(i, x, w)$ denotes the sequence which is identical to w, except that x appears in the i-th place.

Definition 3.3.6 We say that $Sat_n(x, w)$ is a *partial definition of truth for formulas of length* $\leq n$ in $I\Delta_0 + \Omega_1$ iff $I\Delta_0 + \Omega_1 \vdash Evalseq_n(x, w) \rightarrow$

$$\begin{array}{l}
\{Sat_n(x, w) \leftrightarrow \\
\quad [\exists t, t' < x(Term(t) \wedge Term(t') \wedge x = t^{\ulcorner} = {}^{\urcorner}t' \wedge Val(t, w) = Val(t', w)) \\
\quad \vee \exists t, t' < x(Term(t) \wedge Term(t') \wedge x = t^{\ulcorner} \leq {}^{\urcorner}t' \wedge Val(t, w) \leq Val(t', w)) \\
\quad \vee \exists y < x(x = {}^{\ulcorner}\neg^{\urcorner}y \wedge \neg Sat_n(y, w)) \\
\quad \vee \exists y, z < x(x = y^{\ulcorner} \rightarrow {}^{\urcorner}z \wedge (Sat_n(y, w) \rightarrow Sat_n(z, w))) \\
\quad \vee \exists y, i < x(x = {}^{\ulcorner}\forall v_i{}^{\urcorner}y \wedge \forall w'(w =_i w' \rightarrow Sat_n(y, w'))) \\
\quad \vee \exists y, i, t < x(Term(t) \wedge x = {}^{\ulcorner}(\forall v_i \leq {}^{\urcorner}t^{\ulcorner}){}^{\urcorner}y \wedge \\
\quad \forall w' \leq s(i, Val(t, w), w)(w =_i w' \wedge \beta(i, w) \leq Val(t, w) \rightarrow Sat_n(y, w')))]\}
\end{array}$$

We denote the part between brackets [ ] on the right hand side of the equivalence by $\Sigma(Sat_n; x, w)$; note that these are just Tarski's conditions.

Similarly, we say that $Sat_{n,\Delta}(x, w)$ is a *partial definition of truth for Δ_0-formulas of length* $\leq n$ in $I\Delta_0 + \Omega_1$ iff $I\Delta_0 + \Omega_1 \vdash Evalseq_{n,\Delta}(x, w) \rightarrow$

$$\begin{array}{l}
\{Sat_{n,\Delta}(x, w) \leftrightarrow \\
\quad [\exists t, t' < x(Term(t) \wedge Term(t') \wedge x = t^{\ulcorner} = {}^{\urcorner}t' \wedge Val(t, w) = Val(t', w)) \\
\quad \vee \exists t, t' < x(Term(t) \wedge Term(t') \wedge x = t^{\ulcorner} \leq {}^{\urcorner}t' \wedge Val(t, w) \leq Val(t', w)) \\
\quad \vee \exists y < x(x = {}^{\ulcorner}\neg^{\urcorner}y \wedge \neg Sat_{n,\Delta}(y, w)) \\
\quad \vee \exists y, z < x(x = y^{\ulcorner} \rightarrow {}^{\urcorner}z \wedge (Sat_{n,\Delta}(y, w) \rightarrow Sat_{n,\Delta}(z, w))) \\
\quad \vee \exists y, i, t < x(Term(t) \wedge x = {}^{\ulcorner}(\forall v_i \leq {}^{\urcorner}t^{\ulcorner}){}^{\urcorner}y \wedge \\
\quad \forall w' \leq s(i, Val(t, w), w)(w =_i w' \wedge \beta(i, w) \leq Val(t, w) \rightarrow Sat_{n,\Delta}(y, w')))]\}
\end{array}$$

We denote the part between brackets [ ] on the right hand side of the equivalence by $\Sigma_\Delta(Sat_{n,\Delta}; x, w)$. Note that the only difference between $\Sigma(Sat_n; x, w)$ and $\Sigma_\Delta(Sat_{n,\Delta}; x, w)$ is that in the latter, the disjunct for the unbounded quantifier $\forall$ is left out.

In the proof of the main theorem of this section, we will reason inside $I\Delta_0+\Omega_1$, and we will need the existence of Gödel numbers representing formulas Sat_n that provably satisfy the conditions of the preceding definition. Therefore, in the unformalized proofs below, we take care that the formulas Sat_n and the proofs that they have the right properties be bounded by suitable terms. The following lemmas provide us with such formulas. In [Pu 86, Pu 87] Pudlák proves similar lemmas for a language without function symbols. Below, we sketch the adaptation of his method to our case. The parallel construction of a $\Delta_0(Val, |\ |, \lfloor\frac{1}{2}x\rfloor, \#)$-formula $Sat_{n,\Delta}$ which works for Δ_0-formulas is particular to this dissertation. We use the formula $Sat_{n,\Delta}$ only in our proof that Sat_n preserves the Δ_0-induction axioms, but there its use is essential.

Lemma 3.3.7 *There exist formulas $Sat_n(x,w)$ for $n = 0, 1, 2, \ldots$ of length linear in n, and such that, by a proof of length linear in n,*

$$I\Delta_0+\Omega_1 \vdash Evalseq_{n+1}(x,w) \rightarrow (Sat_{n+1}(x,w) \leftrightarrow \Sigma(Sat_n; x, w)).$$

Proof. Sat_n is constructed by recursion. We can define Sat_0 arbitrarily, as there are no formulas of length ≤ 0. If we have the formula Sat_k, we obtain Sat_{k+1} by substituting Sat_k for Sat_n in the formula $\Sigma(Sat_n; x, w)$ defined in Definition 3.3.6.

Remember that we have to ensure that the length of the formula Sat_n grows linearly in n. However, if we straightforwardly used $\Sigma(Sat_n; x, w)$ as defined above, the length of Sat_n would grow exponentially in n, because $\Sigma(Sat_n; x, w)$ contains more than one occurrence of Sat_n.

Ferrante and Rackoff (in [FR 79, Chapter 7]) describe a general technique for writing short formulas, due to Fischer and Rabin. Using these techniques, one can replace $\Sigma(Sat_n; x, w)$ by a formula $\Sigma'(Sat_n; x, w)$ which contains only one occurrence of Sat_n, and which is equivalent to $\Sigma(Sat_n; x, w)$ in a very weak theory - say predicate logic plus the axiom $S0 \neq 0$.

Ferrante and Rackoff use the inclusion of $\leftrightarrow$ in the language of the theory in an essential way. However, Solovay sent us a different construction of short formulas which circumvents the use of $\leftrightarrow$. With his kind permission, we present a sketch of his proof.

Solovay's basic idea is to shift attention from sets to characteristic functions. Without restriction of generality, we may assume that we work with unary predicates $Sat_n(x)$ instead of $Sat_n(x,w)$. Let

$$F_n(x,y) \ := \ (y = S0 \wedge Sat_n(x)) \vee (y = 0 \wedge \neg Sat_n(x)).$$

If we can find a formula H_n equivalent to F_n of length proportional to n, it will be easy to define using this formula our desired formula Sat_{n+1}.

Let L be the language of $I\Delta_0+\Omega_1$ enriched with a new binary predicate letter G. We can find a formula Φ of L in prenex normal form, having only the variables x and y free, such that if G is interpreted as F_n, then Φ is interpreted as F_{n+1}. We show how to find a formula Ψ which is equivalent to Φ and which has only one occurrence of G. Assume that Φ starts with the string of quantifiers $(Q_1x_1)\ldots(Q_rx_r)$, and that there are k occurrences of G in the matrix of Φ, say $G(t_1,m_1),\ldots,G(t_n,m_k)$. The formula Ψ will have the form

$$(Q_1x_1)\ldots(Q_rx_r)(\exists y_1)\ldots(\exists y_k)[M \wedge S].$$

Here $y_1,\ldots,y_k$ are fresh variables (for the moment - in the final definition we will be less liberal with variables). The formula M is obtained from the matrix of Φ by replacing each occurrence of $G(t_i, m_i)$ by $m_i = y_i$. S' job is to ensure that the y_i's are chosen correctly. It is defined as follows.

$$S := \forall w_1 \exists w_2 [G(w_1, w_2) \wedge \bigwedge_{i=1}^{k} (w_1 = t_i \rightarrow w_2 = y_i)].$$

If we define H_{n+1} from H_n using Ψ, we get a formula of length proportional to $n \log n$, because at every step we introduce fresh variables in order to avoid clashes. There are however tricks to get by with a finite set of variables, as the reader may enjoy to figure out (or to look up in [FR 79, Chapter 7]).

We will write $\Sigma'(Sat_n; x, w)$ for the equivalent of $\Sigma(Sat_n; x, w)$ resulting from an application of the techniques described above. The length of Sat_n thus constructed via iterated application of Σ' to Sat_0 is indeed linear in n. Moreover, for all n, the *shape* of the proof of $\Sigma(Sat_n; x, w) \leftrightarrow \Sigma'(Sat_n; x, w)$ is the same for all n. Thus, the proofs of $\Sigma(Sat_n; x, w) \leftrightarrow \Sigma'(Sat_n; x, w)$ grow linearly in n. Hence, as $Sat_{n+1}(x, w) \equiv \Sigma'(Sat_n; x, w)$, we have the following by proofs of length linear in n:

$$\mathrm{I}\Delta_0 + \Omega_1 \vdash Sat_{n+1}(x, w) \leftrightarrow \Sigma(Sat_n; x, w) \tag{3.4}$$

QED

Lemma 3.3.8 $\mathrm{I}\Delta_0 + \Omega_1$ *proves by a proof of length of the order of n^2 that the formula Sat_n as constructed in Lemma 3.3.7 is a partial definition of truth for formulas of length $\leq n$.*

Proof. We want short proofs showing that Sat_n is a partial definition of truth for formulas of length $\leq n$ in $\mathrm{I}\Delta_0 + \Omega_1$, i.e.

$$\mathrm{I}\Delta_0 + \Omega_1 \vdash Evalseq_n(x, w) \rightarrow (Sat_n(x, w) \leftrightarrow \Sigma(Sat_n; x, w)).$$

By (3.4), it suffices to show that, by proofs of length of the order n^2,

$$\mathrm{I}\Delta_0 + \Omega_1 \vdash Evalseq_n(x, w) \rightarrow (Sat_n(x, w) \leftrightarrow Sat_{n+1}(x, w)).$$

This can be proved by external induction on n. In fact, when we define

$$\Phi_n := \forall x \forall w (Evalseq_n(x, w) \rightarrow (Sat_n(x, w) \leftrightarrow Sat_{n+1}(x, w))),$$

the proofs of $\Phi_n \rightarrow \Phi_{n+1}$ in $\mathrm{I}\Delta_0 + \Omega_1$ will have a shape which does not depend on n. (We refer those readers who seek elucidation by examples to [Pu 86, Lemma 5.1].) We can observe that every proof in $\mathrm{I}\Delta_0 + \Omega_1$ of $\Phi_n \rightarrow \Phi_{n+1}$ is the instantiation of a single proof scheme. Thus, the length of the proofs of $\Phi_n \rightarrow \Phi_{n+1}$ increases only linearly in n, so that the length of the proof in $\mathrm{I}\Delta_0 + \Omega_1$ of

$$\forall x \forall w (Evalseq_n(x, w) \rightarrow (Sat_n(x, w) \leftrightarrow Sat_{n+1}(x, w))),$$

is of the order n^2. QED

Lemma 3.3.9 *There exist formulas $Sat_{n,\Delta}(x, w)$ for $n = 0, 1, 2, \ldots$ of lengths linear in n, and such that $\mathrm{I}\Delta_0 + \Omega_1$ proves by proofs of length linear in n that $Sat_{n+1,\Delta}(x, w) \leftrightarrow \Sigma_\Delta(Sat_{n,\Delta}; x, w)$. The resulting formulas $Sat_{n,\Delta}(x, w)$ are $\Delta_0(Val)$-formulas.*

Proof. Completely analogous to the proof of Lemma 3.3.7. Because $\Sigma_\Delta(Sat_{n,\Delta}; x, w)$ contains only bounded quantifiers, and because all quantifiers introduced by the Solovay method can be bounded, the resulting formulas are indeed $\Delta_0(Val)$. QED

Lemma 3.3.10 $I\Delta_0 + \Omega_1$ *proves by a proof of length of the order of* n^2 *that the formula* $Sat_{n,\Delta}(x, w)$ *as constructed in Lemma 3.3.9 is a partial definition of truth for* Δ_0*-formulas of length* $\leq n$.

Proof. We adapt the proof of Lemma 3.3.8, incorporating the fact that we are concerned with Δ_0-formulas only. Thus instead of Φ_n, we define

$$\Phi_{n,\Delta} := \forall x \forall w(Evalseq_{n,\Delta}(x, w) \rightarrow (Sat_{n,\Delta}(x, w) \leftrightarrow Sat_{n+1,\Delta}(x, w))).$$

The proof of $\Phi_{n,\Delta} \rightarrow \Phi_{n+1,\Delta}$ runs along the same lines as the proof of $\Phi_n \rightarrow \Phi_{n+1}$, using the extra fact that if $x = y^{\ulcorner} \rightarrow {}^{\urcorner}z$ and $Fmla_{n+1,\Delta}(x)$, then $Fmla_{n,\Delta}(y)$ and $Fmla_{n,\Delta}(z)$, etc. QED

We now show that the partial definitions of truth can, by proofs of quadratic length, be proven to satisfy Tarski's conditions, which justifies their name.

Lemma 3.3.11 (cf. [Pu 86, Pu 87]) *There exists a constant c such that for every formula* φ *with free variables among* $v_{i_1}, \ldots, v_{i_m}$ *and for every n with* $Len(\ulcorner\varphi\urcorner) \leq n$, *we can prove the following by proofs of length* $\leq c \cdot n^2$:

$$I\Delta_0 + \Omega_1 \vdash \forall w(Evalseq(\ulcorner\varphi\urcorner, w) \rightarrow (Sat_n(\ulcorner\varphi\urcorner, w) \leftrightarrow \varphi[\beta(i_1, w), \ldots, \beta(i_m, w)])) \quad (3.5)$$

and, if φ *is a* Δ_0*-formula, we can also prove the following by proofs of length* $\leq c \cdot n^2$:

$$I\Delta_0 + \Omega_1 \vdash \forall w(Evalseq(\ulcorner\varphi\urcorner, w) \rightarrow (Sat_{n,\Delta}(\ulcorner\varphi\urcorner, w) \leftrightarrow \varphi[\beta(i_1, w), \ldots, \beta(i_m, w)]))(3.6)$$

Proof. By cases. If φ is an atomic formula $t \leq t'$ of length $\leq n$ and with free variables among $v_{i_1}, \ldots, v_{i_m}$, Lemma 3.3.8 implies that we can prove the following by proofs of length linear in n:

$$\begin{aligned} I\Delta_0 + \Omega_1 \vdash \forall w \quad & (Evalseq(\ulcorner t \leq t'\urcorner, w) \\ & \rightarrow (Sat_n(\ulcorner t \leq t'\urcorner, w) \leftrightarrow Val(\ulcorner t\urcorner, w) \leq Val(\ulcorner t'\urcorner, w))) \end{aligned}$$

By Lemma 3.3.4, we can then conclude that we can prove the following by proofs of length linear in n:

$$\begin{aligned} I\Delta_0 + \Omega_1 \vdash \forall w \quad & (Evalseq(\ulcorner t \leq t'\urcorner, w) \\ & \rightarrow (Sat_n(\ulcorner t \leq t'\urcorner, w) \leftrightarrow (t \leq t')[\beta(i_1, w), \ldots, \beta(i_m, w)])) \end{aligned}$$

The case for $t = t'$ is analogous.

For the non-atomic cases, we define

$$\Psi_k(\psi) := \forall w(Evalseq(\ulcorner\psi\urcorner, w) \rightarrow (Sat_k(\ulcorner\psi\urcorner, w) \leftrightarrow \psi[\beta(i_1, w), \ldots, \beta(i_m, w)])).$$

Every formula φ of length $\leq n$ is constructed from atomic formulas in at most n steps. Therefore, we would like to prove the following in $I\Delta_0 + \Omega_1$ by proofs of length linear in k:

1. $\Psi_{k-1}(\psi) \rightarrow \Psi_k(\neg\psi)$ for $Len(\ulcorner\neg\psi\urcorner) \leq k$;

2. $\Psi_{k-1}(\psi) \wedge \Psi_{k-1}(\chi) \to \Psi_k(\psi \to \chi)$ for $Len(\ulcorner \psi \to \chi \urcorner) \leq k$;

3. $\Psi_{k-1}(\psi) \to \Psi_k(\forall v_i \psi)$ for $Len(\ulcorner \forall v_i \psi \urcorner) \leq k$;

4. $\Psi_{k-1}(\psi) \to \Psi_k((\forall v_i \leq t)\psi)$ for $Len(\ulcorner (\forall v_i \leq t)\psi \urcorner) \leq k$.

If we can find these short proofs, then we have for every formula φ of length $\leq n$ a proof of $\Psi_n(\varphi)$ of length of the order of n^2, and we are done. We will leave the easy proofs of the four cases to the reader. QED

Lemma 3.3.12 $I\Delta_0 + \Omega_1$ *proves by a proof of length of the order of n^2 that Sat_n preserves the logical rules (Modus Ponens and Generalization) for formulas of length $\leq n$, i.e.*

$$I\Delta_0 + \Omega_1 \vdash Evalseq_n(y \ulcorner \to \urcorner z, w) \wedge Sat_n(y, w) \wedge Sat_n(y \ulcorner \to \urcorner z, w) \to Sat_n(z, w)$$

and

$$I\Delta_0 + \Omega_1 \vdash Evalseq_n(\ulcorner \forall v_i \urcorner y, w) \wedge \forall w'(w =_i w' \to Sat_n(y, w')) \to Sat_n(\ulcorner \forall v_i \urcorner y, w)$$

Proof. Immediately from Lemma 3.3.8. QED

Lemma 3.3.13 $I\Delta_0 + \Omega_1$ *proves by a proof of length of the order of n^2 that Sat_n preserves the logical axioms and the equality axioms for formulas of length $\leq n$, e.g. axiom scheme (1) of [WP 87]:*

PW1 $I\Delta_0 + \Omega_1 \vdash Evalseq_n(y \ulcorner \to (\urcorner z \ulcorner \to \urcorner y \ulcorner) \urcorner, w) \to Sat_n(y \ulcorner \to (\urcorner z \ulcorner \to \urcorner y \ulcorner) \urcorner, w)$

Similarly, the other propositional schemes (2) and (3) are preserved. Corresponding to axiom schemes (4), (5), and (6) we have the following:

PW4 *(Corresponding to axiom (4) of [WP 87])*

$$\begin{aligned} I\Delta_0 + \Omega_1 \vdash \quad & Evalseq_n(\ulcorner \forall v_i \urcorner y \to Sub(y, \ulcorner v_i \urcorner, t), w) \wedge SubOK(y, \ulcorner v_i \urcorner, t) \\ & \to Sat_n(\ulcorner \forall v_i \urcorner y \to Sub(y, \ulcorner v_i \urcorner, t), w), \end{aligned}$$

where $SubOK(y, \ulcorner v_i \urcorner, t)$ is Buss' formalization of "the term with Gödel number t is free for the variable v_i in the (term or) formula with Gödel number y".

PW5 *(Corresponding to axiom (5) of [WP 87])*

$$\begin{aligned} I\Delta_0 + \Omega_1 \vdash \quad & Evalseq_n(\ulcorner \forall v_i(\urcorner y \ulcorner \to \urcorner z \ulcorner) \to (\urcorner y \ulcorner \to \forall v_i \urcorner z \ulcorner) \urcorner, w) \\ & \wedge \text{“}v_i \textit{ does not appear free in the formula with Gödel nr. } y\text{”} \\ & \to Sat_n(\ulcorner \forall v_i(\urcorner y \ulcorner \to \urcorner z \ulcorner) \to (\urcorner y \ulcorner \to \forall v_i \urcorner z \ulcorner) \urcorner, w). \end{aligned}$$

PW 6 *(Corresponding to axiom (6) of [WP 87])*

$$I\Delta_0 + \Omega_1 \vdash Evalseq_n(v_1 \ulcorner = \urcorner v_1, w) \to Sat_n(v_1 \ulcorner = \urcorner v_1, w)$$

and

$$\begin{aligned} I\Delta_0 + \Omega_1 \vdash \quad & Evalseq_n(v_i \ulcorner = \urcorner v_j \ulcorner \to (\urcorner y \ulcorner \to \urcorner z \ulcorner) \urcorner, w) \\ & \wedge SubOK(y, \ulcorner v_i \urcorner, \ulcorner v_j \urcorner) \wedge Somesub(z, y, \ulcorner v_i \urcorner, \ulcorner v_j \urcorner) \\ & \to Sat_n(v_i \ulcorner = \urcorner v_j \ulcorner \to (\urcorner y \ulcorner \to \urcorner z \ulcorner) \urcorner, w), \end{aligned}$$

where $Somesub(z, y, \ulcorner v_i \urcorner, \ulcorner v_j \urcorner)$ is the formalization of "the formula with Gödel number z is the result of substituting the term v_j for some of the occurrences of v_i in the formula with Gödel number y".

Proof. For the propositional axiom schemes (PW1), (PW2) and (PW3), the results follow almost immediately from Lemma 3.3.8. For (PW4), we need proofs in $\mathrm{I}\Delta_0 + \Omega_1$ of length of the order of n^2 of the following "call by name = call by value" lemma:

$$\begin{aligned}&\mathit{Evalseq}_n(\ulcorner\forall v_i\urcorner y \to \mathit{Sub}(y,\ulcorner v_i\urcorner,t),w)\land \mathit{SubOK}(y,\ulcorner v_i\urcorner,t)\\ &\to \mathit{Sat}_n(\mathit{Sub}(y,\ulcorner v_i\urcorner,t),w)\leftrightarrow \mathit{Sat}_n(y,s(i,\,\mathit{Val}(t,w),w)).\end{aligned}$$

This can be proved by induction on n, in a way similar to the proof of Lemma 3.3.8. The rest of (PW4) then follows by Lemma 3.3.8 itself.

For (PW5), we need proofs in $\mathrm{I}\Delta_0 + \Omega_1$ of length of the order of n^2 of the following:

$$\begin{aligned}&\mathit{Evalseq}_n(\ulcorner\forall v_i(\urcorner y\ulcorner \to \urcorner z\ulcorner) \to (\urcorner y\ulcorner \to \forall v_i\urcorner z\ulcorner)\urcorner, w)\\ &\land \text{“}v_i \textit{ does not appear free in the formula with Gödel number } y\text{”} \land w =_i w'\\ &\to [\mathit{Sat}_n(y,w)\leftrightarrow \mathit{Sat}_n(y,w')].\end{aligned}$$

This can also be proved by induction on n; again, the rest of (PW5) follows by Lemma 3.3.8.

The first equality axiom of (PW6) is proved immediately by Lemma 3.3.8. The second one has a proof similar to that of (PW4). QED

Lemma 3.3.14 $\mathrm{I}\Delta_0+\Omega_1$ *proves by a proof of length of the order of* n^2 *that* Sat_n *preserves the basic non-logical axioms for formulas of length* $\leq n$*, e.g.*

$$\mathrm{I}\Delta_0 + \Omega_1 \vdash \mathit{Evalseq}_n(\ulcorner 0 \leq 0 \land \neg S0 \leq 0\urcorner, w) \to \mathit{Sat}_n(\ulcorner 0 \leq 0 \land \neg S0 \leq 0\urcorner, w).$$

Similarly for the other five basic axioms relating the symbols 0,S,$+$, $\cdot$ *and* $\leq$ *of the language.*

Proof. Immediately by Lemma 3.3.8 and Lemma 3.3.4. QED

Lemma 3.3.15 $\mathrm{I}\Delta_0+\Omega_1$ *proves by a proof of length of the order of* n^2 *that* $\mathit{Sat}_{n,\Delta}$ *agrees with* Sat_n *on* Δ_0*-formulas of length* $\leq n$*, i.e.*

$$\mathit{Evalseq}_{n,\Delta}(x,w) \to [\mathit{Sat}_{n,\Delta}(x,w)\leftrightarrow \mathit{Sat}_n(x,w)].$$

Proof. By induction on n as in the proof of Lemma 3.3.10. Here, we take

$$\Phi_n \ := \ \forall x\forall w(\mathit{Evalseq}_{n,\Delta}(x,w) \to (\mathit{Sat}_{n,\Delta}(x,w)\leftrightarrow \mathit{Sat}_n(x,w))).$$

As in Lemma 3.3.10, we use the fact that if $x = y\ulcorner \to \urcorner z$ and $\mathit{Fmla}_{n+1,\Delta}(x)$, then $\mathit{Fmla}_{n,\Delta}(y)$ and $\mathit{Fmla}_{n,\Delta}(z)$, etc. QED

Lemma 3.3.16 $\mathrm{I}\Delta_0+\Omega_1$ *proves by a proof of length of the order of* n^2 *that* Sat_n *preserves the* Δ_0*-induction axioms of length* $\leq n$*, i.e.*

$$\begin{aligned}&\mathit{Fmla}_{n,\Delta}(y)\land \mathit{Evalseq}_n(\mathit{Sub}(y,\ulcorner v_1\urcorner,0)\ulcorner \land \forall v_1(\urcorner y\ulcorner \to \urcorner\mathit{Sub}(y,\ulcorner v_1\urcorner,Sv_1)\ulcorner) \to \forall v_1\urcorner y, w)\\ &\quad\to \mathit{Sat}_n(\mathit{Sub}(y,\ulcorner v_1\urcorner,0)\ulcorner \land \forall v_1(\urcorner y\ulcorner \to \urcorner\mathit{Sub}(y,\ulcorner v_1\urcorner,Sv_1)\ulcorner) \to \forall v_1\urcorner y, w).\end{aligned}$$

Proof. We work in $I\Delta_0 + \Omega_1$ and assume

$$Fmla_{n,\Delta}(y) \wedge Evalseq_n(Sub(y, \ulcorner v_1 \urcorner, 0)^{\ulcorner} \wedge \forall v_1({}^{\urcorner}y^{\ulcorner} \rightarrow {}^{\urcorner}Sub(y, \ulcorner v_1 \urcorner, Sv_1)^{\ulcorner}) \rightarrow \forall v_1{}^{\urcorner}y, w).$$

Because Sat_n is a partial satisfaction predicate for formulas of length $\leq n$, we can, by a proof of length of the order of n^2, prove that the formula

$$Sat_n(Sub(y, \ulcorner v_1 \urcorner, 0)^{\ulcorner} \wedge \forall v_1({}^{\urcorner}y^{\ulcorner} \rightarrow {}^{\urcorner}Sub(y, \ulcorner v_1 \urcorner, Sv_1)^{\ulcorner}) \rightarrow \forall v_1{}^{\urcorner}y, w)$$

is equivalent to the following formula:

$$Sat_n(Sub(y, \ulcorner v_1 \urcorner, 0), w) \wedge \forall w'(w' =_1 w \rightarrow (Sat_n(y, w') \rightarrow Sat_n(Sub(y, \ulcorner v_1 \urcorner, Sv_1), w'))) \\ \rightarrow \forall w'(w' =_1 w \rightarrow Sat_n(y, w')).$$

This formula in turn is equivalent to:

$$Sat_n(Sub(y, \ulcorner v_1 \urcorner, 0), w) \wedge \forall x(Sat_n(y, s(1, x, w)) \rightarrow Sat_n(Sub(y, \ulcorner v_1 \urcorner, Sv_1), s(1, x, w))) \\ \rightarrow \forall x Sat_n(y, s(1, x, w)),$$

where $s(1, x, w)$ is as defined in Definition 3.3.5. This last formula is then, by a proof of length of the order of n^2 of a "call by name = call by value" lemma analogous to the one proved in Lemma 3.3.13, equivalent to the following formula:

$$Sat_n(y, s(1, 0, w)) \wedge \forall x(Sat_n(y, s(1, x, w)) \rightarrow Sat_n(y, s(1, Sx, w))) \\ \rightarrow \forall x Sat_n(y, s(1, x, w)).$$

This looks almost like an instance of induction. However, because Sat_n is not Δ_0,we replace it by its $\Delta_0(Val, \#, |\ |, \lfloor \frac{1}{2}x \rfloor)$-equivalent $Sat_{n,\Delta}$, as is allowed by Lemma 3.3.15 and the assumption $Fmla_{n,\Delta}(y)$, and we obtain the equivalent formula

$$Sat_{n,\Delta}(y, s(1, 0, w)) \wedge \forall x(Sat_{n,\Delta}(y, s(1, x, w)) \rightarrow Sat_{n,\Delta}(y, s(1, Sx, w))) \\ \rightarrow \forall x Sat_{n,\Delta}(y, s(1, x, w)).$$

As a true instance of $\Delta_0(Val, \#, |\ |, \lfloor \frac{1}{2}x \rfloor)$-induction, the formula above is at last provable from the assumptions. QED

Now that we have the partial truth predicates in hand, we can proceed with the proof proper of the main theorem of this chapter. We suppose that the reader is familiar with $I\Delta_0 + \Omega_1$-cuts and $I\Delta_0 + \Omega_1$-initial segments, and also with Solovay's method of shortening cuts (see definition 2.6.1, definition 2.6.2 and lemma 2.6.6).

We have the following:

Lemma 3.3.17 *If K is an $I\Delta_0 + \Omega_1$-initial segment, then*

$$I\Delta_0 + \Omega_1 \vdash \forall x Prov(\ulcorner K(\overline{x}) \urcorner),$$

where $\overline{x}$ stands for the "efficient numeral" based on the binary expansion of x.

Proof. See lemma 2.6.8. It is not difficult to see that the proofs of $K(\bar{x})$ are of length of the order $|x|^2$.

However, in the formalized context in which we will use the result, the length of the formula K and the length of the proof $p_1(K)$ of $\forall y(K(y) \rightarrow K(Sy))$ and the proof $p_2(K)$ of $\forall y(K(y) \rightarrow K(SS0 \cdot y))$ also play a part in the computation of the length of the total proof, thereby making the length of the total proof of the order $|x|^2 \cdot |K| + |p_1(K)| + |p_2(K)|$.

In fact, if we analyze the proof we find that

$$\mathrm{I}\Delta_0 + \Omega_1 \vdash \forall J \forall x(\Box(J \text{ "is an initial segment" }) \rightarrow \Box(J(\bar{x}))).$$

QED

Definition 3.3.18 We formally define the following:

$$\mathit{LPrf}_v(u, \ulcorner \chi \urcorner) \;:=\; \text{"}\textit{u codes a proof of } \chi \textit{in } \mathrm{I}\Delta_0 + \Omega_1 \textit{ involving only formulas of length } \leq v\text{"}.$$

Lemma 3.3.19 *The following is provable in* $\mathrm{I}\Delta_0 + \Omega_1$*:*

$$\forall x \mathit{Prov}(\ulcorner \forall y \leq \bar{x}(\mathit{Prf}(y, \overline{\ulcorner \varphi \urcorner}) \leftrightarrow \mathit{LPrf}_{|x|}(y, \overline{\ulcorner \varphi \urcorner}))\urcorner)$$

Proof. Formalize the following observation: if a formula v occurs in a proof y where $y \leq x$, then $\mathit{Len}(v) \leq |v| \leq |y| \leq |x|$. QED

Theorem 3.3.20 (Small reflection) *For all sentences φ the following holds:*

$$\mathrm{I}\Delta_0 + \Omega_1 \vdash \forall x \mathit{Prov}(\ulcorner \forall y \leq \bar{x}(\mathit{Prf}(y, \overline{\ulcorner \varphi \urcorner}) \rightarrow \varphi)\urcorner)$$

Proof. By Lemma 3.3.19, it suffices to prove

$$\mathrm{I}\Delta_0 + \Omega_1 \vdash \forall x \mathit{Prov}(\ulcorner \forall y \leq \bar{x}(\mathit{LPrf}_{|x|}(y, \overline{\ulcorner \varphi \urcorner}) \rightarrow \varphi)\urcorner).$$

We reason inside $\mathrm{I}\Delta_0 + \Omega_1$, and we take an x which we shall use to make a cut. The idea behind the proof is to find a Gödel number K_x standing for a formalized "*Prov*-initial segment" such that we have

$$\mathit{Prov}(K_x(\bar{x})^{\ulcorner} \rightarrow \forall y \leq \bar{x}(\mathit{LPrf}_{|x|}(y, \overline{\ulcorner \varphi \urcorner}) \rightarrow \varphi)\urcorner).$$

(By abuse of notation we write $K_x(\bar{x})$ for the Gödel number that results by the appropriate application of the substitution function to K_x). In the construction of the *Prov*-initial segment K_x, we will need the formalized versions of the lemmas which we proved above about the existence and the properties of partial satisfaction predicates for formulas of length smaller than some standard numeral n. In our formalized context, $|x|$ plays the rôle of "standard numeral", as will become clear when we define K_x. Again by abuse of notation, we let $\mathit{Sat}_{|x|}(v, w)$ stand for a Gödel number instead of a formula; we will use the appropriate formalizations of lemmas we proved about the formulas $\mathit{Sat}_n(v, w)$ to derive formalized facts about the Gödel number $\mathit{Sat}_{|x|}(v, w)$.

Keeping these cautionary remarks in mind, we start the proof by defining the Gödel number J_x of a formalized "*Prov*-cut" (later to be shortened to the *Prov*-initial segment K_x that we need) as follows:

$$J_x(s) \;:=\; \ulcorner \forall y, v \leq s(\mathit{LPrf}_{|x|}(y, v) \rightarrow \forall w(\mathit{Evalseq}(v, w) \rightarrow \urcorner \mathit{Sat}_{|x|}(v, w)^{\ulcorner}))\urcorner.$$

By the formalized version of Lemma 3.3.7, we may assume that this Gödel number exists, because the length of $\mathit{Sat}_{|x|}(v,w)$ is linear in $|x|$. (Note that we are reasoning inside $\mathrm{I}\Delta_0+\Omega_1$ all the time!) It is not difficult to prove directly from the definition of J_x (and from the fact that J_x is small enough) that the following holds:

$$\mathit{Prov}(J_x(\bar{0})^{\ulcorner} \wedge \forall y \forall z({}^{\urcorner}J_x(z)^{\ulcorner} \wedge y \leq z \rightarrow {}^{\urcorner}J_x(y)^{\ulcorner})^{\urcorner}).$$

To prove that J_x is closed under successor, we remark that

$$\mathit{Prov}(\ulcorner \mathit{LPrf}_{|x|}(y,v) \rightarrow \mathit{Len}(v) \leq |x| \urcorner).$$

Therefore, we can formalize Lemmas 3.3.12, 3.3.13, 3.3.14 and 3.3.16 to conclude by a proof of length of the order $|x|^2$ that $\mathit{Sat}_{|x|}(v,w)$ is preserved by all logical and non-logical axioms and rules for formulas of length $\leq |x|$, and thus indeed,

$$\mathit{Prov}(\ulcorner \forall y({}^{\urcorner}J_x(y)^{\ulcorner} \rightarrow {}^{\urcorner}J_x(Sy)^{\ulcorner})^{\urcorner}),$$

proving J_x to be a *Prov*-cut.

By a formalization of the proof of Lemma 2.6.6, we can shorten the *Prov*-cut J_x to a *Prov*-initial segment K_x of length linear in $|x|$. The proof that K_x is a *Prov*-initial segment is of length polynomial in $|x|$.

Carefully analyzing the proof of Lemma 3.3.17 (see the remark at the end of that proof), we find, by proofs of length polynomial in $|x|$,

$$\mathit{Prov}(K_x(\bar{x})) \wedge \mathit{Prov}(K_x(\overline{\ulcorner\varphi\urcorner})).$$

And thus, because we have $\mathit{Prov}(\ulcorner \forall y({}^{\urcorner}K_x(y)^{\ulcorner} \rightarrow {}^{\urcorner}J_x(y)^{\ulcorner})^{\urcorner})$, we conclude that, by definition of J_x, we have the following:

$$\mathit{Prov}(\ulcorner \forall y \leq \bar{x}(\mathit{LPrf}_{|x|}(y,\overline{\ulcorner\varphi\urcorner}) \rightarrow \forall w(\mathit{Evalseq}(\overline{\ulcorner\varphi\urcorner},w) \rightarrow {}^{\urcorner}\mathit{Sat}_{|x|}(\overline{\ulcorner\varphi\urcorner},w)^{\ulcorner}))^{\urcorner}).$$

Because we have $\mathit{Prov}(\ulcorner \forall y \leq \bar{x}(\mathit{LPrf}_{|x|}(y,\overline{\ulcorner\varphi\urcorner}) \rightarrow \mathit{Fmla}_{|x|}(\overline{\ulcorner\varphi\urcorner}))^{\urcorner})$, we can apply the formalized version of Lemma 3.3.11, taking note that φ is a sentence. Therefore,

$$\mathit{Prov}(\ulcorner \forall y \leq \bar{x}(\mathit{LPrf}_{|x|}(y,\overline{\ulcorner\varphi\urcorner}) \rightarrow \forall w(\mathit{Evalseq}(\overline{\ulcorner\varphi\urcorner},w) \rightarrow \varphi))^{\urcorner}).$$

This in turn is equivalent to the desired

$$\mathit{Prov}(\ulcorner \forall y \leq \bar{x}(\mathit{LPrf}_{|x|}(y,\overline{\ulcorner\varphi\urcorner}) \rightarrow \varphi)^{\urcorner}).$$

Stepping out of $\mathrm{I}\Delta_0+\Omega_1$ again, we conclude that indeed

$$\mathrm{I}\Delta_0+\Omega_1 \vdash \forall x \mathit{Prov}(\ulcorner \forall y \leq \bar{x}(\mathit{LPrf}_{|x|}(y,\overline{\ulcorner\varphi\urcorner}) \rightarrow \varphi)^{\urcorner}).$$

QED

Remark 3.3.21 Looking carefully at the proof of Theorem 3.3.20, we notice that it is also possible to derive the following result, which is a little bit stronger:

$$\mathrm{I}\Delta_0+\Omega_1 \vdash \forall v(\mathit{Sent}(v) \rightarrow \forall x \mathit{Prov}(\ulcorner \forall y \leq \bar{x}(\mathit{LPrf}_{|x|}(y,\overline{\ulcorner v\urcorner}) \rightarrow {}^{\urcorner}v^{\ulcorner})^{\urcorner}).$$

Theorem 3.3.20 and its proof can also be adapted for the case that φ is a formula instead of a sentence (or in the stronger result mentioned above: $\mathit{Fmla}(v)$ instead of $\mathit{Sent}(v)$).

Corollary 3.3.22 (Švejdar's principle is provable in $\mathrm{I}\Delta_0 + \Omega_1$) *For all sentences φ, ψ, we have the following:*

$$\mathrm{I}\Delta_0 + \Omega_1 \vdash \Box\varphi \to \Box(\Box\psi \leq \Box\varphi \to \psi),$$

i.e.

$$\mathrm{I}\Delta_0 + \Omega_1 \vdash \exists x Prf(x, \ulcorner\varphi\urcorner) \to Prov(\ulcorner \exists y(Prf(y, \overline{\ulcorner\psi\urcorner}) \wedge \forall z \leq y \neg Prf(z, \overline{\ulcorner\varphi\urcorner})) \to \psi\urcorner).$$

Proof. We work inside $\mathrm{I}\Delta_0{+}\Omega_1$ and suppose $Prf(x, \ulcorner\varphi\urcorner)$. By provable Σ_1^b-completeness, this implies

$$Prov(\ulcorner Prf(\dot{x}, \overline{\ulcorner\varphi\urcorner})\urcorner).$$

Hence, we have

$$Prov(\ulcorner \exists y(Prf(y, \overline{\ulcorner\psi\urcorner}) \wedge \forall z \leq y \neg Prf(z, \overline{\ulcorner\varphi\urcorner})) \to \exists y \leq \dot{x} Prf(y, \overline{\ulcorner\psi\urcorner})\urcorner).$$

Theorem 3.3.20 gives

$$Prov(\ulcorner \exists y \leq \dot{x} Prf(y, \overline{\ulcorner\psi\urcorner}) \to \psi\urcorner);$$

therefore, we have the following:

$$Prov(\ulcorner \exists y(Prf(y, \overline{\ulcorner\psi\urcorner}) \wedge \forall z \leq y \neg Prf(z, \overline{\ulcorner\varphi\urcorner})) \to \psi\urcorner).$$

Jumping outside $\mathrm{I}\Delta_0 + \Omega_1$ again, we conclude that

$$\mathrm{I}\Delta_0 + \Omega_1 \vdash \exists x Prf(x, \ulcorner\varphi\urcorner) \to Prov(\ulcorner \exists y(Prf(y, \overline{\ulcorner\psi\urcorner}) \wedge \forall z \leq y \neg Prf(z, \overline{\ulcorner\varphi\urcorner})) \to \psi\urcorner).$$

QED

Remark 3.3.23 Analogously to remark 3.3.21, we may strengthen Švejdar's principle to the following:

$$\mathrm{I}\Delta_0 + \Omega_1 \vdash Sent(u) \wedge Sent(v) \wedge Prov(u) \to Prov(\ulcorner Prov(v) \leq Prov(u) \to \urcorner v).$$

Švejdar introduced a modal system in order to study generalized Rosser sentences, and he derived the formalized version of Rosser's Theorem in it [Šv 83]. Because of Corollary 3.3.22, Švejdar's system is sound with respect to $\mathrm{I}\Delta_0{+}\Omega_1$, and Rosser's Theorem holds in $\mathrm{I}\Delta_0 + \Omega_1$.

Below, we use an argument similar to Švejdar's to derive a more general theorem. For the case of PA, this theorem has been proved by Montagna and Bernardi (see [JM 87]).

Theorem 3.3.24 (Montagna-Bernardi in $\mathrm{I}\Delta_0 + \Omega_1$) *For every function h which is Σ_1^b-definable in $\mathrm{I}\Delta_0 + \Omega_1$ and maps sentences to sentences, there is a sentence C such that*

$$\mathrm{I}\Delta_0 + \Omega_1 \vdash Prov(\ulcorner C\urcorner) \leftrightarrow Prov(h(\ulcorner C\urcorner)).$$

Proof. Define C by diagonalization such that

$$\mathrm{I}\Delta_0 + \Omega_1 \vdash C \leftrightarrow Prov(h(\ulcorner C \urcorner)) \leq Prov(\ulcorner C \urcorner).$$

Reason inside $\mathrm{I}\Delta_0 + \Omega_1$ and assume first that $Prov(\ulcorner C \urcorner)$. Then by definition,

$$Prov(\ulcorner Prov(h(\ulcorner C \urcorner)) \leq Prov(\ulcorner C \urcorner)\urcorner).$$

Meanwhile Corollary 3.3.22 gives

$$Prov(\ulcorner C \urcorner) \rightarrow Prov(\ulcorner Prov(h(\ulcorner C \urcorner)) \leq Prov(\ulcorner C \urcorner) \rightarrow h(\ulcorner C \urcorner)\urcorner).$$

Combined, these two yield $Prov(\ulcorner C \urcorner) \rightarrow Prov(h(\ulcorner C \urcorner))$.

For the other side, we assume that $Prov(h(\ulcorner C \urcorner))$. This implies $Prov(\ulcorner Prov(h(\ulcorner C \urcorner))\urcorner)$, and thus

$$Prov(\ulcorner Prov(h(\ulcorner C \urcorner)) \leq Prov(\ulcorner C \urcorner) \vee Prov(\ulcorner C \urcorner) \leq Prov(h(\ulcorner C \urcorner))\urcorner).$$

By definition of C, we derive

$$Prov(\ulcorner C \vee Prov(\ulcorner C \urcorner) \leq Prov(h(\ulcorner C \urcorner))\urcorner).$$

Now we apply Corollary 3.3.22 to conclude that, because

$$Prov(h(\ulcorner C \urcorner)) \rightarrow Prov(\ulcorner Prov(\ulcorner C \urcorner) \leq Prov(h(\ulcorner C \urcorner)) \rightarrow C \urcorner),$$

indeed $Prov(h(\ulcorner C \urcorner)) \rightarrow Prov(\ulcorner C \urcorner)$. QED

Note that the formalized version of Rosser's Theorem follows immediately from this construction. If we take R such that

$$\mathrm{I}\Delta_0 + \Omega_1 \vdash R \leftrightarrow Prov(\ulcorner \neg R \urcorner) \leq Prov(\ulcorner R \urcorner),$$

we derive $\mathrm{I}\Delta_0 + \Omega_1 \vdash Prov(\ulcorner R \urcorner) \leftrightarrow Prov(\ulcorner \neg R \urcorner)$, and thus $\mathrm{I}\Delta_0 + \Omega_1 \vdash Prov(\ulcorner R \urcorner) \rightarrow Prov(\ulcorner \bot \urcorner)$ and $\mathrm{I}\Delta_0 + \Omega_1 \vdash Prov(\ulcorner \neg R \urcorner) \rightarrow Prov(\ulcorner \bot \urcorner)$.

3.4 Injection of small (but not too small) inconsistency proofs

Using the small reflection principle, we can strengthen Hájek's, Solovay's and Krajíček and Pudlák's results on the injection of inconsistencies into models of $\mathrm{I}\Delta_0 + EXP$ [Há 83, So 89, KP 89]. Instead of only injecting an inconsistency proof, we also take care to respect a fair number of consistency statements. Moreover, we do not need full exponentiation in our original model.

We cannot immediately apply the lemmas of [KP 89], but the essential steps in our proof are the same as in their article. We first apply Pudlák's version of Gödel's Second Incompleteness Theorem (see [Pu 86, Theorem 3.6]) to show that we can indeed inject an inconsistency proof; then we use the Omitting Types Theorem to prevent extra elements from creeping into the lower part of the new model that contains our injected inconsistency proof.

Theorem 3.4.1 *Let* $\mathbf{T} \supseteq \mathbf{I\Delta_0} + \mathbf{\Omega_1}$ *be a* Σ_1^b*-axiomatized theory for which the small reflection principle (see Theorem 3.3.20) is provable in* $\mathrm{I}\Delta_0 + \Omega_1$. *Let* $Con_T(x)$ *be a formalization of the consistency of* $\mathbf{T}$ *up to proofs of length* x. *Let* $\mathcal{M}$ *be a nonstandard countable model of* $\mathrm{I}\Delta_0 + \Omega_1$. *Let* a, c *be nonstandard elements of* $\mathcal{M}$ *such that the following conditions hold:*

- $\exp(a^c) \in \mathcal{M}$,
- $\mathcal{M} \models Con_T(a^k)$ *for all* $k < \omega$.

Then there exists a countable model $\mathcal{K}$ *of* $\mathbf{T}$ *such that* $a \in \mathcal{K}$ *and*

1. $\mathcal{M} \restriction a = \mathcal{K} \restriction a$,
2. $\mathcal{M} \restriction \exp(a^k) \subseteq \mathcal{K}$ *for all* $k < \omega$,
3. $\mathcal{K} \models \neg Con_T(a^c)$,
4. $\mathcal{K} \models Con_T(a^k)$ *for all* $k < \omega$,
5. $\mathcal{K} \models 2^{a^c} \downarrow$.

Proof. Define $\mathcal{N} := \{x \in \mathcal{M} | x < \exp(a^k)$ *for some* $k < \omega\}$. Then $\exp(a^c) \in \mathcal{M} \setminus \mathcal{N}$, thus $\mathcal{M}$ is a proper end-extension of $\mathcal{N}$. Therefore, by Theorem 1 of [WP 89], $\mathcal{N} \models \mathrm{B}\Sigma_1$. (Remember that $\mathrm{B}\Sigma_1$ is $\mathrm{I}\Delta_0$ + the scheme $\forall t(\forall x < t \exists y \varphi(x,y) \rightarrow \exists a \forall x < t \exists y < a \varphi(x,y))$ for $\varphi \in \Sigma_1^0$.) Also, it is easy to see that $\mathcal{N} \models \Omega_1$.

On the other hand, one of our assumptions is that $\mathcal{M} \models Con_T(a^k)$ for all $k < \omega$. By Δ_0-overspill we conclude that there is a nonstandard $d < c$ in $\mathcal{M}$ such that $\mathcal{M} \models Con_T(a^d)$. Thus, by Theorem 3.6 of [Pu 86], there is a $k < \omega$ such that $\mathcal{M} \models Con_{T+\neg Con_T(a^d)}(a^{\frac{d}{k}})$, so certainly $\mathcal{M} \models Con_{T+\neg Con_T(a^c)}(a^{\frac{d}{k}})$. Indeed, because $\frac{d}{k}$ is nonstandard, we even have $\mathcal{N} \models Con(\mathbf{U})$, where $\mathbf{U} := \mathbf{T} + \neg Con_T(a^c)$.

At this point we need some definitions analogous to the ones in [KP 89]. Let $L(\mathcal{N})$ be the language of arithmetic expanded with domain constants for the elements of $\mathcal{N}$. We define a translation t from $L(\mathcal{N})$ to $\mathcal{N}$ by $t(A(a_1, \ldots, a_k)) := \ulcorner A(\overline{a_1}, \ldots, \overline{a_k}) \urcorner$, where $\bar{a}_i$ is the efficient numeral of a_i. We need one more definition:

$$\mathbf{U}^* := \{A(\vec{a}) \in L(\mathcal{N}) | \mathcal{N} \models Prov_U(t(A(\vec{a})))\}.$$

It is easy to show that $\mathbf{U}^*$ is closed under the rules of predicate logic; that $\mathbf{U} \subseteq \mathbf{U}^*$; and that $Diag(\mathcal{N}) \subseteq \mathbf{U}^*$. Also, because $\mathcal{N} \models Con(\mathbf{U})$, we can conclude that $\mathbf{U}^*$ is consistent.

Moreover, by the small reflection principle for $\mathrm{I}\Delta_0 + \Omega_1$, we have

$$\mathcal{N} \models \forall x Prov_U(\ulcorner Con_T(|\bar{x}|) \urcorner),$$

thus for all $k < \omega$, $Con_T(a^k) \in \mathbf{U}^*$.

Finally, using Solovay's cuts, we can show that $\mathcal{N} \models \forall x Prov_U(\ulcorner 2^x \downarrow \urcorner)$, thus $2^{a^c} \downarrow \in \mathbf{U}^*$.

We construct the required model $\mathcal{K}$ by the Omitting Types Theorem in order to take care that $\mathcal{K}$ will contain no new elements below a. Let τ be the type in $L(\mathcal{N})$ defined by

$$\tau(x) := \{x \leq a\} \cup \{x \neq b | b \in \mathcal{M} \restriction a\}.$$

Claim 1: $\mathbf{U}^*$ locally omits τ.

Proof. Take any $A(x)$, and suppose that for all $b \leq a$ in $\mathcal{N}$ we have $\mathbf{U}^* \vdash \neg A(b)$, and that $\mathbf{U}^* \vdash A(x) \rightarrow x \leq a$. We want to show that $\mathbf{U}^* \vdash \neg\exists x A(x)$. By definition of $\mathbf{U}^*$, it is sufficient to prove the following:

$$\mathcal{N} \models \forall b \leq a Prov_U(\ulcorner \neg A(\bar{b})\urcorner) \rightarrow Prov_U(\ulcorner \forall x \leq a \neg A(\bar{x})\urcorner).$$

So, suppose $\mathcal{N} \models \forall b \leq a Prov_U(\ulcorner \neg A(\bar{b})\urcorner)$. By $\mathrm{B}\Sigma_1$, there is a $q \in \mathcal{N}$ such that

$$\mathcal{N} \models \forall b \leq a \exists p < q Prf_U(p, \ulcorner \neg A(\bar{b})\urcorner).$$

Now we can use $\Delta_0(\omega_1)$-induction to show that we can combine these proofs for all $b \leq a$ into one proof p of $\forall x \leq a \neg A(x)$, where $|p| \leq a \cdot (|q| + k \cdot |a|) \leq a^m$ for some standard k, n, m, thus $p \in \mathcal{N}$. We conclude that indeed $\mathcal{N} \models Prov_U(\ulcorner \forall x \leq a \neg A(x)\urcorner)$. QED

At last we can construct a model $\mathcal{K}$ of $\mathbf{U}^*$ omitting τ. Using the facts that we proved about $\mathbf{U}^*$, we conclude that $\mathcal{K}$ satisfies all the properties that we want.

QED

In Theorem 3.4.1, we require that $\mathbf{T} \supseteq \mathrm{I}\Delta_0 + \Omega_1$ is a Σ_1^b-axiomatized theory for which the small reflection principle is provable in $\mathrm{I}\Delta_0 + \Omega_1$. Examples of such theories are finite extensions of $\mathrm{I}\Delta_0 + \Omega_1$ itself, $\mathrm{I}\Delta_0 + EXP$ and PA. We hope to give an exact characterization of theories amenable to methods analogous to those of section 3.3, [Pu 86] and [Pu 87] in a later paper.

Theorem 3.4.1 is only a slight extension of [KP 89, Theorem 2.1]. We use the small reflection principle only to show that the length of injected inconsistency proofs can be bounded from below as well as from above.

A variation on the proof of theorem 3.4.1 gives the following theorem. Its proof contains a more surprising use of the small reflection theorem than the proof of theorem 3.4.1: In theorem 3.4.3 we use it even in our application of the Omitting Types Theorem.

Recently, some papers (see [WP 89, Ad 90, Ad 93]) appeared that partially answer the end extension problem, which was formulated by Kirby and Paris in 1977 as follows: does every model of $\mathrm{I}\Delta_0 + \mathrm{B}\Sigma_1$ have a proper end extension to a model of $\mathrm{I}\Delta_0$? The theorem below gives a sufficient condition for a countable model of $\mathrm{I}\Delta_0 + \mathrm{B}\Sigma_1$ to have a proper end extension to a model of $\mathrm{I}\Delta_0$: if the model additionally satisfies $\Omega_1 + Con(I\Delta_0)$ and provable completeness for Π_2^b-formulas, then it does have such an end extension.

First we need a definition.

Definition 3.4.2 $\mathrm{C}\Pi_2^b(\mathbf{U})$ is the scheme

$$A(a_1, \ldots, a_k) \rightarrow Prov_U(\ulcorner A(\overline{a_1}, \ldots, \overline{a_k})\urcorner)$$

for $A(a_1, \ldots, a_k) \in \Pi_2^b$.

Theorem 3.4.3 *Let* $\mathbf{U} \supseteq \mathbf{Q}$ *be a* Σ_1^b*-axiomatized theory, and suppose* $\mathcal{N}$ *is a countable model of* $\mathrm{B}\Sigma_1 + \Omega_1 + \mathrm{C}\Pi_2^b(\mathbf{U}) + Con(\mathbf{U})$, *then there exists a countable model* $\mathcal{K}$ *of* $\mathbf{U}$ *such that* $\mathcal{K}$ *is an end-extension of* $\mathcal{N}$.

Proof. Define $\mathbf{U}^*$ from $\mathbf{U}, \mathcal{N}$ exactly as in the proof of Theorem 3.4.1. Again, we construct the required model $\mathcal{K}$ of $\mathbf{U}^*$ using the Omitting Types Theorem. This time, we define for all $a \in \mathcal{N}$ the type τ_a in $L(\mathcal{N})$ by:

$$\tau_a(x) := \{x \leq a\} \cup \{x \neq b | b \in \mathcal{M} \upharpoonright a\}.$$

Claim 2: $\mathbf{U}^*$ locally omits τ_a for all $a \in \mathcal{N}$.

Proof. Take any $a \in \mathcal{N}$ and any formula $A(x)$. As in the proof of Claim 1, it is sufficient to show the following:

$$\mathcal{N} \models \forall b \leq a Prov_U(\ulcorner \neg A(\bar{b}) \urcorner) \rightarrow Prov_U(\ulcorner \forall x \leq \bar{a} \neg A(\bar{x}) \urcorner).$$

So, suppose

$$\mathcal{N} \models \forall b \leq a Prov_U(\ulcorner \neg A(\bar{b}) \urcorner).$$

By $\mathrm{B}\Sigma_1$, there is a $q \in \mathcal{N}$ such that

$$\mathcal{N} \models \forall b \leq a \exists p < q Prf_U(p, \ulcorner \neg A(\bar{b}) \urcorner).$$

Now by $\mathrm{C}\Pi_2^b(\mathbf{U})$, we derive

$$\mathcal{N} \models \exists q Prov_U(\ulcorner \forall b \leq \bar{a} \exists p < \bar{q} Prf_U(p, \ulcorner \neg A(\bar{b}) \urcorner) \urcorner).$$

Therefore by the small reflection principle,

$$\mathcal{N} \models Prov_U(\ulcorner \forall b \leq \bar{a} \neg A(\bar{b}) \urcorner).$$

QED

We can now construct a countable model $\mathcal{K}$ of $\mathbf{U}^*$ omitting all τ_a for $a \in \mathcal{N}$. As before, it is easy to see that $\mathbf{U} \subseteq \mathbf{U}^*$ so $\mathcal{K} \models \mathbf{U}$.

By the way, note that by the small reflection principle for $\mathrm{I}\Delta_0 + \Omega_1$, or simply by the isomorphism, we have $Con_U(|\bar{x}|) \in \mathbf{U}^*$ and thus $\mathcal{K} \models Con_U(|\bar{x}|)$ for all $x \in \mathcal{N}$. QED

Chapter 4

Provable completeness for Σ_1-sentences implies something funny, even if it fails to smash the polynomial hierarchy

> But what's so blessed-fair that fears no blot?
> Thou mayst be false and yet I know it not.
>
> Shakespeare, *Sonnets*, no. 92

4.1 Introduction

In chapter 3, we proved that, if $NP \neq co\text{-}NP$, then Σ-completeness for witness comparison *formulas* is not provable in bounded arithmetic, i.e.

$$\begin{aligned} I\Delta_0 + \Omega_1 \nvdash \forall b \forall c \quad & (\exists a(Prf(a,c) \wedge \forall z \leq a \neg Prf(z,b)) \\ & \rightarrow Prov(\ulcorner \exists a(Prf(a,\bar{c}) \wedge \forall z \leq a \neg Prf(z,\bar{b}))\urcorner)). \end{aligned}$$

The above result does not give any information about Σ_1-*sentences.* If bounded arithmetic would prove completeness for Σ_1-sentences, then we could adapt Solovay's Completeness Theorem and prove that L is the provability logic of bounded arithmetic.

In this chapter we show that provable completeness for all Σ_1-sentences is unlikely. Unfortunately we have to work under an assumption (namely $P \neq NP \cap co\text{-}NP$) in which complexity theorists have less faith than in the assumption $NP \neq co\text{-}NP$ that we used in chapter 3.

4.2 If S^1_2 proves completeness for all Σ_1-sentences, then NP $\cap$ co-NP=P

Theorem 4.2.1 *Let $k \geq 1$. If $\Delta^p_k \neq \Sigma^p_k \cap \Pi^p_k$, then there is a sentence σ of the form $\exists x \varphi(x)$, where φ is a Π^b_k-formula, such that*

$$S^k_2 \nvdash \sigma \rightarrow Prov_{S^k_2}(\ulcorner \sigma \urcorner).$$

Proof. We prove the theorem for $k = 1$. For $k > 1$ the proofs are analogous.

Suppose that $P \neq NP \cap co - NP$. Let the Σ^b_1-formula $A(x)$ represent a predicate in $NP \cap co - NP$, but not in P. Thus there is a Σ^b_1-formula $B(x)$ that represents the complement of A. Now define $C(x, y) := (y = 0 \wedge A(x)) \vee (y = 1 \wedge B(x))$.

It is easy to see that $\omega \models \forall x \exists y \leq 1 C(x, y)$. Now let σ be the Σ_1 sentence defined by $\sigma := \exists x \forall y \leq 1 \neg C(x, y)$, and suppose that

$$S^1_2 \vdash \sigma \rightarrow Prov(\ulcorner \sigma \urcorner),$$

which is equivalent to

$$S^1_2 \vdash \forall x (\exists y \leq 1 C(x, y) \vee \exists y Prf(y, \ulcorner \sigma \urcorner)).$$

Next, by Buss' main theorem, we find a polynomial time function f such that

$$\omega \models \forall x (C(x, f(x)) \vee Prf(f(x), \ulcorner \sigma \urcorner)).$$

But σ, being a false sentence, is not provable in S^1_2, so we have actually

$$\omega \models \forall x C(x, f(x)).$$

This means that f is the characteristic function of A, hence A is in P, contrary to our assumption. QED

Chapter 5

On the provability logic of bounded arithmetic

С доказуемости мысом крайним–

(Марина Цветаева, Новогоднее, 1925)

Happy furthermost cape of provability–

(Marina Tsvetaeva, New Year's Poem, 1925,
translation Joseph Brodsky)

Abstract. Let PLΩ be the provability logic of $I\Delta_0 + \Omega_1$. We prove some containments of the form $L \subseteq$ PL$\Omega \subset Th(\mathcal{C})$ where L is the provability logic of PA and $\mathcal{C}$ is a suitable class of Kripke frames.

5.1 Introduction

In this chapter we develop techniques to build various sets of highly undecidable sentences in $I\Delta_0+\Omega_1$. Our results stem from an attempt to prove that the modal logic of provability in $I\Delta_0 + \Omega_1$, here called PLΩ, is the same as the modal logic L of provability in PA. It is already known that $L \subseteq$ PLΩ. We prove here some strict containments of the form PL$\Omega \subset Th(\mathcal{C})$ where $\mathcal{C}$ is a class of Kripke frames.

Stated informally the problem is whether the provability predicates of $I\Delta_0 + \Omega_1$ and PA share the same modal properties. It turns out that while $I\Delta_0 + \Omega_1$ certainly satisfies all the properties needed to carry out the proof of Gödel's second incompleteness theorem (namely $L \subseteq$ PLΩ), the question whether $L =$ PLΩ might depend on difficult issues of computational complexity. In fact if PL$\Omega \neq L$, it would follow that $I\Delta_0 + \Omega_1$ does not prove its completeness with respect to Σ^0_1-formulas, and a fortiori $I\Delta_0 + \Omega_1$ does not prove the Matijasevič–Robinson–Davis–Putnam theorem (every r.e. set is diophantine, see [Ma 70], [DPR 61]). On the other hand if $I\Delta_0 + \Omega_1$ did prove its completeness with respect to Σ^0_1-formulas, it would follow not only that $L =$ PLΩ, but also that $NP = co - NP$. The possibility remains that $L =$ PLΩ and that one could give a proof of this fact without making use of provable Σ^0_1-completeness in its full generality. Such a project is not without challenge due to the ubiquity of Σ^0_1-completeness in the whole area of provability logic.

We begin by giving the definitions of PLΩ. For the definitions of L and T-interpretation, we refer the reader to section 2.2.

Definition 5.1.1 Let PLΩ be the provability logic of $I\Delta_0 + \Omega_1$, i.e. PLΩ is the set of all those modal formulas A such that for all $I\Delta_0 + \Omega_1$-interpretations *, $I\Delta_0 + \Omega_1 \vdash A^*$.

It is easy to see that PLΩ is deductively closed (with respect to modus ponens and necessitation), so we can write PL$\Omega \vdash A$ for $A \in$ PLΩ. Our results arise from an attempt to answer the following:

Question 5.1.2 *Is PL$\Omega = L$? (Where we have identified L with the set of its theorems.)*

The soundness side of the question, namely $L \subseteq$ PLΩ, has already been answered positively. This depends on the fact that any reasonable Σ_1^b-axiomatized theory which is at least as strong as Buss' theory S_2^1 satisfies the derivability conditions needed to prove Gödel's incompleteness theorems (provided one uses efficient coding techniques and employs binary numerals). For the completeness side of the question, namely PL$\Omega \subseteq L$, we will investigate whether we can adapt Solovay's proof that L is the provability logic of PA.

We assume that the reader is familiar with the Kripke semantics for L (see definition 2.2.2) and with the method of Solovay's proof as described in [So 76]. In particular we need the following:

Theorem 5.1.3 *$L \vdash A$ iff A is forced at the root of every finite tree-like Kripke model. (It is easy to see that A will then be forced at every node of every finite tree-like Kripke model.)*

Solovay's method is the following: if $L \nvdash A$, then the countermodel $(K, \prec, \Vdash)$ provided by the above theorem is used to construct a PA-interpretation * for which $PA \nvdash A^*$.

The reason Solovay's proof is difficult to adapt to $I\Delta_0 + \Omega_1$ is that it is not known whether $I\Delta_0 + \Omega_1$ satisfies provable Σ_1^0-completeness (see definition 5.2.1) which is used in an essential way in Solovay's proof.

5.2 Arithmetical preliminaries

Definition 5.2.1 Let Γ be a set of formulas. We say that a (Σ_1^b-axiomatized) theory T satisfies *provable Γ-completeness*, if for every formula $\sigma(\vec{x}) \in \Gamma$, $T \vdash \sigma(x_1, \ldots, x_n) \rightarrow Prov_T(\ulcorner \sigma(\overline{x_1}, \ldots, \overline{x_n}) \urcorner)$.

It is known that PA, as well as any reasonable theory extending $I\Delta_0 +$ EXP, satisfies provable Σ_1^0-completeness.

De Jongh, Jumelet and Montagna [JMM 91] showed that Solovay's result can be extended to all reasonable Σ_1^0-sound theories T satisfying provable Σ_1^0-completeness. More precisely it is sufficient that the provability predicate of T provably satisfies the axioms of Guaspari and Solovay's modal witness comparison logic R^-. So Solovay's result holds for ZF, $I\Sigma_n$ and $I\Delta_0 +$ EXP.

On the other hand it is known that if $NP \neq co-NP$, then $I\Delta_0 + \Omega_1$ does not satisfy provable Σ_1^0-completeness or even provable Δ_0-completeness. In chapter 3 we proved that, if $NP \neq co-NP$, $I\Delta_0 + \Omega_1$ does not even satisfy provable completeness for the single

Σ_1^0-formula $\sigma(u,v) \equiv \exists x(Prf_{I\Delta_0+\Omega_1}(x,u) \wedge \forall y < x \neg Prf_{I\Delta_0+\Omega_1}(y,v))$. (See also chapter 4 for a related result.)

In view of the above difficulties, we try to do without Σ_1^0-completeness. In the rest of this section we state some results about $I\Delta_0+\Omega_1$ which in some cases allow us to dispense with the use of Σ_1^0-completeness. The following proposition is proved by [WP 87] (see also theorem 2.3.24):

Theorem 5.2.2 $I\Delta_0+\Omega_1$ *satisfies provable* Σ_1^b*-completeness.*

By abuse of notation we will denote by $\Box A$ both the arithmetization of the provability predicate of $I\Delta_0+\Omega_1$ and the corresponding modal operator. $\Diamond A$ is defined as $\neg\Box\neg A$ and $\Box^+ A$ as $\Box A \wedge A$. If $A(x)$ is an arithmetical formula, we will write $\forall x \Box(A(x))$ as an abbreviation for the arithmetical sentence which formalizes the fact that for all x there is a $I\Delta_0+\Omega_1$-proof of $A(\bar{x})$, where $\bar{x}$ is the binary numeral for x. If A and B are arithmetical sentences, $\Box A \leq \Box B$ denotes the witness comparison sentence

$$\exists x(Prf_{I\Delta_0+\Omega_1}(x, \ulcorner A \urcorner) \wedge \forall y < x \neg Prf_{I\Delta_0+\Omega_1}(y, \ulcorner B \urcorner)).$$

Similarly $\Box A < \Box B$ denotes

$$\exists x(Prf_{I\Delta_0+\Omega_1}(x, \ulcorner A \urcorner) \wedge \forall y \leq x \neg Prf_{I\Delta_0+\Omega_1}(y, \ulcorner B \urcorner)).$$

$\Box_k A$ is a formalization of the fact that A has a proof in $I\Delta_0+\Omega_1$ of Gödel number $\leq k$. So $\Box A < \Box B$ can be written as $\exists x(\Box_x A \wedge \neg\Box_x B)$. (Note that all the above definitions are only abbreviations for some arithmetical formulas and are not meant to correspond to an enrichment of the modal language.)

Remark 5.2.3 Since the proof predicate can be formalized by a Σ_1^b-formula, we have $I\Delta_0+\Omega_1 \vdash \Box A \rightarrow \Box\Box A$ and $I\Delta_0+\Omega_1 \vdash \Box_x A \rightarrow \Box\Box_x A$.

We suppose that the reader is familiar with $I\Delta_0+\Omega_1$-initial segments (see definition 2.6.2). Given an $I\Delta_0+\Omega_1$-initial segment I, $I\Delta_0+\Omega_1$ can formalize the fact that I defines a model of $I\Delta_0+\Omega_1$. It follows that for any arithmetical sentence θ we have:

Proposition 5.2.4 $I\Delta_0+\Omega_1 \vdash \Box(\theta) \rightarrow \Box(\theta^I)$, *where* θ^I *is obtained from* θ *by relativizing all the quantifiers to* I.

Note that if a Σ_1^0-formula is witnessed in an initial segment, then it is witnessed in the universe. Thus we have:

Remark 5.2.5 For every $I\Delta_0+\Omega_1$-initial segment I, and every Σ_1^0-formula $\sigma(x_1,\ldots,x_n)$, $I\Delta_0+\Omega_1 \vdash x_1 \in I \wedge \ldots \wedge x_n \in I \wedge \sigma^I(x_1,\ldots,x_n) \rightarrow \sigma(x_1,\ldots,x_n)$.

The use of binary numerals is essential for the following proposition (see lemma 2.6.8 and [Pu 86]):

Proposition 5.2.6 *For any* $I\Delta_0+\Omega_1$*-initial segment* I, $I\Delta_0+\Omega_1 \vdash \forall x \Box(x \in I)$.

Making use of an efficient truth predicate, we proved the following result in section 3.3:

Theorem 5.2.7 (Small reflection principle) $I\Delta_0+\Omega_1 \vdash \forall k \Box(\Box_k A \rightarrow A)$.

An immediate corollary is the following principle (originally stated by Švejdar for PA; see corollary 3.3.22 for a proof):

Corollary 5.2.8 (Švejdar's principle) $\mathrm{I}\Delta_0 + \Omega_1 \vdash \Box A \rightarrow \Box(\Box B \leq \Box A \rightarrow B)$.

Using Solovay's technique of shortening of cuts, it is easy to prove the following:

Proposition 5.2.9 *There is an* $\mathrm{I}\Delta_0+\Omega_1$*-initial segment* J*, such that for each* Σ_1^0*-formula* $\sigma(x_1, \ldots, x_n)$ *we have:* $\mathrm{I}\Delta_0 + \Omega_1 \vdash J(x_1) \wedge \ldots \wedge J(x_n) \wedge \sigma^J(x_1, \ldots, x_n) \rightarrow \Box\sigma(x_1, \ldots, x_n)$.

Proof. The proof is similar to the proof of provable Σ_1^b-completeness for $\mathrm{I}\Delta_0 + \Omega_1$ (see theorem 2.3.24 and [WP 87]). Therefore we only give a sketch of the proof. By induction on the structure of the formula, one can prove that for each Δ_0- formula A with free variables $x_1, \ldots, x_n$, there are k, l and m such that

$$\begin{aligned} \mathrm{I}\Delta_0 + \Omega_1 \vdash\ & \forall x \forall x_1, \ldots, x_n \leq x(y = exp(exp(|\ulcorner A \urcorner|^k \cdot |x|^l) + m) \\ & \wedge A(x_1, \ldots, x_n) \rightarrow \exists z \leq y Prf_{I\Delta_0+\Omega_1}(z, \ulcorner A(\bar{x}_1, \ldots, \bar{x}_2)\urcorner)). \end{aligned}$$

Now let J be the initial segment, which can be obtained by Solovay's shortening methods (cf. lemma 2.6.6, lemma 2.6.9), such that

- $\mathrm{I}\Delta_0 + \Omega_1 \vdash \forall x(J(x) \rightarrow \exists z(z = 2^x))$ and
- $\mathrm{I}\Delta_0 + \Omega_1 \vdash \forall x, y(J(x) \wedge J(y) \rightarrow J(x+y) \wedge J(x \cdot y) \wedge J(2^{|x| \cdot |y|}))$.

For this initial segment, we have for all Δ_0-formulas A,

$$\begin{aligned} \mathrm{I}\Delta_0 + \Omega_1 \vdash\ & \forall x_1, \ldots, x_n(J(x_1) \wedge \ldots \wedge J(x_n) \\ & \wedge A(x_1, \ldots, x_n) \rightarrow \exists z Prf_{I\Delta_0+\Omega_1}(z, \ulcorner A(\bar{x}_1, \ldots, \dot{x}_2)\urcorner)). \end{aligned}$$

The result immediately follows. QED

In the sequel of this chapter 'J' will always refer to the initial segment of proposition 5.2.9.

Corollary 5.2.10 *If* S_i $(i = 1, \ldots, k)$ *are* Σ_1^0*-sentences, then*

$$\mathrm{I}\Delta_0 + \Omega_1 \vdash \Box(\bigvee_i S_i) \rightarrow \Box(\bigvee_i \Box^+ S_i).$$

Proof. Let J be as in proposition 5.2.9. Work in $\mathrm{I}\Delta_0 + \Omega_1$ and suppose $\Box(\bigvee_i S_i)$ holds. Since J (provably) defines a model of $\mathrm{I}\Delta_0 + \Omega_1$, it follows $\Box(\bigvee_i S_i^J)$. By proposition 5.2.9 and remark 5.2.5 $\Box(S_i^J \rightarrow \Box^+ S_i)$ and the desired result follows. QED

The above corollary was originally proved by Albert Visser [Vi 91b] as a consequence of the following more general result:

Theorem 5.2.11 (Visser's principle) *If* S *and* $S_i (i = 1, \ldots, k)$ *are* Σ_1^0*-sentences, then* $\mathrm{I}\Delta_0 + \Omega_1 \vdash \Box(\bigwedge_i (S_i \rightarrow \Box S_i) \rightarrow S) \rightarrow \Box S$.

5.3 Trees of undecidable sentences

We will rephrase the problem of whether $\mathrm{PL}\Omega = L$ as a problem concerning the existence of suitable trees of undecidable senctences.

Let C be a class of finite tree-like strict partial orders. Without loss of generality we assume that for all $(K, \prec) \in C$, $K = \{1, \ldots, n\}$ for some $n \in \omega$, and 1 is the root (i.e. the least element of K). By $Th(C)$ we denote the set of all those modal formulas that are forced at the root of every Kripke model whose underlying tree belongs to C. Let $\preceq$ be the non-strict partial order associated to $\prec$.

Definition 5.3.1 Given a tree $(K, \prec)$ with root 1 and underlying set $K = \{1, \ldots, n\}$, we say that $(K, \prec)$ can be *embedded* (or *simulated*) in $\mathrm{I}\Delta_0 + \Omega_1$ if there are arithmetical sentences $L_1, \ldots, L_n$ (one for each node) such that, letting $\Box$ denote formalized provability from $\mathrm{I}\Delta_0 + \Omega_1$, the conjunction of the following sentences is consistent with $\mathrm{I}\Delta_0 + \Omega_1$:

1. L_1

2. $\Box^+(L_1 \vee \ldots \vee L_n)$

3. $\Box^+(L_i \to \neg L_j)$ for $i \neq j$ in K

4. $\Box^+(L_i \to \Diamond L_j)$ for $i \prec j$ in K

5. $\Box^+(L_i \to \Box \neg L_j)$ for $i \not\prec j$ in K

The following lemma is inspired by Solovay's proof of the fact that L is the provability logic of PA.

Lemma 5.3.2 *In order for $PL\Omega \subseteq Th(C)$ to be the case it suffices that every tree $(K, \prec) \in C$ can be embedded in* $\mathrm{I}\Delta_0 + \Omega_1$.

Proof. Suppose $A \notin Th(C)$. Then there is a Kripke model $(K, \prec, \Vdash)$ such that $(K, \prec) \in C$, $K = \{1, \ldots, n\}$, 1 is the least element of K, and $1 \Vdash \neg A$. By our hypothesis there exists a model M of $\mathrm{I}\Delta_0 + \Omega_1$ and sentences $L_1, \ldots, L_n$ satisfying, inside the model M, the conditions 1 - 5 of definition 5.3.1. Define a $\mathrm{I}\Delta_0 + \Omega_1$-interpretation * by setting, for every atomic propositional letter p, $p^* \equiv \bigvee_{i \Vdash p} L_i$. It is then easy to verify by induction on the complexity of the modal formula B, that for every $i \in K$:

- $i \Vdash B \Rightarrow M \models \Box^+(L_i \to B^*)$;
- $i \Vdash \neg B \Rightarrow M \models \Box^+(L_i \to \neg B^*)$.

The induction step for $\Box$ is based on condition 4 and the following consequence of conditions 2 and 5 of definition 5.3.1:

$$M \models \Box^+(L_i \to \Box(\bigvee_{j \succ i} L_j)).$$

Since $1 \Vdash \neg A$, it follows that $M \models \neg A^*$, hence $\mathrm{I}\Delta_0 + \Omega_1 \nvdash A^*$ as desired. QED

Corollary 5.3.3 *If every finite tree $(K, \prec)$ can be embedded in* $\mathrm{I}\Delta_0 + \Omega_1$, *then $PL\Omega = L$.*

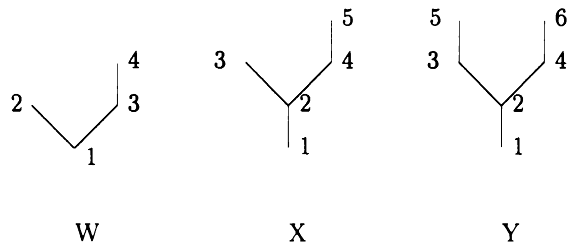


Figure 5.1: The trees W, X, Y

Proof. Let C be the class of all finite trees. If our hypothesis is satisfied, then $L \subseteq \mathrm{PL}\Omega \subseteq Th(C) = L$. QED

It can be easily verified that the sufficient condition of lemma 5.3.2 is also necessary. For, suppose that some tree $(K, \prec) \in C$ with root 1 and underlying set $K = \{1, \ldots, n\}$ cannot be embedded in $\mathrm{I}\Delta_0 + \Omega_1$. Then the negation of the conjunction of 1-5 (see definition 5.3.1) is easily seen to be in $\mathrm{PL}\Omega \setminus Th(C)$.

Thus $\mathrm{PL}\Omega \subseteq Th(C)$ iff every $(K, \prec) \in C$ can be embedded in $\mathrm{I}\Delta_0 + \Omega_1$. Hence a very natural question to ask is:

Question 5.3.4 *Which finite trees can be embedded in* $\mathrm{I}\Delta_0 + \Omega_1$*?*

Note that a complete answer to the above question, although interesting by itself, may not suffice to characterize $\mathrm{PL}\Omega$. In fact if C is the set of *all* finite trees that can be embedded in $\mathrm{I}\Delta_0 + \Omega_1$, we can in general only conclude $\mathrm{PL}\Omega \subseteq Th(C)$.

In order to describe the results proved in this chapter and in previous papers, we need to define what it means for a tree to omit another tree.

Definition 5.3.5 Let $(T_1, \prec_1)$ and $(T_2, \prec_2)$ be (strict) partial orders. An homomorphic embedding of $(T_1, \prec_1)$ into $(T_2, \prec_2)$ is an injective map $f : T_1 \rightarrow T_2$ such that for all $x, y \in T_1$, $x \prec_1 y \leftrightarrow f(x) \prec_2 f(y)$. If there is no homomorphic embedding of T_1 into T_2 we say that T_2 *omits* T_1.

If we try to adapt Solovay's proof to $\mathrm{I}\Delta_0 + \Omega_1$ in the most straightforward manner, the only trees that we can embed in $\mathrm{I}\Delta_0 + \Omega_1$ are the linear trees, namely trees omitting $(K, \prec)$ where $K = \{1, 2, 3\}$, $1 \prec 2$, $1 \prec 3$ and 2 is incomparable with 3.

A first improvement can be achieved using Švejdar's principle: let C_1 be the class of all trees that omit the tree $\mathbf{W} = (W, \prec)$, the least strict partial order with underlying set $W = \{1, 2, 3, 4\}$ such that $1 \prec 2$, $1 \prec 3 \prec 4$ (see Figure 1). In [Ve 88] it was proved that for trees in C_1 Solovay's proof can be adapted using Švejdar's principle. In other words, $\mathrm{PL}\Omega \subseteq Th(C_1)$. Moreover it was proved that the inclusion is a strict one.

In subsequent unpublished work I showed, using both Švejdar's and Visser's principles, that $\mathrm{PL}\Omega$ is included in the modal theory of C_2, the class of all trees of height ≤ 3.

A new improvement [BV 91] was achieved by analogous techniques but using a different definition of the Solovay constants. In this way it was proved that $\mathrm{PL}\Omega \subseteq Th(C_3)$, where C_3 is the class of all trees that omit the tree $\mathbf{X} = (X, \prec)$, the least strict partial order with underlying set $X = \{1, 2, 3, 4, 5\}$ such that $1 \prec 2 \prec 4 \prec 5$, $1 \prec 2 \prec 3$.

Finally in Section 5.4 of the present chapter, we improve these earlier results, by proving:

Theorem 5.3.6 *PL$\Omega \subseteq Th(C_4)$, where C_4 is the class of trees that omit the tree* $\mathbf{Y} = (Y, \prec)$*, the least strict partial order with underlying set $Y = \{1,2,3,4,5,6\}$ such that* $1 \prec 2 \prec 3 \prec 5$, $1 \prec 2 \prec 4 \prec 6$.

In particular, theorem 5.3.6 implies that we can embed $\mathbf{X}$. Note that the trees in C_4 can have an arbitrarily large number of bifurcation points, but each bifurcation point except the root can have at most one immediate successor which is not a leaf. The root can have any number of immediate successors which are not leaves.

On the other hand, we prove in Section 5.5 that for many classes C of trees (and especially for the classes $C_1, \dots, C_4$ defined above), we cannot have PL$\Omega = Th(C)$. Therefore, all inclusions mentioned above are strict. More precisely we prove that *if* PL$\Omega = Th(C)$, then every binary tree can be homomorphically embedded in some tree belonging to C. So it is unlikely that PLΩ is the theory of a class of trees, unless PL$\Omega = L$.

5.4 Upper bounds on PLΩ

Our task in this section will be to prove PL$\Omega \subseteq Th(C_4)$ using lemma 5.3.2.

Definition 5.4.1 Given $(K, \prec) \in C_4$, we say that $i \in K$ is a *special* node, iff i is a leaf, and some brother of i is not a leaf.

For example, in the tree $\mathbf{X}$ of Figure 1, the only special node is 3.

Definition 5.4.2 Let $(K, \prec) \in C_4$. Without loss of generality assume that
$K = \{1, \dots, n\}$ and 1 is the root. Let J be the initial segment of proposition 5.2.9. By a self-referential construction based on the diagonal lemma, we can simultaneously define sentences
$L_1, \dots, L_n$, and auxiliary functions v, w, S, such that the following holds:

1. If $i \in K$ is not special, let $w(i) = \mu x \Box_x \neg L_i$ (with the convention that $w(i) = \infty$ if $\Diamond L_i$); if $i \in K$ is special $w(i) = \mu x \in J \Box_x \neg L_i$ (with the convention that $w(i) = \infty$ if $\Diamond^J L_i$). We agree that ∞ is a specific element greater than any integer. Note that the definition of w can be formalized in $\mathrm{I}\Delta_0 + \Omega_1$.

2. If j is an immediate successor of i in $(K, \prec)$, let $v(i,j) = w(j)$; otherwise $v(i,j) = \infty$.

3. $S : K \longrightarrow K$ is defined as follows: $S(i) = i$ if for no $j \in K$ we have $v(i,j) < \infty$; otherwise among all the $j \in K$ with $v(i,j) < \infty$, pick one for which $v(i,j)$ is minimal, and set $S(i) = S(j)$. (Note that there exists at most one such j because if $w(j) = w(j') < \infty$, then there is one single proof of both $\neg L_j$ and $\neg L_{j'}$, so $j = j'$.)

4. $\mathrm{I}\Delta_0 + \Omega_1 \vdash L_i \leftrightarrow \Box \neg L_1 \wedge i = S(1)$.

The important point to observe, is that the definition of S can be formalized in $\mathrm{I}\Delta_0 + \Omega_1$ and that $\mathrm{I}\Delta_0 + \Omega_1$ proves that $S(1)$ is always defined. This depends on the fact that, although S is defined in a recursive way, to compute $S(1)$ one only needs a standard number of recursive calls, namely at most d where d is the height of the tree $(K, \prec)$ (in fact at each recursive call we climb one step up in the tree). Note also that S depends self-referentially on $L_1, \dots, L_n$. Finally note that, if a, b are distinct immediate successors of i, then the statement $v(i,a) < v(i,b)$ is equivalent to a witness comparison sentence in which some quantifiers are relativised to J. In particular, if a and b are not special, then $v(i,a) < v(i,b)$ is equivalent to the Σ^0_1-sentence $\Box \neg L_a < \Box \neg L_b$.

Remark 5.4.3 The main differences with Solovay's construction are the following: 1) we do not use an extra node 0 (but this is a minor point since we could define L_0 as $\Diamond L_1$). 2) In our construction we can only jump one step at a time, namely at each recursive call S we can only move from one point to some immediate successor (but this is also an inessential difference). 3) While Solovay employs a primitive recursive function from ω to K whose definition is not directly formalizable in $\mathrm{I}\Delta_0 + \Omega_1$, we use instead a function $S : K \rightarrow K$ which is provably total in $\mathrm{I}\Delta_0 + \Omega_1$. 4) We jump to a special node $i \in K$ only if we find a proof of $\neg L_i$ belonging to the initial segment J.

Given $(K, \prec)$ as above, we will show that $L_1, \ldots, L_n$ constitute an embedding of $(K, \prec)$ in $\mathrm{I}\Delta_0 + \Omega_1$. We need the following lemma.

Lemma 5.4.4 *Let $L_1, \ldots, L_n$ and $(K, \prec)$ be as in definition 5.4.2. Then:*

1. $\vdash \Box\neg L_1 \rightarrow L_1 \vee \ldots \vee L_n$.
2. $\vdash L_i \rightarrow \neg L_j$ *for* $i \neq j$ *in* K;
3. $\vdash L_i \rightarrow \Box\neg L_i$ *for* $i \in K$.
4. L_1 *is consistent with* $\mathrm{I}\Delta_0 + \Omega_1$;
5. *If* $j, j' \in K$ *are brothers, then* $\vdash \Box\neg L_j \leftrightarrow \Box\neg L_{j'}$.
6. $\vdash L_i \rightarrow \Diamond L_j$ *for* $i \prec j$ *in* K.
7. $\vdash L_j \rightarrow \Box\neg L_i$ *for* $i \prec j$ *in* K;
8. *If i is above (i.e.* $\succeq$*) a brother of j, then* $\vdash L_i \rightarrow \Box\neg L_j$; *if moreover j is a leaf, then* $\vdash L_j \rightarrow \Box\neg L_i$.
9. *Let* $j \succ 1$ *be an immediate successor of the root* 1. *Then* $\vdash L_1 \rightarrow \Box\Box(\neg L_j)$;
10. $\vdash L_1 \rightarrow \Box^+(L_i \rightarrow \Box\neg L_j)$ *whenever* i, j *are incomparable nodes of* K;

where ' $\vdash$ *' stands for '*$\mathrm{I}\Delta_0 + \Omega_1 \vdash$ *'.*

Proof. It will be clear from the context at which places we reason inside $\mathrm{I}\Delta_0 + \Omega_1$.

(1) and (2) are clear from the definition of the sentences L_i and the fact that $S : K \longrightarrow K$ is a total function.

(3). L_i implies that $\Box\neg L_1 \wedge i = S(1)$. If $i = 1$, $\Box\neg L_i$ follows immediately; otherwise we have $w(i) < \infty$, and therefore $\Box\neg L_i$.

(4). If L_1 is inconsistent with $\mathrm{I}\Delta_0 + \Omega_1$, then $\Box\neg L_1$ holds in the standard model, so by (1), one of the sentences L_i must hold in the standard model. This is absurd since each of these sentences implies its own inconsistency.

(5). First note that $\vdash \Box_x\neg L_j \rightarrow \Box(x \in J \wedge \Box_x\neg L_j)$. Thus, regardless of whether j is special or not, $\vdash \Box\neg L_j \rightarrow \Box(w(j) = \mu x \Box_x \neg L_j)$. Since j and j' are brothers, $\vdash L_{j'} \rightarrow$

$w(j') < w(j)$ (because $j' = S(1)$ implies $w(j') < w(j)$). Therefore $\vdash \Box\neg L_j \to \Box(L_{j'} \to \Box\neg L_{j'} < \Box\neg L_j)$. On the other hand by Švejdar's principle

$$\vdash \Box\neg L_j \to \Box(\Box\neg L_{j'} < \Box\neg L_j \to \neg L_{j'})$$

and we can conclude $\vdash \Box\neg L_j \to \Box\neg L_{j'}$.

(6). Löb's logic L proves $\Diamond A \wedge \Box(A \to \Diamond B) \to \Diamond B$. Hence by arithmetical soundness $\vdash \Diamond L_u \wedge \Box(L_u \to \Diamond L_v) \to \Diamond L_v$. It follows that in the proof of (6) we can assume without loss of generality that j is an immediate successor of i. Working inside $\mathrm{I}\Delta_0+\Omega_1$, assume L_i. Then $i = S(1)$. Hence $w(j) = \infty$. Now if j is not a special node, then $w(j) = \infty \leftrightarrow \Diamond L_j$ and we are done. If j is a special node, from $w(j) = \infty$ we can only conclude $\Diamond^J L_j$, so we need an additional argument. This is provided by point (5). In fact by definition of special node, i has certainly one immediate successor j' which is not special. Hence from L_i we can derive $\Diamond L_{j'}$ reasoning as above. By point (5), $\Diamond L_j \leftrightarrow \Diamond L_{j'}$ and we are done.

(7) can be derived through the chain of implications: $L_j \to \Box\neg L_j \to \Box\Box\neg L_j \to \Box\neg L_i$, where the last implication uses point (6).

(8). Let i be above a brother of j. Then by (5), (7) and (3) $\vdash L_i \to \Box\neg L_j$ as desired. To prove the second part, assume further that j is a leaf. We need to show $\vdash L_j \to \Box\neg L_i$. We can assume that i is *strictly* above a brother j' of j (for if i itself is a brother of j the desired result follows from (3) and (5)). But then j must be a special node, and therefore $w(j) = \mu x \in J\Box_x\neg L_j$. So $w(j) < w(j')$ is equivalent to a Σ^0_1-formula relativized to J, namely $w(j) < w(j') \leftrightarrow \exists x \in J(Prf_{I\Delta_0+\Omega_1}(x, \ulcorner\neg L_j\urcorner) \wedge \forall y \leq x \neg Prf_{I\Delta_0+\Omega_1}(y, \ulcorner\neg L_{j'}\urcorner))$. Thus by the properties of the initial segment J (and by theorem 5.2.6),

$$\vdash w(j) < w(j') \to \Box w(j) < w(j').$$

Now the desired result follows by observing that $\vdash L_j \to w(j) < w(j')$ (as $\vdash j = S(1) \to w(j) < w(j')$) and $\vdash L_i \to w(j') < w(j)$.

(9). By (1) and (3), $\vdash L_1 \to \Box(\bigvee_{i \succ 1} L_i)$. So to prove $\vdash L_1 \to \Box\Box\neg L_j$, it suffices to show that for each $i \succ 1$ we have $\vdash \Box(L_i \to \Box\neg L_j)$. This follows from (8) , (3) and (7).

(10). If the incomparable nodes i and j are in one of the situations covered by point (8), then $\vdash L_i \to \Box\neg L_j$, and a fortiori $\vdash L_1 \to \Box^+(L_i \to \Box\neg L_j)$ as desired. Since $(K, \prec)$ omits $\mathbf{Y}$, (8) can always be applied except when the largest node (with respect to $\preceq$) below i and j is 1 (the root). So assume that this is the case. By (2), we have $\vdash L_1 \to (L_i \to \Box\neg L_j)$. In order to show that also $\vdash L_1 \to \Box(L_i \to \Box\neg L_j)$, we will take in account the properties of the initial segment J (see proposition 5.2.9). Let i', j' be the least nodes with $1 \prec i' \preceq i$ and $1 \prec j' \preceq j$. So i' and j' are brothers. It follows from (9) that $\vdash L_1 \to \Box(\Box\neg L_{i'})$. Therefore, by proposition 5.2.4, $\vdash L_1 \to \Box(\Box^J\neg L_{i'})$. In the presence of $\Box^J\neg L_{i'}$, the sentence $w(i') < w(j')$ is equivalent to a Σ^0_1-sentence relativized to J. Therefore, by proposition 5.2.9, $\vdash L_1 \to \Box(w(i') < w(j') \to \Box(w(i') < w(j')))$. The desired result now follows from the fact that L_i provably implies $i = S(1)$ which entails $w(i') < w(j')$, while L_j provably implies $w(j') < w(i')$. QED

Corollary 5.4.5 *If $(K, \prec)$ and $L_1, \ldots, L_n$ are as above, then the conjunction of the following sentences is consistent with $\mathrm{I}\Delta_0 + \Omega_1$:*

1. L_1

2. $\Box^+(L_1 \vee \ldots \vee L_n)$.

3. $\Box^+(L_i \to \neg L_j)$ *for* $i \neq j$ *in* K.

4. $\Box^+(L_i \to \Diamond L_j)$ *for* $i \prec j$ *in* K.

5. $\Box^+(L_i \to \Box\neg L_j)$ *for* $i \not\prec j$ *in* K.

Proof. By (1) and (3) of lemma 5.4.4, and noting that $\vdash \Box\neg L_1 \to \Box\Box\neg L_1$, we derive $\vdash L_1 \to \Box^+(L_1 \vee \ldots \vee L_n)$. Next, (3) implies that $\vdash L_1 \to \Box^+(L_i \to \neg L_j)$ for $i \neq j$ in K. By (6) we have $L_1 \to \Box^+(L_i \to \Diamond L_j)$ for $i \prec j$ in K. Finally the corollary follows by (10) and (4).

Note that the derivation of corollary 5.4.5 from lemma 5.4.4 follows from a straightforward argument which can even be formalized in the decidable theory L^ω. (The axioms of L^ω are all the theorems of L and all the instances of $\Box A \to A$. The only rule is modus ponens.) QED

We have thus shown that every tree of C_4 can be embedded in $\mathrm{I}\Delta_0 + \Omega_1$. Thus:

Theorem 5.4.6 *PLΩ* $\subseteq Th(C_4)$.

5.5 Disjunction property

In this section we prove the following:

Theorem 5.5.1 *If PLΩ = $Th(C)$, where C is a class of finite trees, then every binary tree can be homomorphically embedded in some tree belonging to C.*

In particular, since the binary tree **Y** cannot be embedded in any member of C_4, it will follow that the inclusion PLΩ $\subseteq Th(C_4)$ is strict.

We will use the fact that PLΩ has the 'disjunction property' as proved by Franco Montagna (personal communication).

Definition 5.5.2 A modal theory P has the disjunction property if for every pair of modal sentences A and B, if $P \vdash \Box A \vee \Box B$, then $P \vdash A$ or $P \vdash B$.

It is known that L has the disjunction property.

Theorem 5.5.3 (Montagna) *PLΩ has the disjunction property.*

Proof. Suppose that for some $\mathrm{I}\Delta_0+\Omega_1$-interpretations $^\circ$ and $^\bullet$ we have $\mathrm{I}\Delta_0+\Omega_1 \nvdash A(\vec{p}^\circ)$ and $\mathrm{I}\Delta_0+\Omega_1 \nvdash B(\vec{p}^\bullet)$, where $\vec{p}$ contains all propositional variables occurring in the modal formulas A and B. We have to prove that there is an $\mathrm{I}\Delta_0+\Omega_1$- interpretation * such that $\mathrm{I}\Delta_0 + \Omega_1 \nvdash (\Box A \vee \Box B)^*$

By multiple diagonalization, define for all $p_i \in \vec{p}$ an arithmetical formula p_i^* such that

$$\mathrm{I}\Delta_0 + \Omega_1 \vdash p_i^* \leftrightarrow (\Box A(\vec{p}^*) \leq \Box B(\vec{p}^*) \wedge p_i^\circ) \vee (\Box B(\vec{p}^*) < \Box A(\vec{p}^*) \wedge p_i^\bullet).$$

We will show that $\mathrm{I}\Delta_0 + \Omega_1 \nvdash (\Box A \vee \Box B)^*$. So suppose, to derive a contradiction, that $\mathrm{I}\Delta_0 + \Omega_1 \vdash \Box A(\vec{p}^*) \vee \Box B(\vec{p}^*)$. Then

$$\mathrm{I}\Delta_0 + \Omega_1 \vdash \Box A(\vec{p}^*) \leq \Box B(\vec{p}^*) \vee \Box B(\vec{p}^*) < \Box A(\vec{p}^*).$$

Thus, because $\mathrm{I}\Delta_0 + \Omega_1$ is a true theory, either

1. $\Box A(\vec{p}^*) \leq \Box B(\vec{p}^*)$ and $\mathrm{I}\Delta_0 + \Omega_1 \vdash p_i^* \leftrightarrow p_i^\circ$ for all i (by definition of $\vec{p}^*$), or

2. $\Box B(\vec{p}^*) < \Box A(\vec{p}^*)$ and $\mathrm{I}\Delta_0 + \Omega_1 \vdash p_i^* \leftrightarrow p_i^\bullet$ for all i.

In case 1, we have $\mathrm{I}\Delta_0 + \Omega_1 \vdash A(\vec{p}^*)$, so $\mathrm{I}\Delta_0 + \Omega_1 \vdash A(\vec{p}^\circ)$, contradicting our assumption. Similarly, case 2 contradicts the assumption $\mathrm{I}\Delta_0 + \Omega_1 \nvdash B(\vec{p}^\bullet)$.

QED

In order to prove theorem 5.5.1 we need the following definition.

Definition 5.5.4 We define D_n by induction.

- $D_0 = \top$

- $D_{i+1}(\vec{p}, r) = \Diamond(D_i(\vec{p}) \wedge \Box^+ r) \wedge \Diamond(D_i(\vec{p}) \wedge \Box^+ \neg r)$, where $\vec{p}$ is of length i, and all propositional variables in $(\vec{p}, r)$ are different.

The main property of the formulas D_n is expressed by the following lemma.

Lemma 5.5.5 *If* $\mathbf{K}$ *is a finite tree-like Kripke model with root* k *such that* $k \Vdash D_n$, *then we can homomorphically embed (see definition 5.3.5) the full binary tree* $\mathbf{T}_n$ *of height* n *(and* $2^{n+1} - 1$ *nodes) into* $\mathbf{K}$.

Proof. By induction on n.

Base case. Trivial: $\mathbf{T}_0$ contains only one point.

Induction step. Suppose that $k \Vdash D_{i+1}(\vec{p}, r)$, i.e.

$$k \Vdash \Diamond(D_i(\vec{p}) \wedge \Box^+ r) \wedge \Diamond(D_i(\vec{p}) \wedge \Box^+ \neg r).$$

Then there are nodes k_1, k_2 such that $k \preceq k_1, k \preceq k_2, k_1 \Vdash D_i(\vec{p}) \wedge \Box^+ r$ and $k_2 \Vdash D_i(\vec{p}) \wedge \Box^+ \neg r$. By the induction hypothesis, we can homomorphically embed a copy of the full binary tree $\mathbf{T}_i$ of bifurcation depth i into the subtree of $\mathbf{K}$ that consists of all points $\succeq k_1$. Analogously, we can homomorphically embed a copy of $\mathbf{T}_i$ into the subtree of $\mathbf{K}$ of points $\succeq k_2$.

Because $k_1 \Vdash \Box^+ r$ and $k_2 \Vdash \Box^+ \neg r$, we may conclude that k_1 and k_2 are incomparable and that the two images of $\mathbf{T}_i$ are disjoint. Therefore, we can combine both homomorphic embeddings into one and subsequently map the root of $\mathbf{T}_{i+1}$ to k. Thus an homomorphic embedding of $\mathbf{T}_{i+1}$ into $\mathbf{K}$ is produced.

QED

Theorem 5.5.1 is now an immediate consequence of the following:

Theorem 5.5.6 *Let* C *be a class of finite trees such that* $Th(C)$ *has the disjunction property. Then for every* n, $Th(C) \cup \{D_n\}$ *is consistent. Therefore every binary tree (thus every tree) can be homomorphically embedded in some member of* C.

Proof. Let $P = Th(C)$. Note that $P \supseteq L$. We prove by induction on n that $P \cup \{D_n\}$ is consistent.

Base case. Trivial.

Induction step. Suppose as induction hypothesis that for $\vec{p}$ consisting of i different propositional variables, $P \cup \{D_i(\vec{p})\}$ is consistent. In order to derive a contradiction, suppose that $P \vdash \neg D_{i+1}(\vec{p}, r)$, that is

$$P \vdash \Box(\Box^+ r \rightarrow \neg D_i(\vec{p})) \vee \Box(\Box^+ \neg r \rightarrow \neg D_i(\vec{p})).$$

Then by the disjunction property, either

1. $P \vdash \Box^+ r \rightarrow \neg D_i(\vec{p})$ or

2. $P \vdash \Box^+ \neg r \rightarrow \neg D_i(\vec{p})$.

We show that 1 cannot hold. By the induction hypothesis, $P \nvdash \neg D_i(\vec{p})$. Since r does not appear in $D_i(\vec{p})$, we can take $r = \top$. But then $P \vdash \Box^+ r$, so $P \nvdash \Box^+ r \rightarrow \neg D_i(\vec{p})$.

By an analogous proof, we can show that 2 cannot hold, which gives the desired contradiction.

QED

Note that in the proof of the fact that $Th(\mathcal{C}) \cup \{D_n\}$ is consistent we have only used the fact that $Th(\mathcal{C})$ is a consistent modal theory extending L and satisfying the disjunction property. The same proof can therefore be applied to PLΩ, yielding:

Proposition 5.5.7 *PL$\Omega \cup \{D_n\}$ is consistent.*

Remark 5.5.8 For a strengthening of proposition 5.5.7 due to Berarducci, we refer the reader to [BV 93].

We are now able to strengthen theorem 5.5.1 as follows:

Theorem 5.5.9 *If there exists a binary tree H which cannot be homomorphically embedded in any member of $\mathcal{C}$, then $Th(\mathcal{C}) \not\subseteq PL\Omega$.*

Proof. Under our assumption there is some n such that the full binary tree of height n cannot be embedded in any member of $\mathcal{C}$. Hence $Th(\mathcal{C}) \cup \{D_n\}$ is inconsistent. On the other hand PL$\Omega \cup \{D_n\}$ is consistent. QED

Part III

Metamathematics for Peano Arithmetic

Chapter 6

Feasible interpretability

Sometimes we see a cloud that's dragonish:
A vapour sometime like a bear or lion,
A tower'd citadel, a pendent rock,
A forked mountain, or blue promontory
With trees upon't, that nod unto the world
And mock our eyes with air...

Shakespeare, *Anthony and Cleopatra*

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

> for some interpretation, U proves the interpretations of all axioms of V by proofs of length polynomial in the length of those axioms.

Here both U and V are Σ_1^b-axiomatized theories.

Many interpretations encountered in everyday mathematics (e.g. the interpretation of $ZF + \mathbf{V} = \mathbf{L}$ 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, $A \rhd B$ is translated as $PA + A^* \rhd_f PA + B^*$, 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.

6.1 Introduction

In this chapter, we investigate a novel concept of interpretability – we call it feasible interpretability – in which the complexity of proofs associated to the interpretation is bounded.

The concept was invented by Albert Visser, who called it effective interpretability in his paper [Vi b].

In order to define this concept, we slightly change the usual definition of interpretability (see section 2.4). First we give a half-formal definition of $U \rhd_f V$ (pronounced as "U feasibly interprets V":

$$U \rhd_f V \leftrightarrow \exists K \exists P(\text{“}K\textit{is an interpretation and }P\textit{is a polynomial”} \wedge \\ \forall a(\alpha_V(a) \rightarrow \exists p(\text{“}|p| \leq P(|a|)\text{”} \wedge Prf_U(p, a^K)))) \tag{6.1}$$

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 in definition 2.3.2. 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 (6.1).

However, it is clear that the formula $U \rhd_f V$ is Σ^0_2. As we know that, for reasonable theories U extending PA, $\{A \mid U \rhd U + A\}$ is a Π^0_2-complete predicate, it would be interesting to find out whether $\{A \mid U \rhd_f U + A\}$ is Σ^0_2-complete. Chapter 7 provides a positive answer to this question.

In [Vi b], 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 [Ber 90] could be adapted to show that ILM is the modal interpretability logic for feasible interpretability over PA^*.

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

The rest of the chapter is organized as follows. In section 6.2, 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 6.3 and section 6.5 that ILM is exactly the modal interpretability logic for feasible interpretability over PA. Section 6.4, 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.

6.2 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.

> $U \rhd_f V$ 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 φ of V, there is a proof of length $\leq P(|\varphi|)$ in U of φ^K

Here $|\varphi|$ denotes the length of φ. 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 will prove an easy lemma which can be used to show that many well-known interpretations are feasible. First we state some conditions on the length of formulas and proofs.

Remark 6.2.1 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 of the Gödel number of φ". In all examples in the rest of the chapter, the length measure is polynomially related to the length of the binary expression of the Gödel number.

In general, 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 6.2.2 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 $\circ \in \{\wedge, \vee, \rightarrow, \leftrightarrow\}$,
3. $|Qx\psi| \geq |\psi| + 1$ for $Q \in \{\forall, \exists\}$, and
4. for all formulas φ, $|\varphi| \geq 2$.

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 $\wedge$-rules and Modus Ponens do not make the proofs explode to gargantuan proportions; e.g. we suppose that we do not use a tableau system or a sequent calculus. A sufficient condition is the following.

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

1. if $|A| \leq |B|$, and we have a proof of length l_A of the formula A, and a proof of length $l_{A\rightarrow B}$ of $A \rightarrow B$, then there is a proof of length $\leq l_A + l_{A\rightarrow B} + c \cdot |B|$ of the formula B; and
2. if $|A|, |B| \leq |C|$, and we have a proof of length l_A of A, a proof of length l_B of B and a proof of length $l_{A\wedge B\rightarrow C}$ of $A \wedge B \rightarrow C$, then we have a proof of length $\leq l_A + l_B + l_{A\wedge B\rightarrow C} + c \cdot |C|$ of the formula C.

Lemma 6.2.2 *Let L be a language and U a theory satisfying the conditions in Remark 6.2.1. Let F be a function from L into L_U such that*

> *there is a polynomial P such that for all $\varphi \in L$, $|F(\varphi)| \leq P(|\varphi|)$. (This is always the case when L is finite.)*

Moreover, suppose that the following four conditions hold:

1. *$U \overset{P(|\varphi|)}{\vdash} F(\varphi)$ for all atomic $\varphi \in L$;*
2. *$U \overset{P(|\neg\psi|)}{\vdash} F(\psi) \rightarrow F(\neg\psi)$ for all $\psi \in L$;*

3. $U \vdash^{P(|\psi \circ \chi|)} F(\psi) \wedge F(\chi) \to F(\psi \circ \chi)$ for all $\psi, \chi \in L$ and $\circ \in \{\wedge, \vee, \to, \leftrightarrow\}$;

4. $U \vdash^{P(|Qx\psi|)} F(\psi) \to F(Qx\psi)$ for all $\psi \in L$ and $Q \in \{\forall, \exists\}$.

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

Proof. We do not need to find the smallest possible polynomial bound R, which makes the proof quite simple. Take a constant $d \geq 2$ such that

1. for all $n \geq 2$, $P(n) \leq n^d$ and
2. for all $\varphi \in L$, $c \cdot |F(\varphi)| \leq |\varphi|^d$, where c is as in Remark 6.2.1 in the condition on the length of proofs.

Define the polynomial $R(n) := n^{d+2}$. We will prove by induction on the construction of φ 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 φ, $U \vdash F(\varphi)$ by a proof of length $\leq P(|\varphi|)$. But by definition of d, $P(|\varphi|) \leq |\varphi|^d \leq |\varphi|^{d+2}$.

¬-step Suppose as induction hypothesis that $U \vdash F(\psi)$ by a proof of length $\leq |\psi|^{d+2}$. By assumption, $U \vdash F(\psi) \to F(\neg\psi)$ 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 6.2.1, we have $U \vdash F(\neg\psi)$ by a proof of length $\leq |\psi|^{d+2} + |\neg\psi|^d + c \cdot |F(\neg\psi)| \leq |\psi|^{d+2} + |\neg\psi|^d + |\neg\psi|^d$ (where the last inequality holds by clause 2 of the definition of d). Since we assume that $|\neg\psi| \geq |\psi| + 1$, we have $|\psi|^{d+2} + |\neg\psi|^d + |\neg\psi|^d \leq |\neg\psi|^{d+2}$ by an easy computation using the binomial theorem and the fact that $d \geq 2$.

Connective step Let $\circ \in \{\wedge, \vee, \to, \leftrightarrow\}$. Suppose as induction hypothesis that $U \vdash F(\psi)$ by a proof of length $\leq |\psi|^{d+2}$, and $U \vdash F(\chi)$ by a proof of length $\leq |\chi|^{d+2}$. By assumption, $U \vdash F(\psi) \wedge F(\chi) \to F(\psi \circ \chi)$ 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 6.2.1 now implies that $U \vdash F(\psi \circ \chi)$ by a proof of length $\leq |\psi|^{d+2} + |\chi|^{d+2} + |\psi \circ \chi|^d + c \cdot |F(\psi \circ \chi)| \leq |\psi|^{d+2} + |\chi|^{d+2} + |\psi \circ \chi|^d + |\psi \circ \chi|^d$ (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 $|\psi|^{d+2} + |\chi|^{d+2} + |\psi \circ \chi|^d + |\psi \circ \chi|^d \leq |\psi \circ \chi|^{d+2}$, as desired.

Quantifier step The quantifier steps are analogous to the ¬-step, so we leave them to the reader.

QED

Remark 6.2.3 For applications of lemma 6.2.2, we usually take $F(\psi)$ to be a schema involving ψ and ψ^K, where K is an interpretation. When we want to prove that some interpretation K of V into U is feasible, we can often use lemma 6.2.2 in the following way. Suppose all but a finite number of axioms of V have the form $\Phi(\psi)$, where Φ is a formula scheme. The feature we need in order to apply lemma 6.2.2 is the fact that both $|\Phi(\psi)|$ and $|\psi^K|$ are polynomial in $|\psi|$.

An adapted version of lemma 6.2.2 works in case attention is restricted to Δ_0-formulas. An example of a function F for which the Δ_0-version of lemma 6.2.2 could be applied is the scheme $F(\psi) := \psi \leftrightarrow \psi^K$ for $\psi \in \Delta_0$, where K is a fixed interpretation such that $|\psi^K|$ is polynomial in $|\psi|$.

As a first example, in which we do not yet need lemma 6.2.2, we will show that the usual interpretation of $\mathrm{I}\Delta_0 + \Omega_1$ into $\mathrm{I}\Delta_0$ by a cut is feasible.

Theorem 6.2.4 $\mathrm{I}\Delta_0 \rhd_f \mathrm{I}\Delta_0 + \Omega_1$ *by a cut.*

Proof. Let J be a cut constructed by Solovay's methods such that $\mathrm{I}\Delta_0$ proves that J is a cut closed under $+, \cdot$, and ω_1 (see lemma 2.6.10). Define φ^J to be the formula φ with all quantifiers restricted to J. It is well-known that J is an interpretation of $\mathrm{I}\Delta_0 + \Omega_1$ into $\mathrm{I}\Delta_0$; so to show that it is a feasible interpretation, it suffices to find a polynomial P such that for all Δ_0-formulas φ, the following holds: by proofs of length $\leq P(|\ulcorner\varphi\urcorner|)$:

$$\mathrm{I}\Delta_0 \vdash^{P(|\ulcorner\varphi\urcorner|)} [\varphi(0) \wedge \forall x(\varphi(x) \rightarrow \varphi(Sx)) \rightarrow \forall x\varphi(x)]^J.$$

First, it is easy to see that there is a polynomial P_1 such that for all Δ_0-formulas φ, $\mathrm{I}\Delta_0 \vdash^{P_1(|\ulcorner\varphi\urcorner|)} J(a) \rightarrow (\varphi(a) \leftrightarrow \varphi(a)^J)$ and $\mathrm{I}\Delta_0 \vdash^{P_1(|\ulcorner\varphi\urcorner|)} \forall x\varphi \rightarrow (\forall x\varphi)^J$. Second, there is a polynomial P_2 such that for all Δ_0-formulas φ, the following holds:

$$\mathrm{I}\Delta_0 \vdash^{P_2(|\ulcorner\varphi\urcorner|)} \forall a\, [\varphi(0) \wedge \forall x \leq a\, (\varphi(x) \rightarrow \varphi(Sx)) \rightarrow \forall x \leq a\, \varphi(x)].$$

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 Δ_0-formulas φ, the following holds:

$$\mathrm{I}\Delta_0 \vdash^{P_3(|\ulcorner\varphi\urcorner|)} (\forall a[\varphi(0) \wedge \forall x \leq a(\varphi(x) \rightarrow \varphi(Sx)) \rightarrow \forall x \leq a\, \varphi(x)])^J.$$

Now it is easy to find a polynomial P from P_3 such that for all Δ_0-formulas φ, the following holds:

$$\mathrm{I}\Delta_0 \vdash^{P(|\ulcorner\varphi\urcorner|)} [\varphi(0) \wedge \forall x(\varphi(x) \rightarrow \varphi(Sx)) \rightarrow \forall x\varphi(x)]^J.$$

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

Next, we will prove that the usual interpretation of $ZF + \mathbf{V} = \mathbf{L}$ 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 $\mathbf{L}$ of constructible sets. We will first prove that the schema of separation relativized to $\mathbf{L}$ follows feasibly from the reflection theorem for $\mathbf{L}$, and then give a feasible proof of the reflection theorem itself. Finally we give a proof of the schema of replacement relativized to $\mathbf{L}$. For the reflection schema, we will try to follow the elegant proof in terms of closed unbounded collections, which unfortunately becomes much less elegant when forced into the strait-jacket 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 φ of ZF is defined as the number of appearances of symbols in φ; 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 6.2.5 *ZF proves the following by proofs of length polynomial in* $|\varphi|$*:*

$$\begin{aligned}\forall z, \vec{v} \in \mathbf{L} \quad & \{x \in z \mid \varphi^{\mathbf{L}}(x, z, \vec{v})\} \in \mathbf{L} \rightarrow \\ & \forall z, \vec{v} \in \mathbf{L}\, \exists y \in \mathbf{L}\, [x \in y \leftrightarrow x \in z \wedge \varphi^{\mathbf{L}}(x, z, \vec{v})]\end{aligned}$$

Proof. Immediate; note that $(x \in y)^{\mathbf{L}} := x \in y$, so the succedent is feasibly equivalent to the comprehension schema for φ, relativized to $\mathbf{L}$. QED

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

Lemma 6.2.6 *ZF proves the following by proofs of length polynomial in* $|\varphi|$*:*

$$\begin{aligned}(\forall \alpha\, \exists \beta > \alpha\, \forall z, x, \vec{v} \in \mathbf{L}_\beta \quad & [\varphi^{\mathbf{L}}(x, z, \vec{v}) \leftrightarrow \varphi^{\mathbf{L}_\beta}(x, z, \vec{v})]) \rightarrow \\ & \forall z, \vec{v} \in \mathbf{L}\, \{x \in z \mid \varphi^{\mathbf{L}}(x, z, \vec{v})\} \in \mathbf{L}\end{aligned}$$

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

1. $\forall \alpha\, \exists \beta > \alpha\, \forall z, x, \vec{v} \in \mathbf{L}_\beta\, [\varphi^{\mathbf{L}}(x, z, \vec{v}) \leftrightarrow \varphi^{\mathbf{L}_\beta}(x, z, \vec{v})]$ and
2. $z, \vec{v} \in \mathbf{L}$

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

$$\{x \in z \mid \varphi^{\mathbf{L}}(x, z, \vec{v})\} = \{x \in \mathbf{L}_\beta \mid (x \in z \wedge \varphi(x, z, \vec{v}))^{\mathbf{L}_\beta}\} \in \mathit{Def}(\mathbf{L}_\beta) = \mathbf{L}_{\beta+1},$$

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

From lemma 6.2.5 and lemma 6.2.6, we conclude that in order to feasibly prove the comprehension schema, we only need polynomial length proofs of

$$\forall \alpha\, \exists \beta > \alpha\, \forall z, x, \vec{v} \in \mathbf{L}_\beta\, [\varphi^{\mathbf{L}}(x, z, \vec{v}) \leftrightarrow \varphi^{\mathbf{L}_\beta}(x, z, \vec{v})].$$

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

Definition 6.2.7 A collection C of ordinals is

- *unbounded* iff $\forall \alpha\, \exists \beta > \alpha\, (\beta \in C)$;
- *closed* iff $\forall a\, (a \neq \emptyset \wedge a \subseteq C \rightarrow \sup\, a \in C)$;
- *closed unbounded* (c.u.b.) iff C is both closed and unbounded.

Lemma 6.2.8 $ZF \vdash$ *"If C and $\mathcal{D}$ are c.u.b., then $C \cap \mathcal{D}$ is c.u.b. as well"*

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

Definition 6.2.9 A collection C of ordinals is a *closed unbounded φ-mirror* iff

1. C is closed unbounded, and
2. $\mathbf{L}_\alpha$ reflects φ for all ordinals $\alpha \in C$, i.e.
   $\forall \alpha\, (\alpha \in C \rightarrow \forall \vec{v} \in \mathbf{L}_\alpha\, [\varphi^{\mathbf{L}}(\vec{v}) \leftrightarrow \varphi^{\mathbf{L}_\alpha}(\vec{v})])$

Suppose φ is a formula and $\mathcal{D}$ is a first-order definable collection of ordinals. Using definition 6.2.7, we are able to construct new first-order formulas $CUB_{\mathcal{D}}$, $CUB_{\mathcal{D},\varphi}$ and REF_φ with the following intended meanings:

1. $CUB_{\mathcal{D}}$:= "$\mathcal{D}$ is closed unbounded"
2. $CUB_{\mathcal{D},\varphi}$:="$\mathcal{D}$ is a closed unbounded φ-mirror"
3. REF_φ:= " there is some collection of ordinals that is a closed unbounded φ-mirror"

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

Lemma 6.2.10 (Reflection theorem) *ZF proves the following by proofs of length polynomial in $|\varphi|$:*

$$\forall\alpha\,\exists\beta>\alpha\,\forall z,x,\vec{v}\in\mathbf{L}_\beta\,[\varphi^{\mathbf{L}}(x,z,\vec{v})\leftrightarrow\varphi^{\mathbf{L}_\beta}(x,z,\vec{v})]$$

Proof. First we note that ZF proves $(\alpha<\beta\to\mathbf{L}_\alpha\subseteq\mathbf{L}_\beta)$, "if γ is a limit ordinal, then $\mathbf{L}_\gamma=\bigcup_{\alpha<\gamma}\mathbf{L}_\alpha$" and $\mathbf{L}=\bigcup_{\alpha\in OR}\mathbf{L}_\alpha$.

We will prove the reflection theorem by induction on the construction of φ. A straightforward application of lemma 6.2.2 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. $\leq P(|Qx\psi|)$, ZF proves the following:

1. for atomic φ:

   $\forall\alpha\,\exists\beta>\alpha\,\forall z,x\in\mathbf{L}_\beta\,[\varphi^{\mathbf{L}}(x,z)\leftrightarrow\varphi^{\mathbf{L}_\beta}(x,z)]\wedge CUB_{OR,\varphi}$

2. the $\neg$-step:

   $\forall\alpha\,\exists\beta>\alpha\,\forall\vec{v}\in\mathbf{L}_\beta\,[\psi^{\mathbf{L}}(\vec{v})\leftrightarrow\psi^{\mathbf{L}_\beta}(\vec{v})]\wedge REF_\psi\quad\to$
   $\forall\alpha\,\exists\beta>\alpha\,\forall\vec{v}\in\mathbf{L}_\beta\,[\neg\psi^{\mathbf{L}}(\vec{v})\leftrightarrow\neg\psi^{\mathbf{L}_\beta}(\vec{v})]\wedge REF_{\neg\psi}$

3. the connective step, where $\circ\in\{\wedge,\vee,\to,\leftrightarrow\}$:

   $\forall\alpha\,\exists\beta>\alpha\,\forall\vec{v}\in\mathbf{L}_\beta\,[\psi^{\mathbf{L}}(\vec{v})\leftrightarrow\psi^{\mathbf{L}_\beta}(\vec{v})]\wedge REF_\psi\wedge$
   $\forall\alpha\,\exists\beta>\alpha\,\forall\vec{w}\in\mathbf{L}_\beta\,[\chi^{\mathbf{L}}(\vec{w})\leftrightarrow\chi^{\mathbf{L}_\beta}(\vec{w})]\wedge REF_\chi\quad\to$
   $\forall\alpha\,\exists\beta>\alpha\,\forall\vec{v}\in\mathbf{L}_\beta\,[\psi^{\mathbf{L}}\circ\chi^{\mathbf{L}}(\vec{v},\vec{w})\leftrightarrow\psi^{\mathbf{L}_\beta}\circ\chi^{\mathbf{L}_\beta}(\vec{v},\vec{w})]\wedge REF_{\psi\circ\chi}$

4. the quantifier step, where $Q\in\{\exists,\forall\}$:

   $\forall\alpha\,\exists\beta>\alpha\,\forall z,\vec{v}\in\mathbf{L}_\beta\,[\psi^{\mathbf{L}}(z,\vec{v})\leftrightarrow\psi^{\mathbf{L}_\beta}(z,\vec{v})]\wedge REF_\psi\quad\to$
   $\forall\alpha\,\exists\beta>\alpha\,\forall\vec{v}\in\mathbf{L}_\beta\,[Qz\in\mathbf{L}\,\psi^{\mathbf{L}}(z,\vec{v})\leftrightarrow Qz\in\mathbf{L}_\beta\,\psi^{\mathbf{L}_\beta}(z,\vec{v})]\quad\wedge REF_{Qz\psi}$

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 $\mathbf{L}_\alpha,\mathbf{L}$, some propositional reasoning independent on the specific ψ,χ, and an application of lemma 6.2.8 for step 3. We will show how the proofs of the $\exists$-case in step 4 can be bounded by a polynomial; we can then find a bound for the $\forall$-step by rewriting $\forall$ as $\neg\exists\neg$ and using the bounds for the $\neg$-step and the $\exists$-step.

Define

$$\mathcal{D}:=\{\beta\mid\forall\vec{v}\in\mathbf{L}_\beta\,[\exists z\in\mathbf{L}\,\psi^{\mathbf{L}}(z,\vec{v})\to\exists z\in\mathbf{L}_\beta\,\psi^{\mathbf{L}}(z,\vec{v})]\}.$$

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

$$\begin{array}{l}\forall\alpha\,\exists\beta>\alpha\,\forall z,\vec{v}\in\mathbf{L}_\beta\,[\psi^{\mathbf{L}}(z,\vec{v})\leftrightarrow\psi^{\mathbf{L}_\beta}(z,\vec{v})]\wedge CUB_{C_\psi,\,\psi}\wedge CUB_{\mathcal{D}}\rightarrow\\ \forall\alpha\,\exists\beta>\alpha\,\forall\vec{v}\in\mathbf{L}_\beta\,[\exists z\in\mathbf{L}\,\psi^{\mathbf{L}}(z,\vec{v})\leftrightarrow\exists z\in\mathbf{L}_\beta\,\psi^{\mathbf{L}}(z,\vec{v})]\wedge CUB_{C_\psi\cap\mathcal{D},\,\exists z\psi}\end{array}$$

In fact, we only use lemma 6.2.8 and the fact that $\forall\beta\,(\mathbf{L}_\beta\subseteq\mathbf{L})$. Thus we need to find a polynomial P such that $ZF\vdash CUB_{\mathcal{D}}$ by a proof of length $\leq P(|\exists z\psi|)$. Immediately from the definition, it is clear that $ZF\vdash$ "$\mathcal{D}$ 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 $\mathcal{D}$ is unbounded, that is:

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

i.e.

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

We will reason in ZF, taking care that all steps are applications of general ZF-theorem schemas that do not depend on the specific formula ψ. Take any ordinal α. We know using only predicate logic that

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

therefore,

$$\forall\vec{v}\in\mathbf{L}_\alpha\,\exists!\alpha_{\vec{v}}\,(\alpha_{\vec{v}}=\bigcap\{\beta>\alpha\mid\exists z\in\mathbf{L}_\beta\,[\exists z\in\mathbf{L}\,\psi^{\mathbf{L}}(z,\vec{v})\rightarrow\psi^{\mathbf{L}}(z,\vec{v})]\}).$$

by the unrelativized replacement and union axioms, there is a β_1 such that $\beta_1=\sup\{\alpha_{\vec{v}}|\vec{v}\in\mathbf{L}_\alpha\}$. Continuing in this way, we can define by recursion a sequence β_p for $p\in\omega$, where for all $p\in\omega$,

$$\forall\vec{v}\in\mathbf{L}_{\beta_p}\,\exists z\in\mathbf{L}_{\beta_{p+1}}\,[\exists z\in\mathbf{L}\,\psi^{\mathbf{L}}(z,\vec{v})\rightarrow\psi^{\mathbf{L}}(z,\vec{v})]\tag{6.2}$$

Define $\beta:=\sup\{\beta_p\mid p\in\omega\}$. Because $\alpha=\beta_0<\beta_1<\beta_2<\ldots$, we infer that β is a limit ordinal $>\alpha$. Now using (6.2) and the fact that $\mathbf{L}_\beta=\bigcup_{\gamma<\beta}\mathbf{L}_\gamma$, we find that

$$\forall\vec{v}\in\mathbf{L}_\beta\,\exists z\in\mathbf{L}_\beta\,[\exists z\in\mathbf{L}\,\psi^{\mathbf{L}}(z,\vec{v})\rightarrow\psi^{\mathbf{L}}(z,\vec{v})],$$

as desired. QED

Lemma 6.2.11 *For all φ, ZF feasibly proves the comprehension schema for φ, relativized to $\mathbf{L}$; i.e. by proofs of length polynomial in $|\varphi|$, ZF proves the following:*

$$\forall z,\vec{v}\in\mathbf{L}\,\exists y\in\mathbf{L}\,\forall x\in\mathbf{L}\,[x\in y\leftrightarrow x\in z\wedge\varphi^{\mathbf{L}}(x,z,\vec{v})]$$

Proof. Combine lemmas 6.2.5, 6.2.6 and 6.2.10. QED

Lemma 6.2.12 *For all φ, ZF feasibly proves the replacement schema for φ, relativized to $\mathbf{L}$; i.e. by proofs of length polynomial in $|\varphi|$, ZF proves the following:*

$$\begin{array}{ll}\forall a,\vec{v}\in\mathbf{L} & [\forall x\in a\,\exists!y\in\mathbf{L}\,\varphi^{\mathbf{L}}(x,y,\vec{v})\rightarrow\\ & \exists c\in\mathbf{L}\,\forall y\in\mathbf{L}\,(y\in c\leftrightarrow\exists x\in a\,\varphi^{\mathbf{L}}(x,y,\vec{v}))]\end{array}$$

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

$$ZF \vdash \forall a, \vec{v} \in \mathbf{L}\, [\forall x \in a\, \exists! y \in \mathbf{L}\, \varphi^{\mathbf{L}}(x, y, \vec{v}) \rightarrow \exists b \in \mathbf{L}\, (\forall x \in a\, \exists y \in b\, \varphi^{\mathbf{L}}(x, y, \vec{v}))]$$

The last proof works, as in lemma 6.2.10, by general theorem schemas of ZF that do not depend on the specific φ. Work in ZF and suppose $a, \vec{v} \in \mathbf{L}$ and $\forall x \in a\, \exists! y \in \mathbf{L}\, \varphi^{\mathbf{L}}(x, y, \vec{v})$. Now

$$\forall x \in a\, \exists! \beta_x\, (\beta_x = \bigcap\{\alpha \mid \exists y \in \mathbf{L}_\alpha\, \varphi^{\mathbf{L}}(x, y, \vec{v})\});$$

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

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

QED

Theorem 6.2.13 $ZF \rhd_f ZF + \mathbf{V} = \mathbf{L}$

Proof. We take the usual interpretation $^{\mathbf{L}}$ of $ZF + \mathbf{V} = \mathbf{L}$ into ZF. Because ZF is axiomatized by a finite list of axioms plus the schemata of comprehension and replacement, the lemmas 6.2.11 and 6.2.12 immediately imply that $^{\mathbf{L}}$ is a feasible interpretation. QED

Remark 6.2.14 Looking carefully at the proofs of the lemmas leading up to theorem 6.2.13 and using an analog of lemma 6.2.2 for polynomial time instead of polynomial length, one can observe that theorem 6.2.13 can be strengthened: the proofs in ZF of the interpreted $ZF + \mathbf{V} = \mathbf{L}$–axioms $\varphi^{\mathbf{L}}$ are not only of length polynomial in $|\varphi|$. There is even a deterministic polynomial time Turing machine M such that if the input of M is the code of an axiom φ of $ZF + \mathbf{V} = \mathbf{L}$, then M outputs the code of a ZF-proof of $\varphi^{\mathbf{L}}$.

It is a known result that $PRA \vdash Con(ZF) \rightarrow Con(ZF + \mathbf{V} = \mathbf{L})$ (see [Sm 77, Corollary 5.2.4]). One of the referees suggested that by theorem 6.2.13 this could perhaps be strengthened to $I\Delta_0 + \exp \vdash Con(ZF) \rightarrow Con(ZF + \mathbf{V} = \mathbf{L})$. The observation about the polynomial time computability of the proofs of the interpreted axioms, however, even leads to the conjecture that $I\Delta_0 + \Omega_1 \vdash \forall a(\alpha_{ZF+\mathbf{V}=\mathbf{L}}(a) \rightarrow \exists p Prf_{ZF}(p, a^{\mathbf{L}}))$, where $\alpha_{ZF+\mathbf{V}=\mathbf{L}}$ is a Δ_1^b-formula axiomatizing $ZF + \mathbf{V} = \mathbf{L}$ (cf. [Bu 86, Theorem 5.6]). Then, by a standard argument (involving Parikh's Theorem), we would have $I\Delta_0 + \Omega_1 \vdash Con(ZF) \rightarrow Con(ZF + \mathbf{V} = \mathbf{L})$.

Contrary to our expectations, the usual interpretation of $ZF + \mathbf{V} \neq \mathbf{L}$ 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 $ZF \rhd ZF + \mathbf{V} \neq \mathbf{L}$ and $ZF + AC \rhd ZF + AC + \neg CH$ 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, and those of $I\Delta_0 + \Omega_n$ into Q [HP 93, Section V.5] 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 $U \rhd V$ but not $U \rhd_f V$ are contrived theories obtained by fixed-point constructions like the ones in section 6.4. 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:

$$\begin{aligned} U \rhd_{fm} V \leftrightarrow \exists K \exists M (&\text{"}K \textit{is an interpretation and } M \textit{is a deterministic} \\ &\textit{polynomial time Turing Machine"} \wedge \forall a(\alpha_V(a) \rightarrow Prf_U(M(a), a^K))). \end{aligned} \tag{6.3}$$

This definition is more in line with the conventional use of the word "feasible" in the context of polynomial time computability. The clause $Prf_U(M(a), a^K)$ in (6.3) is a P-like formula, while the clause $\exists p(\text{"}|p| \leq P(|a|)\text{"} \wedge Prf_U(p, a^K))$ in the definition of feasible interpretability used in this chapter is an NP-like formula. However, all interpretations considered in this section can also be shown to be feasible in Buss' sense: as in remark 6.2.14, we only need an easy analog of lemma 6.2.2.

6.3 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 $A \rhd B$ is "$PA + A$ feasibly interprets $PA + B$". The definition of ILM can be found in section 2.5.

Definition 6.3.1 A *feasibility interpretation* is a map * which assigns to every propositional variable p a sentence p^* of the language of PA, and which is extended to all modal formulas as follows:

1. $(A \rhd B)^* = PA + A^* \rhd_f PA + B^*$
2. $(\Box A)^* = Prov_{PA}(A^*)$
3. * distributes over the boolean connectives.

Here $\rhd_f$ 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 interpretions *, $PA \vdash A^*$. Thus, we have to check that the axioms J1 to J5 are valid in PA when $A \rhd B$ is read as $PA + A \rhd_f PA + B$. 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 Σ_1^0-sentences can be added on both sides instead of one $\Box$-sentence only.

Lemma 6.3.2 *PA proves all feasibility translations of J1 to J5.*

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

J1 Suppose for some theory V and some p that $Prf_V(p, \ulcorner A \urcorner)$. Then by the identity interpretation and the polynomial bound $P(n) = n + |p|$, $V \rhd_f V + A$. So in particular, if $\Box_{PA}(A \to B)$, then $PA + A \rhd_f PA + A + B$, and surely $PA + A \rhd_f PA + B$.

J2 Suppose

- $U \rhd_f V$ by interpretation K_1 and polynomial P_1, and
- $V \rhd_f W$ by the interpretation K_2 and polynomial P_2.

As in the usual case, $U \rhd W$ by the interpretation $K_1 \circ K_2$. 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; the number of steps in this proof will be $\leq k \cdot |p|$, 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 $PA+A \rhd_f PA+B$ and $PA+B \rhd_f PA+C$, then $PA+A \rhd_f PA+C$.

J3 Suppose

- $U + A \rhd_f V$ by interpretation K_1 and polynomial P_1, and
- $U + B \rhd_f V$ by interpretation K_2 and polynomial P_2.

As in the usual case, we have $U+A \vee B \rhd V$ 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, $\vdash A \to (C^M \leftrightarrow C^{K_1})$ and $\vdash \neg A \to (C^M \leftrightarrow C^{K_2})$ 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 $Prf_U(p_1', \ulcorner A \to \urcorner c^M)$ with $|p_1'| \leq k \cdot (|c| + P(|c|) + P_1(|c|))$; and
- we can find p_2' such that $Prf_U(p_2', \ulcorner \neg A \wedge B \to \urcorner c^M)$ with $|p_2'| \leq k \cdot (|c| + P(|c|) + P_2(|c|))$.

Combining p_1' and p_2' and their respective polynomial bounds, we find p and P' such that $Prf_U(p, \ulcorner A \vee B \to \urcorner c^M)$ with $|p| \leq P'(|c|)$. Thus $U + A \vee B \rhd_f V$.

In particular, we have: if $PA + A \rhd_f PA + C$ and $PA + B \rhd_f PA + C$, then $PA + A \vee B \rhd_f PA + C$.

J4 Because $(PA + A \rhd_f PA + B) \rightarrow (PA + A \rhd PA + B)$, we have by the soundness of J4 for normal interpretability immediately $(PA + A \rhd_f PA + B) \rightarrow (\Diamond A \rightarrow \Diamond B)$.

J5 In an easier variation of lemma 6.5.11, we use a claim proved in [Vi 91b], which is stated in our chapter as lemma 6.5.10. Suppose β is a Σ_1^b-formula axiomatizing a subset U of a Σ_1^b-language L. We will prove that $\mathrm{I}\Delta_0 + \Omega_1 + \Diamond_\beta \top \rhd_f U$ i.e. $\mathrm{I}\Delta_0 + \Omega_1 + \Diamond_U \top \rhd_f U$.

By lemma 6.5.10, there is a polynomial P_1 such that

$$\begin{aligned} PA \vdash \; & \Box_{\mathrm{I}\Delta_0+\Omega_1+Con(\beta)} Con(\beta) \rightarrow \\ & \exists K \forall a \in Sent(L)[\mathrm{I}\Delta_0 + \Omega_1 + Con(\beta) \overset{P_1(|a|)}{\vdash\!\!\!-\!\!\!-} (\ulcorner \Box_\beta a \rightarrow \urcorner a^K)]. \end{aligned}$$

Of course we also know that $PA \vdash \Box_{\mathrm{I}\Delta_0+\Omega_1+Con(\beta)} Con(\beta)$, so

$$PA \vdash \exists K \forall a \in Sent(L)[\mathrm{I}\Delta_0 + \Omega_1 + Con(\beta) \overset{P_1(|a|)}{\vdash\!\!\!-\!\!\!-} (\ulcorner \Box_\beta a \rightarrow \urcorner a^K)].$$

On the other hand, by provable Σ_1^b-completeness there exists a polynomial P_2 such that

$$PA \vdash \forall a(\beta(a) \rightarrow [\mathrm{I}\Delta_0 + \Omega_1 + Con(\beta) \overset{P_2(|a|)}{\vdash\!\!\!-\!\!\!-} (\ulcorner \Box_\beta a \urcorner)]).$$

Combining the last two results, we have a polynomial P_3 such that

$$PA \vdash \exists K \forall a(\beta(a) \rightarrow [\mathrm{I}\Delta_0 + \Omega_1 + Con(\beta) \overset{P_3(|a|)}{\vdash\!\!\!-\!\!\!-} (a^K)]),$$

so $PA \vdash (\mathrm{I}\Delta_0 + \Omega_1 + \Diamond_\beta \top) \rhd_f U$. In particular, we have for any sentence A:

$$PA \vdash (\mathrm{I}\Delta_0 + \Omega_1 + \Diamond_{PA} A) \rhd_f PA + A,$$

thus

$$PA \vdash (PA + \Diamond_{PA} A) \rhd_f PA + A.$$

QED

We want to prove that Montagna's principle M holds for feasible interpretability over PA in its general version, where we can add an infinite set of Σ_1^0-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 6.4.1 of a set $\mathcal{S}$ of formulas such that $PA \vdash PA \rhd PA + \mathcal{S}$ but $\omega \not\models PA \rhd_f PA + \mathcal{S}$ also relies heavily on these details.

Suppose $U \supseteq PA$, $V \supseteq PA$. Now suppose $U \rhd_f V$ by the interpretation K (preserving $=$) with domain δ, and polynomial P. We want to find a polynomial Q such that for every Σ_1^0-sentence σ there is a $U + \sigma$-proof p of σ^K with $|\ulcorner p \urcorner| \leq Q(|\ulcorner \sigma \urcorner|)$. First, we need some definitions and lemmas. Fix U, V, K, P as given above.

Definition 6.3.3 Define *pism(s)* for "s is a partial isomorphism" and the function $G(j, y)$ as follows:

$$\begin{array}{lll} \mathit{pism}(s) & := & \mathit{seq}(s) \wedge (s)_0 = 0^K \wedge \forall i < \mathit{lh}(s) - 1((s)_{i+1} = S^K(s)_i) \\ G(j,y) & := & \exists s(\mathit{pism}(s) \wedge \mathit{lh}(s) = j + 1 \wedge (s)_j = y) \end{array}$$

Lemma 6.3.4 *$U \vdash \forall j \exists ! s(\mathrm{pism}(s) \wedge \mathrm{lh}(s) = j+1)$ 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 6.3.5 *U proves that g is injective, and $U \vdash \forall j \forall y(G(j, y) \rightarrow \delta(y))$.*

Proof. By induction. QED

Lemma 6.3.6 *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 *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 proof for $\cdot$ is analogous, and in the case of $\leq$ we can use the fact that $\leq$ is expressible via $+$.)

We have $U \vdash g(x + 0) = g(x) = g(x) +^K 0^K = g(x) +^K g(0)$ and $U \vdash g(x + y) = g(x) +^K g(y) \rightarrow g(x + Sy) = g(S(x + y)) = S^K(g(x + y)) = S^K(g(x) +^K g(y)) = g(x) +^K S^K(g(y)) = g(x) +^K g(Sy)$, So by induction on y, $U \vdash \forall x \forall y(g(x + y) = g(x) +^K g(y))$. QED

Lemma 6.3.7 *The range of g is 'closed downwards', i.e. $U \vdash \forall x \forall u(\delta(u) \wedge u <^K g(x) \rightarrow \exists y < x(u = g(y)))$.*

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. $PA \vdash \forall x \forall u(u < x + 1 \rightarrow u < x \vee u = x)$ and
2. $U \vdash \forall x(g(x) +^K 1^K = g(x + 1))$, we also have
3. $U \vdash \forall x \forall u(\delta(u) \wedge u <^K g(x + 1) \rightarrow u <^K g(x) \vee u = g(x))$.

Now we can start with the proof by induction on x that $U \vdash \forall x \forall u(\delta(u) \wedge u <^K g(x) \rightarrow \exists y < x(u = g(y)))$.

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

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

QED

Remark 6.3.8 Let $I(x)$ be the formula $\exists y(x = g(y))$. Note that if $U \nvdash \forall x\delta(x)$, then U does not prove that I defines a cut. For, suppose that $U \vdash I(0) \wedge \forall x(I(x) \rightarrow I(x+1))$. Then induction gives $U \vdash \forall x\exists y(x = g(y))$, thus, by lemma 6.3.5, $U \vdash \forall x\delta(x)$, contradicting our assumption.

On the other hand, by the previous lemma we do have $U \vdash \forall x\forall u(\delta(u) \wedge I(x) \wedge u <^K x \rightarrow I(u))$.

Lemma 6.3.9 *For all formulas $\varphi \in \Delta_0$, U proves the following by proofs of length polynomial in $|\ulcorner\varphi(x_1,\ldots,x_n)\urcorner|$:*

$$\varphi(x_1,\ldots,x_n) \leftrightarrow (\varphi^K)(g(x_1),\ldots,g(x_n)).$$

Proof. By induction on the construction of φ. We will see below that the proofs for the atomic formulae ψ are obviously of length linear in $|\ulcorner\psi\urcorner|$, and that all induction steps follow a given proof scheme in which the particular formulas at hand can be plugged in. So, because every φ has at most $|\ulcorner\varphi\urcorner|$ subformulas, there is a polynomial R such that for all φ, the U- proof of $\varphi(x_1,\ldots,x_n) \leftrightarrow (\varphi^K)(g(x_1),\ldots,g(x_n))$ is of length $\leq R(|\ulcorner\varphi\urcorner|)$. 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 6.3.6, we have for all terms t by proofs of length polynomial in $|\ulcorner t\urcorner|$:

$$U \vdash \forall x_1,\ldots,x_n(g(t(x_1,\ldots,x_n)) = (t^K)(g(x_1),\ldots,g(x_n))).$$

So suppose φ is the formula $t_1(x_1,\ldots,x_n) = t_2(x_1,\ldots,x_n)$ where $x_1,\ldots,x_n$ include all variables appearing in t_1 and t_2. Then, because U proves that g is an injective function,

$$\begin{array}{lrcl}
U \vdash & t_1(x_1,\ldots,x_n) & = & t_2(x_1,\ldots,x_n) \\
& \leftrightarrow g(t_1(x_1,\ldots,x_n)) & = & g(t_2(x_1,\ldots,x_n)) \\
& \leftrightarrow (t_1^K)(g(x_1),\ldots,g(x_n)) & = & (t_2^K)(g(x_1),\ldots,g(x_n)) \\
& \leftrightarrow ((t_1 & = & t_2)^K)(g(x_1),\ldots,g(x_n))
\end{array}$$

$\forall$x $\leq$ t-step Suppose that $\varphi(x_1,\ldots,x_n) = \forall x \leq t(x_1,\ldots,x_n)\psi(x,x_1,\ldots,x_n)$, and that $U \vdash \psi(x,x_1,\ldots,x_n) \leftrightarrow (\psi^K)(g(x),g(x_1),\ldots,g(x_n))$ (induction hypothesis). We will use the fact that, because of lemmas 6.3.5, 6.3.6 and 6.3.7, by proofs of length polynomial in $|\ulcorner t\urcorner|$:

$$\begin{array}{ll}
U \vdash \forall u, x_1,\ldots,x_n & (\exists x\,[x \leq t(x_1,\ldots,x_n) \wedge u = g(x)] \\
& \leftrightarrow \delta(u) \wedge u \leq^K t^K(g(x_1),\ldots,g(x_n))).
\end{array}$$

Thus, we have the following equivalences by proofs of length polynomial in $|\ulcorner\varphi(x_1,\ldots,x_n)\urcorner|$:

$$\begin{array}{ll}
U \vdash & \varphi(x_1,\ldots,x_n) \\
& \leftrightarrow \forall x \leq t(x_1,\ldots,x_n)\psi(x,x_1,\ldots,x_n) \\
& \leftrightarrow \forall x \leq t(x_1,\ldots,x_n)(\psi^K)(g(x),g(x_1),\ldots,g(x_n)) \qquad \text{(by ind. hyp.)} \\
& \leftrightarrow \forall u(\delta(u) \wedge u \leq^K t^K(g(x_1),\ldots,g(x_n)) \rightarrow \psi^K(u,g(x_1),\ldots,g(x_n))) \\
& \leftrightarrow (\forall x \leq t\,\psi)^K(g(x_1),\ldots,g(x_n)) \qquad \text{(by def. of } K\text{)} \\
& \leftrightarrow (\varphi)^K(g(x_1),\ldots,g(x_n)). \quad \text{QED}
\end{array}$$

Now we can finish the proof of the uniform version of Montagna's property M for feasible interpretability.

Theorem 6.3.10 *Suppose*

- *U satisfies full induction,*
- *V extends PA in its language and*
- *$U \rhd_f V$ by interpretation K (preserving $=$) and polynomial P.*

Then there is a polynomial Q such that for every Σ_1^0-sentence σ there is a $U + \sigma$-proof p of σ^K with $|\ulcorner p \urcorner| \leq Q(|\ulcorner \sigma \urcorner|)$.

Thus, $U + \mathcal{S} \rhd_f V + \mathcal{S}$ where $\mathcal{S}$ is a finite or infinite Σ_1^b-set of Σ_1^0-sentences.

Proof. Suppose $\sigma \in \mathcal{S}$ is the Σ_1^0-sentence $\exists x \varphi(x)$, where $\varphi \in \Delta_0$. By lemma 6.3.9, there is a polynomial R such that we can prove the following by a proof of length $\leq R(|\ulcorner \sigma \urcorner|)$:

$$\begin{aligned} U \vdash \exists x \varphi(x) &\rightarrow \exists x \varphi^K(g(x)) \\ &\rightarrow \exists y(\delta(y) \wedge \varphi^K(y)) \\ &\rightarrow (\exists x \varphi(x))^K. \end{aligned}$$

Now we have $U + \mathcal{S} \rhd_f V + \mathcal{S}$ 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 6.4.1. Let g be as defined in lemma 6.3.4. We will only give the result which needs some adaptation. The following preservation lemma corresponds to lemma 6.3.6:

Lemma 6.3.11 *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 6.3.6. Preservation of *exp* then follows in the same way as preservation of $+$ was proved from preservation of S in lemma 6.3.6. QED

6.4 Interpretability does not imply feasible interpretability

Theorem 6.4.1 *There is a set $\mathcal{S}$ of $\Delta_0(exp)$-sentences such that $PA \vdash PA \rhd PA + \mathcal{S}$, but $\omega \not\models PA \rhd_f PA + \mathcal{S}$.*

Proof. Define by Gödel's diagonalization theorem (or rather by the free variable version as formulated by Montague) a $\Delta_0(exp)$-formula $\varphi(y)$ such that

$$PA \vdash \varphi(y) \leftrightarrow \forall x \leq y \, \neg Prf(x, \ulcorner \varphi(\overline{y}) \urcorner).$$

It is easy to see that if we diagonalize directly, there is a polynomial O such that for each n, $|n| < |\ulcorner \varphi(\overline{n}) \urcorner| \leq O(|n|)$. Moreover, if $\varphi(\overline{n})$ were false, then by definition we would have a proof of the $\Delta_0(exp)$-sentence $\varphi(\overline{n})$; so $\varphi(\overline{n})$ must be true. But then, since $\varphi(\overline{n})$ is $\Delta_0(exp)$, we have the following:

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

Define $\mathcal{S} := \{\varphi(\bar{n}) \mid n \in \omega\}$. Then, by the identity interpretation, $\omega \models PA \rhd PA + \mathcal{S}$. Actually, as in [JM 88, section 6], we even have $PA \vdash \forall y Prov(\ulcorner \varphi(\bar{y}) \urcorner)$, so $PA \vdash PA \rhd PA + \mathcal{S}$.

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

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

We also know by lemma 6.3.9 (with $U = V = PA$) that there is a polynomial R such that for every n,

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

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

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

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

A salient feature of the counterexample above is the trivial identity interpretation by which PA interprets $PA + \mathcal{S}$. To prove that interpretability does not imply feasible interpretability, it is not essential that the set of formulas added to PA be infinite like $\mathcal{S}$ above. In theorem 6.4.2 we 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 cases of feasible interpretability; i.e. it is not true that if $U \rhd_f V$, $U \rhd_f B$ and $U \rhd V + B$, then $U \rhd_f V + B$ (take $U = V = PA$, $B = A(\bar{n})$ or $B = E^*$ as below in the proof of theorem 6.4.2).

Theorem 6.4.2 *There is a sentence A such that $\omega \models PA \rhd PA + A$, but $\omega \not\models PA \rhd_f PA + A$.*

In order to prove theorem 6.4.2, we need a well-known result and its proof, given in theorem 6.4.4 below. Solovay proved that the set $\{A | PA \rhd PA + A\}$ is Π^0_2-complete [So 76b]. This result inspired Hájek to prove that, for every n, the set $\{A \mid A \text{ is } \Pi^0_{n+1}\text{-conservative over } PA\}$ is also Π^0_2-complete [Há 79b].

We have adapted the proof of theorem 6.4.4 from Visser's unpublished rendition of an alternative proof by Lindström of Hájek's general result (see [Vi 90b]). First we need a definition.

Definition 6.4.3 Define $\Box_{U,x} B$ for "there is a proof of the formula B which only uses those axioms of U with Gödel number $\leq x$." Then define $\Box^*_U B := \exists x \Box_{U,x} B$.

Theorem 6.4.4 *Suppose U is a theory extending PA in the language of PA, such that for all B, $PA \vdash \forall x \Box_U (\Box_{U,x} B \rightarrow B)$ (reflection for U). Then for every Π_2-predicate $P(x)$, there is a formula A such that*

$$PA \vdash \Diamond_U \top \rightarrow \forall x((U \rhd U + A(x)) \leftrightarrow P(x)).$$

Proof.

The proof is taken almost verbatim from [Vi 90b].

Let $P(x)$ be any Π_2-predicate, say $P(x) = \forall x S(x,y)$, with $S \in \Sigma^0_1$. Pick R by diagonalization such that $PA \vdash R(x,y) \leftrightarrow S(x,y) \preceq \Box_U R(x,y)$. Let $Q(x,y) := \Box_U R(x,y) \preceq S(x,y)$. Now we can prove the following:

$$PA \vdash \forall x \forall y (\Box_U R(x,y) \leftrightarrow S(x,y) \vee \Box_U \bot). \tag{6.4}$$

In order to prove (6.4), work inside PA and suppose $\Box_U R(x,y)$. Then either $R(x,y)$ or $Q(x,y)$ holds. In case that $R(x,y)$ holds we have $S(x,y)$ by definition. In case that $Q(x,y)$ holds we have $\Box_U Q(x,y)$ by Σ^0_1-completeness, and hence by definition both $\Box_U R(x,y)$ and $\Box_U \neg R(x,y)$, thus $\Box_U \bot$.

For the other direction, suppose $S(x,y)$. Again we have either $R(x,y)$ or $Q(x,y)$. From $R(x,y)$ we find $\Box_U R(x,y)$ by Σ^0_1-completeness. From $Q(x,y)$ we immediately derive $\Box_U R(x,y)$. Finally $\Box_U \bot$ gives $\Box_U R(x,y)$ as well. This finishes the proof of (6.4).

Define A by diagonalization such that $PA \vdash A(x) \leftrightarrow \Box^*_U \neg A(x) \preceq \exists y \neg R(x,y)$. Note that by (6.4) we have $PA \vdash \Diamond_U \top \rightarrow \forall x[\forall y \Box_U R(x,y) \rightarrow P(x)]$ and $PA \vdash \forall x[P(x) \rightarrow \forall y \Box_U R(x,y)$. For this A, we can prove

$$PA \vdash \Diamond_U \top \rightarrow \forall x((U \ \rhd U + A(x)) \leftrightarrow P(x)). \tag{6.5}$$

First note that reflection for U allows us to apply the Orey–Hájek theorem in order to conclude that $PA \vdash \forall x(\forall y \Box_U \Diamond_{U,y} A(x) \leftrightarrow (U \ \rhd U + A(x)))$.

We start the proof proper of (6.5). Work in PA and suppose $\Diamond_U \top$.

→-side Pick any x and suppose $U \ \rhd U + A(x)$. Then by the Orey-Hájek theorem $\forall y \Box_U \Diamond_{U,y} A(x)$. We will prove $\forall y \Box_U R(x,y)$. Pick any y. We have $\Box_U[Q(x,y) \rightarrow \neg R(x,y)]$; therefore by definition of A,

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

and hence by reflection

$$\Box_U[Q(x,y) \rightarrow \neg A(x)].$$

But then there is a v such that

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

so by Σ^0_1- completeness

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

Also by Σ^0_1-completeness, there is a w such that

$$\Box_U[Q(x,y) \rightarrow \Box_{U,w} Q(x,y)].$$

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

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

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

$$\Box_U[\Box_U R(x,y) \rightarrow R(x,y)],$$

hence by Löb's theorem $\Box_U R(x,y)$. We may conclude $\forall y \Box_U R(x,y)$, thus, because we have $\Diamond_U \top$, we conclude $P(x)$.

$\leftarrow$-side Pick an x and suppose $P(x)$. Then $\forall y \Box_U R(x,y)$ and thus $\forall y \Box_U (\forall z < y\, R(x,z))$. It follows by definition of A that

$$\forall y \Box_U (\Box_{U,y} \neg A(x) \rightarrow A(x)).$$

On the other hand, we have

$$\forall y \Box_U (\Box_{U,y} \neg A(x) \rightarrow \neg A(x))$$

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

This finishes the proof of (6.5), and thus of the theorem. QED

Proof of theorem 6.4.2. Let $P(x)$ be some Π^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

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

and

$$PA \vdash A(x) \leftrightarrow \Box^*_{PA} \neg A(x) \preceq \exists y \neg R(x,y),$$

where $\Box^*$ is as defined in definition 6.4.3.

Carrying out the proof of theorem 6.4.4 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

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

Now suppose, to derive a contradiction, that

$$\omega \models \forall x[(PA \rhd PA + A(x)) \leftrightarrow (PA \rhd_f PA + A(x))].$$

Then

$$\omega \models \forall x[(PA \rhd_f PA + A(x)) \leftrightarrow P(x)].$$

However, it is easy to see that $PA \rhd_f PA + A(x)$ is a Σ^0_2-predicate, contradicting the Π^0_2-completeness of P. Therefore, there is an $n \in \omega$ such that

- $\omega \models PA \rhd PA + A(\overline{n})$ but
- $\omega \not\models PA \rhd_f PA + A(\overline{n})$.

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 $\{e \mid e$ is the Gödel number of a sentence E such that $\neg(PA \rhd_f PA + E)\} \in \Pi^0_2$, we can construct a formula A as in theorem 6.4.4 for which the following holds: for all sentences E,

$$\omega \models \neg(PA \rhd_f PA + E) \leftrightarrow PA \rhd PA + A(\ulcorner E \urcorner).$$

Now let E^* be the sentence constructed by the fixed point theorem such that

$$PA \vdash E^* \leftrightarrow A(\ulcorner E^* \urcorner).$$

Then

$$\omega \models \neg(PA \rhd_f PA + E^*) \leftrightarrow PA \rhd PA + E^*.$$

Therefore,

$$\omega \models PA \rhd PA + E^* \text{ and } \omega \not\models PA \rhd_f PA + E^*.$$

Incidentally, such a specific counterexample can also be constructed by a straightforward adaptation to Π^0_2 of a theorem of Myhill (see [Od 89, Proposition III.6.2]).

6.5 *ILM* is the 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 section 6.3 that for all modal formulas in the language of *ILM* we have:

if $ILM \vdash \varphi$, then for all feasibility interpretations *, $PA \vdash \varphi^*$.

Therefore, we will only need to show the converse:

if $ILM \not\vdash \varphi$, then there is a feasibility interpretation * such that $PA \not\vdash \varphi^*$.

We suppose that the reader has a copy of [Ber 90] at hand in order to follow the original proofs. For the lemmas 6.5.5 up to 6.5.7, knowledge of [Pu 86], [Pu 87] or chapter 3 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 Gödel 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 to chapter 3.

We apologize to the reader that in this chapter $Fmla_n$, Sat_n and $True_n$ have different meanings than in chapter 3: here the complexity measure is logical complexity, not length.

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

Definition 6.5.1 Formally, we define the following concepts:

- *Sent*(a) for "a is the Gödel number of a sentence";
- *Fmla*(a) for "a is the Gödel number of a formula";
- *Fmla*$_n(a)$ for "a is the Gödel number of a formula of logical complexity $\leq n$";
- *Cl*(a) for "the Gödel number of the universal closure of the formula with Gödel number a"; note that *Cl* denotes a function;
- *Indax*$_n(b)$ for "b is the Gödel number of an induction axiom of logical complexity $\leq n$", i.e.

$$\begin{aligned} Indax_n(b) \Longleftrightarrow\ & Fmla_n(b) \wedge \exists y[Fmla(y) \wedge \\ & b = Sub(y, \ulcorner v_1 \urcorner, \ulcorner 0 \urcorner)^\ulcorner \wedge \forall v_1(^\urcorner y^\ulcorner \rightarrow {}^\urcorner Sub(y, \ulcorner v_1 \urcorner, \ulcorner Sv_1 \urcorner)^\ulcorner) \rightarrow \forall v_1{}^\urcorner y. \end{aligned}$$

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 6.5.2 We formally define the following:

- *BPrf*$_n(x, y)$ for "x codes a proof of the formula coded by y, where only formulas of logical complexity $\leq n$ appear in the proof";
- $W \overset{P(n)}{\vdash} x$ for "x codes a formula that is provable in W by a proof of length $\leq P(n)$" where P is a polynomial;
- $W \overset{|n|}{\underset{*}{\vdash}} (x)$ for "there is a polynomial P such that" $\forall n \exists p(|p| \leq P(|n|) \wedge Prf_W(p, x))$;
- *Prov*$_n(x)$ for "x codes a formula that is provable by a proof which only uses those axioms of PA with Gödel number $\leq n$"; abbreviation $\Box_n \varphi$ for *Prov*$_n(\ulcorner \varphi \urcorner)$;
- *Prov*$_{W,v}(x)$ for "x codes a formula that is provable by a proof which only uses those axioms of W with Gödel number $\leq n$"; abbreviation $\Box_{W,n} \varphi$ for *Prov*$_{W,n}(\ulcorner \varphi \urcorner)$.

In the context of satisfaction predicates *Sat*$_n(x, w)$, we need a few more concepts.

Definition 6.5.3 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;
- *True*$_n(x)$ for $\forall w(Evalseq(w, x) \rightarrow Sat_n(x, w))$, where *Sat*$_n$ is as in [Pu 87];

Remark 6.5.4 When we prove formalized results, we read *True*$_n$ as a Gödel number just as *Sat*$_n$. So in that case the appropriate definition is as follows:

True$_n(x)$ for $\ulcorner \forall w(Evalseq(w, x) \rightarrow {}^\urcorner Sat_n(x, w)^\ulcorner) \urcorner$.

Lemma 6.5.5 (feasible subformula property) *There are polynomials P and Q such that*

$$PA \vdash \forall k \forall a(Fmla(a) \rightarrow PA \overset{P(|k|+|a|)}{\vdash} [Prov_k(a) \rightarrow \exists q(BPrf_{Q(|k|+|a|)}(q,a))])$$

Proof. In [Ta 75] Takeuti gives a proof of the free-cut elimination theorem for PA, where PA is formulated as a Gentzen system. In order to be able to use Takeuti's method, we first transform Hilbert proofs that use only axioms of Gödel number $\leq k$ into associated Gentzen proofs of length linear in the length of the original proofs, in which all non-logical axioms have Gödel number $\leq k$ and the induction rule is only applied to formulas of Gödel number $\leq k$.

Free cut-elimination then works in such a way that all principal formulas of induction inferences in the new free cut-free proof are substitution instances of principal formulas of induction inferences in the old proof. From this result one derives a proof of the corresponding subformula property: all formulas in the free cut-free proof of a are substitution instances of subformulas of either a principal formula of the induction rule for a formula of length $\leq |k|$, or an axiom of Q, or a itself.

At this point we can transform the Gentzen proof back into a Hilbert proof, again with only a linear increase in the length of the proofs and of the axioms occuring. We can then formalize the proof of the subformula property in PA: we find a polynomial Q such that

$$PA \vdash \forall k \forall a(Prov_k(a) \rightarrow \exists q(BPrf_{Q(|k|+|a|)}(q,a))).$$

But then it is easy to see that there is a polynomial P such that

$$PA \vdash \forall k \forall a(Fmla(a) \rightarrow PA \overset{P(|k|+|a|)}{\vdash} [Prov_k(a) \rightarrow \exists q(BPrf_{Q(|k|+|a|)}(q,a))]),$$

as desired. QED

Lemma 6.5.6 *There is a polynomial P such that*

$$PA \vdash \forall k \forall a(Fmla(a) \rightarrow PA \overset{P(|k|+|a|)}{\vdash} [\exists q BPrf_{|k|+|a|}(q,a) \rightarrow True_{|k|+|a|}(a)])$$

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 in advance) that $True_{|k|+|a|}$ 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_1 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. $b = (Sub(y, \ulcorner v_1 \urcorner, \ulcorner 0 \urcorner)\ulcorner \wedge \forall v_1(\urcorner y \ulcorner \rightarrow \urcorner Sub(y, \ulcorner v_1 \urcorner, \ulcorner S v_1 \urcorner)\ulcorner) \rightarrow \forall v_1 \urcorner y)$. We have to prove the following by a proof of length polynomial in $n := |k| + |a|$:

$$True_n(Sub(y, \ulcorner v_1 \urcorner, \ulcorner 0 \urcorner)\ulcorner \wedge \forall v_1(\urcorner y \ulcorner \rightarrow \urcorner Sub(y, \ulcorner v_1 \urcorner, \ulcorner S v_1 \urcorner)\ulcorner) \rightarrow \forall v_1 \urcorner y). \qquad (6.6)$$

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 3.3.12 and 3.3.16, but remember that a different complexity measure was used there), we can find a proof of length polynomial in n that (6.6) is equivalent to the following:

$$\forall w[Sat_n(y, s^*(1,0,w)) \wedge \forall x(Sat_n(y, s^*(1,x,w)) \rightarrow Sat_n(y, s^*(1,Sx,w))) \rightarrow \\ \forall x(Sat_n(y, s^*(1,x,w))]. \qquad (6.7)$$

The formulas (6.7) 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 $P(n) = c \cdot n^3$ should now suffice to carry out the proofs of (6.6).

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

$$PA \vdash \forall k \forall a(Fmla(a) \rightarrow PA \overset{P(|k|+|a|)}{\vdash} [\forall b(Indax_{|k|+|a|}(b) \rightarrow True_{|k|+|a|}(b))]).$$

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 6.5.7 *There is a polynomial P such that*

$$PA \vdash \forall k \forall a(Fmla(a) \rightarrow PA \overset{P(|k|+|a|)}{\vdash} (True_{|k|+|a|}(a)^{\ulcorner} \rightarrow \neg Cl(a))$$

Proof. By a formalized Tarski's snowing lemma; cf. lemma 3.3.11. QED

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

Theorem 6.5.8 (feasible reflection theorem)
There is a polynomial P such that

$$PA \vdash \forall k \forall a(Sent(a) \rightarrow PA \overset{P(|k|+|a|)}{\vdash} (\ulcorner Prov_k(a) \rightarrow \neg a))$$

Proof. Combine lemmas 6.5.5, 6.5.6 and 6.5.7. QED

In the following lemmas and theorems, $\exists K$ abbreviates
$\exists K$("K codes an interpretation" $\wedge \ldots$).

The next lemma was proved by Albert Visser [Vi 91a, section 6, Claim 3] in the course of a formalized Henkin construction in $\mathrm{I}\Delta_0 + \Omega_1$.

Lemma 6.5.9 *Suppose $U \supseteq \mathrm{I}\Delta_0 + \Omega_1$ and β axiomatizes some subset of a Σ_1^b-language L. Then there is an r such that*

$$\mathrm{I}\Delta_0 + \Omega_1 \vdash \Box_U Con(\beta) \rightarrow \exists K \forall a \in Sent(L) \exists p < \omega_1^r(a) Prf_U(p, \ulcorner \Box_\beta a \rightarrow \neg a^K).$$

Proof. See [Vi 91b]. QED

In remark 2.3.3, we pointed out that the values of ω_1-terms in a correspond to *exp*(the values of polynomials in $|a|$). Therefore, lemma 6.5.9 implies the following lemma:

Lemma 6.5.10 *Suppose $U \supseteq \mathrm{I}\Delta_0 + \Omega_1$ and β axiomatizes some subset of a Σ_1^b-language L. Then there is a polynomial P such that*

$$\mathrm{I}\Delta_0 + \Omega_1 \vdash \Box_U Con(\beta) \rightarrow \exists K \forall a \in Sent(L) U \overset{P(|a|)}{\vdash} (\ulcorner \Box_\beta a \rightarrow \neg a^K).$$

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

Theorem 6.5.11 (feasible Orey's theorem) *Suppose that $U \supseteq PA$ and W is given by a set of axioms defined by the Σ_1^b-formula α. Then*

$$PA \vdash \forall x[U \vdash_{*}^{|x|} (\ulcorner \Diamond_{\alpha,x} \top \urcorner)] \rightarrow U \ \rhd_f W.$$

Proof. Work in PA and suppose

$$\forall x[U \vdash_{*}^{|x|} (\ulcorner \Diamond_{\alpha,x} \top \urcorner)].$$

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

$$\beta(x) := \alpha(x) \wedge \Diamond_{\alpha,x+1} \top.$$

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

$$\Box_U Con(\beta).$$

> (For, reason in U and suppose $Prf_\beta(x, \bot)$, then for the axiom of β coded by the biggest Gödel number y to appear in x we have $\alpha(y) \wedge \neg \Diamond_{\alpha,y+1} \top$, thus $\neg\beta(y)$: a contradiction.)

On the other hand, by provable Σ_1^b-completeness for $\alpha(a)$ and by the assumption $\forall x[U \vdash_{*}^{|x|} (\ulcorner \Diamond_{\alpha,x} \top \urcorner)]$, we have a polynomial P_1 given in advance such that:

$$\forall a(\alpha(a) \rightarrow [U \vdash^{P_1(|a|)} (\ulcorner \alpha(a) \wedge \Diamond_{\alpha,x+1} \top \urcorner))].$$

So, by definition of β, we have the following for a polynomial P_2 fixed in advance:

$$\forall a(\alpha(a) \rightarrow [U \vdash^{P_2(|a|)} (\ulcorner \beta(a) \urcorner)]. \tag{6.8}$$

But, using $\Box_U Con(\beta)$ we can apply lemma 6.5.10 to first derive, for a polynomial P_3 fixed in advance:

$$\exists K \forall a \in Sent(L)[U \vdash^{P_3(|a|)} (\ulcorner \Box_\beta a \rightarrow \urcorner a^K)],$$

and thus for a polynomial P_4 fixed in advance:

$$\exists K \forall a \in Sent(L)[U \vdash^{P_4(|a|)} (\ulcorner \beta(a) \rightarrow \urcorner a^K)]. \tag{6.9}$$

Finally we can combine 6.8 and 6.9 to get the desired conclusion that there is a polynomial P given in advance such that

$$\exists K \forall a(\alpha(a) \rightarrow [U \vdash^{P(|a|)} (a^K))],$$

i.e. $U \ \rhd_f W$. QED

Now we can start the proof of the arithmetical completeness of ILM with respect to feasible interpretations (cf. definition 6.3.1) over PA.

Theorem 6.5.12 *If $ILM \nvdash B$, then there is a feasibility interpretation * such that $PA \nvdash B^*$.*

The proof will in most places be identical to the one in [Ber 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 [Ber 90].

Proof sketch. Suppose $ILM \nvdash B$, and take, by modal completeness of ILM with respect to simplified models (see definition 2.5.6), a provably primitive recursive ILM-Kripke model $V =< V, R, S, b, \Vdash>$, with $b = 1$ and $1 \nVdash B$. Extend V with a new root 0 with $0Rx$ for all $x \in V$, as in definition 5.1 of [Ber 90]. Adapting definition 5.2 of [Ber 90], we define a *feasibility interpretation* * such that for all propositional variables p,

$$p^* := \text{“}\exists x \in V \cup \{0\} : L = x \wedge x \Vdash p\text{”},$$

where L is defined as the limit of the Solovay function F, which is in turn defined in definition 5.7 of [Ber 90]. We want to prove the following:

$$\text{whenever } 1 \nVdash A, \text{then } PA \nvdash A^*, \tag{6.10}$$

Then we will be done, as we have chosen V such that $1 \nVdash B$. To prove (6.10), 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 * *respects* A, i.e.

$$PA \vdash \forall x \in V(x \Vdash A \wedge L = x \rightarrow A^*) \text{ and}$$

$$PA \vdash \forall x \in V(x \Vdash \neg A \wedge L = x \rightarrow \neg A^*).$$

It is clear from the definition of F that * is faithful on atomic formulas. Moreover, the induction steps for the propositional connectives and $\Box$ immediately follow from the proofs in [Ber 90]. Even the “negative” induction step for $\rhd$ has a straightforward proof:

> Work in PA and suppose $x \in V$, $x \Vdash \neg(A \rhd B)$, and $L = x$; then by part 2 in the proof of lemma 5.6 of [Ber 90] and by the induction hypothesis, $\neg(A^* \rhd B^*)$. But then surely $\neg(A^* \rhd_f B^*)$, thus, as * is a feasibility interpretation, $\neg(A \rhd B)^*$.

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 $\rhd$.

For $x \in V$, let *rank*(x, n), the rank of x at stage n, be defined as in definition 5.7 of [Ber 90]. The following proposition is an analog of proposition 5.14 in [Ber 90].

Proposition 6.5.13 (F has feasible property S) *PA proves the following:*

$$\begin{aligned} PA \vdash\ & \forall x \in V \cup \{0\}[L = x \rightarrow \\ & PA \vdash^{|k|}_{*} (\ulcorner \forall y, z \in V \cup \{0\}(L = y \wedge xRz \wedge ySz \rightarrow \Diamond_k L = z)\urcorner)] \end{aligned}$$

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 $\forall x, \forall y, \forall z$.

Fact 1 $PA \vdash PA \vdash^{|k|}_{*} (\ulcorner \forall y(L = y \rightarrow \Diamond_{\bar{k}} L = y)\urcorner)$

> Proof. Immediately from the reflection theorem 6.5.8. 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 $PA \vdash PA \overset{|k|}{\underset{*}{\vdash}} (\ulcorner \forall y(L = y \rightarrow \forall n(\bar{k} < \mathit{rank}(y,n)))\urcorner)$

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 $PA \vdash PA \overset{|k|}{\underset{*}{\vdash}} (\ulcorner \forall z(\Box_{\bar{k}} L \neq z \rightarrow \exists m \forall n \geq m(\mathit{rank}(z,n) \leq \bar{k}))\urcorner)$

Proof. Immediately from the definition of rank. QED

Fact 4 $PA \vdash PA \overset{|k|}{\underset{*}{\vdash}} (\ulcorner \forall y \forall z(L = y \wedge \Box_{\bar{k}} L \neq z \rightarrow \exists n(F(n) = y \wedge n \textit{ codes } y \wedge \mathit{rank}(z,n) \leq \bar{k} \wedge \mathit{rank}(z,n) < \mathit{rank}(y,n)))\urcorner))$

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 $\mathit{rank}(z,n) < \mathit{rank}(y,n)$, we use fact 2. QED

Fact 5 We have the following:

$$\begin{aligned} PA \vdash \; & \forall x(L = x \rightarrow PA \overset{|k|}{\underset{*}{\vdash}} (\ulcorner \forall y \forall z(L = y \wedge \Box_{\bar{k}} L \neq z \wedge xRz \wedge ySz \rightarrow \\ & \exists n(n \textit{ codes } y \wedge \mathit{rank}(z,n) < \mathit{rank}(y,n) \wedge \mathit{rank}(z,n) \leq \bar{k} \\ & \wedge F(\mathit{rank}(z,n))SxRz \wedge F(\mathit{rank}(z,n))Rz))\urcorner)) \end{aligned}$$

Proof. For the part up to $\mathit{rank}(z,n) \leq \bar{k}$, we use fact 4. For the last two conjuncts, we use the S-monotonicity of F and the property corresponding to M of Veltman ILM–frames. QED

Fact 6 We have the following:

$$\begin{aligned} PA \vdash \forall x(L = x \rightarrow PA \overset{|k|}{\underset{*}{\vdash}} \; & (\ulcorner \forall y \forall z(L = y \wedge \Box_{\bar{k}} L \neq z \wedge xRz \wedge ySz \rightarrow \\ & \exists n(F(n) = y \wedge F(n+1) = z))\urcorner) \end{aligned}$$

Proof. Immediate from fact 5 and the definition of the function F, clause 2. QED

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

$$PA \vdash \forall x[L = x \rightarrow PA \overset{|k|}{\underset{*}{\vdash}} (\ulcorner \forall y \forall z(L = y \wedge xRz \wedge ySz \rightarrow \Diamond_{\bar{k}} L = z)\urcorner)]$$

QED

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

Proposition 6.5.14 (positive induction step for $\rhd$) *Let * be the feasibility interpretation defined in the proof sketch of theorem 6.5.12. Suppose as induction hypothesis that*

$$PA \vdash \forall y(L = y \rightarrow (y \Vdash A \leftrightarrow A^*)) \textit{ and}$$

$$PA \vdash \forall z(L = z \rightarrow (z \Vdash B \leftrightarrow B^*)).$$

Then

$$PA \vdash \forall x(L = x \wedge x \Vdash A \rhd B \rightarrow (A \rhd B)^*).$$

Proof. Let b be such that

$$PA \vdash \forall y(L = y \to (y \Vdash A \leftrightarrow A^*)) \text{ and}$$

$$PA \vdash \forall z(L = z \to (z \Vdash B \leftrightarrow B^*)),$$

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

$$PA \vdash \forall z(z \Vdash B \to \Box_c(z \Vdash B));$$

for this, any $c \geq$ the Gödel number of the biggest axiom of Robinson's arithmetic Q will do. Define $d := max(b, c)$. By theorem 6.5.11, the feasible version of Orey's theorem, it is sufficient to prove the following:

$$PA \vdash \forall x(L = x \wedge x \Vdash A \rhd B \to \forall k \geq d\ PA \vdash^{|k|}_{*} (\ulcorner A^* \to \Diamond_{\bar{k}} B^{*\urcorner}).$$

Again, we will state a list of easily provable facts from which the result immediately follows.

Fact 1 $PA \vdash \forall x(L = x \wedge x \Vdash A \rhd B \to \Box[A^* \to \exists y(L = y \wedge xRy \wedge y \Vdash A \wedge x \Vdash A \rhd B)])$

Proof. $L = x \to \Box\exists y(L = y \wedge xRy)$ by property $(\neg R)$, $\Box(A^* \wedge L = y \to y \Vdash A)$ by the induction hypothesis, and $\Box(x \Vdash A \rhd B)$ by provable Σ^0_1-completeness. QED

Fact 2 $PA \vdash \forall x(L = x \wedge x \Vdash A \rhd B \to \Box[A^* \to \exists y\exists z(L = y \wedge xRy \wedge y \Vdash A \wedge x \Vdash A \rhd B \wedge xRz \wedge ySz \wedge z \Vdash B)])$

Proof. From fact 1 and the definition of $x \Vdash A \rhd B$. QED

Fact 3 $PA \vdash \forall z \forall k \geq d\ PA \vdash^{|k|}_{*} (z \Vdash B \to \Box_{\bar{k}} z \Vdash B)$

Proof. From the definition of d, and the fact that k appears only as efficient numeral. QED

Fact 4 $PA \vdash \forall x(L = x \wedge x \Vdash A \rhd B \to \forall k \geq d\ PA \vdash^{|k|}_{*} (\ulcorner A^* \to \exists y\exists z(L = y \wedge xRy \wedge xRz \wedge ySz \wedge \Diamond_{\bar{k}} L = z \wedge \Box_{\bar{k}} z \Vdash B)^\urcorner))$

Proof. From fact 2 for $A^* \to \exists y\exists z(L = y \wedge xRy \wedge xRz \wedge ySz \wedge z \Vdash B)$; fact 3 for a proof of length polynomial in k of $z \Vdash B \to \Box_{\bar{k}} z \Vdash B)$, and proposition 6.5.13 for a proof of length polynomial in $|k|$ of $L = y \wedge xRy \wedge xRz \wedge ySz \to \Diamond_{\bar{k}} L = z$. QED

Fact 5 $PA \vdash \forall x(L = x \wedge x \Vdash A \rhd B \to \forall k \geq d\ PA \vdash^{|k|}_{*} (\ulcorner A^* \to \exists z \Diamond_{\bar{k}}(L = z \wedge z \Vdash B)^\urcorner))$

Proof. If k is large 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 $|k|$: $PA \vdash \forall z([\Box_k(z \Vdash B \to L \neq z) \wedge \Box_k z \Vdash B] \to \Box_k L \neq z)$, and thus $PA \vdash \forall z(\Diamond_k L = z \wedge \Box_k z \Vdash B \to \Diamond_k(L = z \wedge z \Vdash B))$. This argument can be formalized and combined with fact 4 to derive fact 5. QED

Fact 6 $PA \vdash \forall x(L = x \wedge x \Vdash A \rhd B \rightarrow \forall k \geq d\ PA \overset{|k|}{\underset{*}{\vdash}} (\ulcorner A^* \rightarrow \Diamond_{\bar{k}} B^* \urcorner)$

Proof. From fact 5 and the induction hypothesis; the fact that $k \geq d$ is used at this place. We also use that $PA \vdash \forall k \geq d[PA \overset{|k|}{\underset{*}{\vdash}} (\exists z \Diamond_{\bar{k}}(L = z \wedge z \Vdash B) \rightarrow \Diamond_{\bar{k}} \exists z(L = z \wedge z \Vdash B))]$.

QED

From fact 6 and the feasible version of Orey's theorem, we may indeed derive

$$PA \vdash \forall x(x \Vdash A \rhd B \wedge L = x \rightarrow (A \rhd B)^*),$$

as desired.

QED

Proof sketch of theorem 6.5.12, continued. Concluding by induction that * respects all formulas A, we have proved that $ILM \not\vdash B^*$. Therefore, ILM is arithmetically complete with respect to feasible interpretability over PA.

QED.

Chapter 7

The complexity of feasible interpretability

How is it that life is orderly and you are content, a little cynical perhaps but on the whole just so, and then without warning you find the solid floor is a trapdoor and you are now in another place whose geography is uncertain and whose customs are strange?

(Jeanette Winterson, *The Passion*)

Abstract. We prove that there is a Σ_1-formula ξ such that

$$\{e \mid PA \text{ feasibly interprets } PA + \xi(\bar{e})\}$$

is Σ_2-complete. The method of proof that we use combines a recursion-theoretical reduction and an adaptation of some lemmas from Lindström's paper [Li 84].

7.1 Introduction

In this chapter, we continue our investigation of feasible interpretability begun in chapter 6. We remind the reader of the half-formal definition of feasible interpretability:

$$U \rhd_f V \leftrightarrow \exists K \exists P(\text{“}K is\ an\ interpretation\ and\ P\ is\ a\ polynomial\text{”} \wedge \forall a(\alpha_V(a) \rightarrow \exists p(\text{“}|p| \leq P(|a|)\text{”} \wedge Prf_U(p, a^K)))). \qquad (7.1)$$

Similarly, we define a concept of *feasible* Π_1*-conservativity*, given as

$$U \rhd_{\Pi_1 f} V \leftrightarrow \exists P(\text{“}P\ is\ a\ polynomial\text{”} \wedge \forall x \forall y(Fmla_{\Pi_1}(x) \wedge Prf_V(y, x) \rightarrow \exists p(\text{“}|p| \leq P(|y|)\text{”} \wedge Prf_U(p, x)))).$$

In section 6.2 we show that many interpretations encountered in everyday mathematics are feasible. For example, we have both $ZFC \rhd_f ZFC + CH$ and $ZFC \rhd_f ZFC + \neg CH$. All in all it seems that the only examples of theories U and V such that $U \rhd V$ but not $U \rhd_f V$ are contrived sets of sentences obtained by fixed-point constructions. Moreover, $\rhd$ and $\rhd_f$ turn out to behave rather similarly with respect to their modal-logical properties.

However, when we study the definitional complexity of feasible interpretability, the difference with normal interpretability is striking. It is clear from (7.1) that the formula $U \rhd_f V$ is Σ^0_2. On the other hand Solovay in [So 76b] proved that $\{A \mid PA \rhd PA + A\}$ is Π^0_2-complete. This result in turn inspired Hájek to prove that, for every n, the set $\{A \mid A$ is Π^0_{n+1}-conservative over $PA\}$ is Π^0_2-complete [Há 79b].

Bearing in mind Rogers' observation in [Rog 67] that "almost all arithmetical sets with intuitively simple definitions that have been studied [...] have proved to be Σ^0_n-complete or Π^0_n-complete (for some n)", we would like to know whether feasible interpretability is complete for some level of the arithmetical hierarchy.

Indeed, it turns out that there is a Σ^0_1-formula ξ such that $\{e \mid PA \rhd_f PA + \xi(\bar{e})\}$ is Σ^0_2-complete, as we prove in section 7.6. From our methods we immediately derive that $\{e \mid PA \rhd PA + \xi(\bar{e})$ but not $PA \rhd_f PA + \xi(\bar{e})\}$ is not only inhabited, but even rather wildly so – to be explicit, it is Π^0_2-complete. Thus the two completeness results provide some precise evidence for the observation that normal interpretability and feasible interpretability over PA have substantially different extensions.

The formula ξ that we use for the Σ^0_2-completeness results is as simple as possible. More precisely, it is easy to show that for any Π^0_1-formula $\xi(x)$, $\{e \mid PA \rhd_f PA + \xi(\bar{e})\}$ is equal to $\{e \mid PA \vdash \xi(\bar{e})\}$, which is recursively enumerable.

The rest of the chapter is organized as follows. In section 7.2 we give some preliminaries on partial truth definitions and some notational conventions. In section 7.3 we characterize feasible interpretability in terms of feasible Π_1-conservativity. Section 7.4 contains the main novelty of this chapter, namely a recursion-theoretical reduction by which we show that the set of (possibly infinite) theories feasibly interpretable over PA is Σ^0_2-complete.

Lindström provided in [Li 84] a general method by which one can replace every recursively enumerable set Y of Σ^0_n-sentences by a single Σ^0_n-sentence σ such that $PA + \sigma$ has the same Π^0_n-consequences as $PA + Y$. In section 7.5, we prove a feasible version of Lindström's lemmas. The proofs in this section are fairly straightforward.

Finally, in section 7.6, we apply the methods of section 7.5 to the possibly infinite, but still suitably simple, sets of formulas constructed in section 7.4. Thus, using the characterization of feasible interpretability over PA as feasible Π_1-conservativity, we prove that there is a Σ^0_1-formula $\xi(x)$ such that $\{e \mid PA \rhd_f PA + \xi(\bar{e})\}$ is Σ^0_2-complete.

The chapter is almost self-contained. However, for some details of proofs we refer the reader to chapter 6, and we suppose that the reader is at least slightly familiar with the terminology used in [Bu 86].

7.2 Preliminaries and notation

We use Pudlák's notation $T \overset{n}{\vdash} \varphi$ and $T \overset{|n|}{\underset{*}{\vdash}} \varphi(n)$ as discussed in notation 2.6.7.

We will also sloppily leave out some Gödel brackets and numeral dots, in particular deeper nested ones. We use efficient numerals $\bar{n}$ of length linear in $|n|$.

We suppose that all theories mentioned in the sequel are Σ^b_1-axiomatized.

Definition 7.2.1 $Prov_{k,T}(\ulcorner A \urcorner)$ stands for "there is a proof of A from T in which only axioms with Gödel number $\leq k$ are used".

$$Con_k(T) \ := \ \neg Prov_{k,T}(\ulcorner \bot \urcorner).$$

In this chapter we use two kinds of partial truth predicates. Pudlák in [Pu 87] introduced the first kind that we need. His $True_n$ are truth predicates of length linear in n

which work for quantifier depth $\leq n$. Pudlák works with a relational language, whereas in chapter 3, the standard language of arithmetic including function symbols was used. In the present chapter, as in chapter 6, we use Pudlák's complexity measure, namely the logical complexity and not the length of formulas. We have the following Tarski lemma:

Lemma 7.2.2 *There is a polynomial P such that for all sentences A of length $\leq |k|$,*

$$PA \vdash^{P(|k|)} True_{|k|}(A) \leftrightarrow A.$$

Proof. See [Pu 87]. (cf. chapter 3. Pudlák works with a relational language, whereas in chapter 3, the standard language of arithmetic including function symbols was used.) QED

Lemma 7.2.2 can be formalized to get:

Lemma 7.2.3

$$PA \vdash \exists P \forall k \forall a (Sent(a) \rightarrow PA \vdash^{P(|k|+|a|)} (True_{|k|+|a|}(a) \leftrightarrow a)).$$

Definition 7.2.4 A theory T in the language of arithmetic is *feasibly essentially reflexive* if there is a polynomial P such that for all sentences A and for all k,

$$T \vdash^{P(|k|+|A|)} (Prov_{k,T}(\ulcorner A \urcorner) \rightarrow A).$$

Lemma 7.2.5 *PA is feasibly essentially reflexive, even provably so, i.e. we have:*

$$PA \vdash \exists P \forall k \forall a (Sent(a) \rightarrow PA \vdash^{P(|k|+|a|)} (Prov_{k,PA}(a) \rightarrow True_{|k|+|a|}(a))),$$

thus

$$PA \vdash \exists P \forall k \forall a (Sent(a) \rightarrow PA \vdash^{P(|k|+|a|)} (Prov_{k,PA}(a) \rightarrow a)).$$

Proof. See theorem 6.5.8 and lemma 7.2.3. QED

We will need a similar result in section 7.5. First we need two definitions.

Definition 7.2.6 We say that a set A is *sparse* if there exists a polynomial P such that for every n the number of elements of A having length $\leq n$ is bounded by $P(n)$.

Definition 7.2.7 A Σ_1^b-formula α defines a *provably sparse* relation if there is a polynomial P such that for all q,

$$PA \vdash^{P(|q|)} \forall z(\alpha(z) \wedge z \leq \bar{q} \rightarrow \bigvee_{x \leq \bar{q} \wedge \alpha(x)} z = \bar{x}).$$

Lemma 7.2.8 *Let A be an extension of PA by a provably sparse set of axioms in the language of PA. Then we have*

$$A \vdash^{|m|}_{*} \forall u, v \leq \overline{m}(Prf_A(v,u) \rightarrow True_{|m|}(u)).$$

Proof. Suppose the set of new axioms over PA is given by the Σ_1^b-formula α. The proof uses lemma 6.5.5 and lemma 6.5.6. In order to apply lemma 6.5.5, we work in PA and first transform proofs with Gödel number $\leq m$ of u from A into PA_m-proofs of (par abus de langage)

$$(\bigwedge_{x\leq\overline{m}\wedge\alpha(x)} x)^{\ulcorner} \rightarrow {}^{\urcorner}u;$$

this can be done polynomially in $|m|$.

Next we apply lemma 6.5.5 and lemma 6.5.6. Finally we note that, due to the provable sparsity of the set of new axioms defined by α, we have

$$A \overset{|m|}{\underset{*}{\vdash}} True_{|m|}(\bigwedge_{x\leq\overline{m}\wedge\alpha(x)} x),$$

Thus we derive the lemma. We leave the details, e.g. of formalization of $\bigwedge_{x\leq\overline{m}\wedge\alpha(x)} x$, to the reader.

QED

The second kind of partial truth predicates that we use are the standard ones related to the levels of the arithmetical hierarchy.

Lemma 7.2.9 *For all $i \geq 1$ there are partial truth definitions $True_{\Sigma_i}$ and $True_{\Pi_i}$ such that the Tarski lemmas have short proofs. More precisely, there is a polynomial P such that:*

- *for all Π_i-sentences π, $PA \overset{P(|\pi|)}{\vdash} \pi \leftrightarrow True_{\Pi_i}(\ulcorner\pi\urcorner)$, and*
- *for all Σ_i-sentences σ, $PA \overset{P(|\sigma|)}{\vdash} \sigma \leftrightarrow True_{\Sigma_i}(\ulcorner\sigma\urcorner)$.*

Proof. Visser in [Vi 92] gives a $\Delta_0(exp)$ definition of satisfaction for Δ_0-formulas. It is easy to construct from this partial truth definitions $True_{\Sigma_i}$ and $True_{\Pi_i}$ of length linear in i. We leave the reader the easy but tedious task of showing that the Tarski lemmas indeed have short proofs. QED

We also need a result relating the two kinds of partial truth definitions.

Lemma 7.2.10 *Let $\Gamma_i \in \{\Sigma_i, \Pi_i\}$. Then*

$$PA \overset{|m|}{\underset{*}{\vdash}} \forall u \leq \overline{m}[Fmla_{\Gamma_i}(u) \wedge True_{|m|}(u) \rightarrow True_{\Gamma_i}(u)].$$

Proof. We leave the proof to the reader. QED

7.3 Characterizations of feasible interpretability

In section 7.6, we will use a characterization which says that, over suitable theories, feasible interpretability is equivalent to feasible Π_1-conservativity. We prove a formalized version of this equivalence in corollary 7.3.3.

Non-feasible versions of all three characterizations below are well-known. We proved a feasible version of Orey's Theorem in chapter 6.

We have the following Orey-Hájek-style characterization of feasible interpretability. (Provably sparse relations are defined in definition 7.2.7.)

Lemma 7.3.1 (Feasible Orey-Hájek characterization) *Suppose $U \supseteq I\Sigma_1$ and V is an extension of PA by a provably sparse set of axioms in the language of PA. Then*

$$PA \vdash [U \vdash^{|x|}_{*} Con_x(U)] \rightarrow (U \rhd_f V \leftrightarrow [U \vdash^{|x|}_{*} Con_x(V)]).$$

Proof.

→ We first state a useful fact. We axiomatize each PA_k using only one induction axiom. To be precise, there is a function $i \mapsto \ulcorner\theta_i\urcorner$, with $|\ulcorner\theta_i\urcorner|$ polynomial in i, definable in $I\Sigma_1$, and a Σ^b_1-formula $\alpha^*_{x,V}$ (standing for the axiom set that contains the induction axiom for $\theta_{|x|}$ plus the sparse set of remaining axioms of V) such that

$$I\Sigma_1 \vdash \forall x, z(\alpha^*_{x,V}(z) \rightarrow z \leq x)$$

and

$$I\Sigma_1 \vdash \forall x(Prov_{x,V}(\perp) \rightarrow Prov_{\alpha^*_{x,V}}(\perp)).$$

Therefore we have, by instantiation,

$$PA \vdash [U \vdash^{|x|}_{*} Prov_{x,V}(\perp) \rightarrow Prov_{\alpha^*_{x,V}}(\perp)]. \tag{7.2}$$

We now start our proof proper by reasoning in PA and assuming that

$$U \vdash^{|x|}_{*} Con_x(U). \tag{7.3}$$

If $U \rhd_f V$ by interpretation K, then there is a polynomial R such that

$$\forall x \forall z \leq x(\alpha^*_{x,V}(z) \rightarrow U \vdash^{R(|x|)} z^K).$$

Because $\alpha^*_{x,V}$ singles out a provably sparse set, we have provable completeness, thus we derive

$$U \vdash^{|x|}_{*} \forall z \leq x(\alpha^*_{x,V}(z) \rightarrow U \vdash^{R(|x|)} z^K). \tag{7.4}$$

Next we want to find a polynomial Q such that

$$U \vdash^{|x|}_{*} Prov_{\alpha^*_{x,V}}(\perp) \rightarrow \neg Con_{2^{Q(|x|)}}(U). \tag{7.5}$$

So reason inside U and suppose $Prf_{\alpha^*_{x,V}}(q, \perp)$. If we take K-translations of all formulas in q and add some intermediate steps, we find a quasi-U-proof p of $\perp$ that still depends on some assumptions z^K where $\alpha^*_{x,V}(z)$. But, since $z \leq x$, we know by (7.4) that these z^K have U-proofs of length $\leq R(|x|)$. We add these proofs to p in order to find a U-proof p' of $\perp$ that uses only U-axioms of length $\leq Q(|x|)$ for some polynomial Q given in advance. Stepping out of U again, we find that indeed (7.5) holds.

However, by (7.3) we have $U \vdash^{|x|}_{*} Con_{2^{Q(|x|)}}(U)$, so (7.5) gives $U \vdash^{|x|}_{*} \neg Prov_{\alpha^*_{x,V}}(\perp)$. By (7.2) we finally conclude $U \vdash^{|x|}_{*} Con_x(V)$.

$\leftarrow$ For this direction, which is a feasible version of Orey's Theorem, we do not need the assumption $U \vdash^{|x|}_{*} Con_x(U)$, nor do we need the provable sparsity of V over PA. We make the desired interpretation by a feasible Henkin construction for a Feferman proof predicate of V. For details of the proof we refer the reader to theorem 6.5.11.

QED

Lemma 7.3.2

$$PA \vdash [V \vdash^{|x|}_{*} Con_x(V)] \rightarrow (U \rhd_{\Pi_1 f} V \leftrightarrow [U \vdash^{|x|}_{*} Con_x(V)]).$$

Proof.

$\rightarrow$ Reason in PA and suppose that $U \rhd_{\Pi_1 f} V$ and $V \vdash^{|x|}_{*} Con_x(V)$. Because $Con_x(V)$ is a Π_1-sentence, this immediately gives us $U \vdash^{|x|}_{*} Con_x(V)$.

$\leftarrow$ This direction of the proof does not depend on the assumption $V \vdash^{|x|}_{*} Con_x(V)$.

So, reason in PA and suppose that $U \vdash^{|x|}_{*} Con_x(V)$, $Fmla_{\Pi_1}(x) \wedge Prf_V(y, x)$. (We will use without mention the fact that $x \leq y$.) First we analyze the proof of Σ^b_1-completeness, and we note that there is a fixed m — to be explicit, m is the Gödel number of the largest axiom of Robinson's Arithmetic Q — for which we have the following:

$$U \vdash^{|x|}_{*} Con_m(V + x) \rightarrow x. \tag{7.6}$$

Because the axioms of V are recognized in a Σ^b_1-way, we can again invoke provable Σ^b_1-completeness to show that $Prf_V(y, x)$ implies

$$U \vdash^{|y|}_{*} \neg Con_y(V + \neg x). \tag{7.7}$$

By our assumption $U \vdash^{|x|}_{*} Con_x(V)$ we have $U \vdash^{|y|}_{*} Con_{max(m,y)}(V)$, which we may combine with (7.7) to derive $U \vdash^{|y|}_{*} Con_m(V + x)$, thus by (7.6) $U \vdash^{|y|}_{*} x$, as desired.

QED

Corollary 7.3.3 *If $U \supseteq I\Sigma_1$ and V is an extension of PA by a provably sparse set of axioms, then*

$$PA \vdash [U \vdash^{|x|}_{*} Con_x(U)] \wedge [V \vdash^{|x|}_{*} Con_x(V)] \rightarrow (U \rhd_f V \leftrightarrow U \rhd_{\Pi_1 f} V).$$

Proof.

$\rightarrow$ By the $\rightarrow$-direction of lemma 7.3.1 and the $\leftarrow$-direction of lemma 7.3.2; we do not need the assumption $V \vdash^{|x|}_{*} Con_x(V)$.

$\leftarrow$ By the $\rightarrow$-direction of lemma 7.3.2 and the $\leftarrow$-direction of lemma 7.3.1; we do not need the assumption $U \vdash^{|x|}_{*} Con_x(U)$.

QED

Because PA is provably feasibly essentially reflexive, we have the following useful characterization for feasible interpretability over PA:

Corollary 7.3.4 *For all formulas A, B in the language of PA,*

$$PA \vdash PA + A \rhd_f PA + B \leftrightarrow PA + A \rhd_{\Pi_1 f} PA + B.$$

Proof. Immediately from lemma 7.2.5 and corollary 7.3.3 QED

7.4 The set of Σ_1^b-axiomatized theories feasibly interpretable over PA is Σ_2-complete

In this section, we will prove that the set of theories that can be feasibly interpreted in PA is Σ_2-complete. We assume that each Σ_1^b-axiomatized theory is given by a code of a non-deterministic polynomial time Turing machine that accepts exactly the Gödel numbers of axioms of the theory in question, and we suppose that the coding of Turing machines is standard, e.g. as in [BDG 87].

By the way, the reduction that we use to prove theorem 7.4.1 can be easily adapted to yield a slightly alternative proof of Hájek's theorem that $\{e \mid$ the deterministic Turing machine coded by e works in polynomial time$\}$ is Σ_2-complete (cf. [Há 79a]).

Let us turn to the technicalities. Take some Gödel numbering which codes formulas in the language of arithmetic (which for our purpose includes a function symbol *exp*) as binary numbers. L_e stands for the language accepted by the Turing machine with code e. By writing out the definitions we see that $E := \{e \mid e$ codes a deterministic Turing machine such that PA feasibly interprets $PA+$ the set of formulas whose codes are in $L_e\} = \{e \mid \exists K\ \exists P\ \forall x \forall y(x$ is an axiom of PA or y is an accepting computation of e on $x \to \exists z(|z| \leq P(|x|) \wedge Prf(z, x^K)))\}$ is in Σ_2.

It is well-known that $\{e \mid W_e$ finite$\}$ is Σ_2-complete (cf [So 87]). So in order to prove that E is in fact Σ_2-complete, the following theorem suffices.

Theorem 7.4.1 *There is a total recursive function F such that for all e:*

W_e is finite
$\Leftrightarrow$ $F(e)$ codes a non-deterministic polynomial time Turing machine such that PA feasibly interprets PA + the set of formulas whose codes are in $L_{F(e)}$.

In order to prove this theorem, we first introduce a definition and prove a lemma.

Definition 7.4.2 Define by Gödel's diagonalization theorem (or rather by the free variable version as formulated by Montague) a $\Delta_0(exp)$-formula $\varphi(y)$ such that

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

This fixed point is a bounded analog to the fixed point that Gödel used to prove his First Incompleteness Theorem. Informally, every $\varphi(\bar{n})$ says "I am not provable by any short proof". Part of the proof of the following lemma is reminiscent of Gödel's argument. It is almost identical to the proof of theorem 6.4.1.

Lemma 7.4.3 *For any NP-subset X of the natural numbers defined by a Σ_1^b-formula α, we have*

$$X \text{ is finite} \Leftrightarrow PA \rhd_f PA + \{\varphi(\bar{n}) | \alpha(n)\}.$$

Proof.

- It is easy to see that there is a polynomial O such that for each n, $|n| < |\ulcorner \varphi(\bar{n}) \urcorner| \leq O(|n|)$.
- If $\varphi(\bar{n})$ were false, then by definition we would have a proof of the $\Delta_0(exp)$-sentence $\varphi(\bar{n})$; so $\varphi(\bar{n})$ must be true. But then, since $\varphi(\bar{n})$ is $\Delta_0(exp)$, we have the following:

1. PA proves $\varphi(\overline{n})$, though
2. because $\varphi(\overline{n})$ is true, PA does not prove $\varphi(\overline{n})$ by any proof whose Gödel number is of length $\leq 2^{|n|}$.

- If X is finite, then it is obvious that $PA \rhd_f PA + \{\varphi(\overline{n})|\alpha(n)\}$.
- Suppose X is infinite and $PA \rhd_f PA + \{\varphi(\overline{n})|\alpha(n)\}$ by interpretation K and polynomial P. Thus, for all $n \in X$,

  $$PA \overset{P(|\ulcorner\varphi(\overline{n})\urcorner|)}{\vdash} \varphi(\overline{n})^K.$$

  Since φ is $\Delta_0(exp)$, we also know by lemma 3.10 of [Ve 93] (with $U = V = PA$) that there is a polynomial R such that for every $n \in X$,

  $$PA \overset{R(|\ulcorner\varphi(\overline{n})\urcorner|)}{\vdash} \varphi(\overline{n}) \leftrightarrow \varphi(\overline{n})^K.$$

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

  $$PA \overset{Q(|\ulcorner\varphi(\overline{n})\urcorner|)}{\vdash} \varphi(\overline{n}).$$

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

QED

Now we can prove the theorem by giving an appropriate reduction F.

Proof. For e, t given, we describe the behavior of the deterministic Turing machine coded by $F(e)$ on input t. In the rest of the proof we will sloppily mention the codes instead of the machines or functions that they code. As usual, $|s|$ stands for the number of symbols that s consists of.

```
IF $t$ is not of the form $\ulcorner\varphi(\bar{s})\urcorner$ for any $s$,
    THEN we halt and reject $t$;
    ELSE we find the $s$ such that $t$ is of the form $\ulcorner\varphi(\bar{s})\urcorner$;
    first we simulate the behavior of $e$ on inputs $1, \ldots, |s|$, for at most $|s|$ steps each.
    IF $|s| > |s - 1|$ AND
        $e$ halts on $|s|$ within $|s|$ steps, OR
        there is an $i \leq |s| - 1$ such that
            $e$ halts on $i$ within $|s|$ steps AND
            $e$ does not halt on $i$ within $|s| - 1$ steps,
        THEN we halt and accept $t$;
        ELSE we halt and reject $t$.
```

We want to show that F is the required function. To this end, first define $s_e(i) :=$ the smallest s such that

- $i \leq |s|$ and
- e halts on i within $|s|$ steps.

Now we prove the theorem.

($\Rightarrow$) Suppose W_e is finite, say $W_e = \{i_1, \dots, i_n\}$.

Then $L_{F(e)} = \{\ulcorner\varphi(\overline{s_e(i_1)})\urcorner, \dots, \ulcorner\varphi(\overline{s_e(i_n)})\urcorner\}$. By lemma 7.4.3 we know that $PA \rhd_f PA + \{\varphi(\overline{s_e(i_1)}), \dots, \varphi(\overline{s_e(i_n)})\}$.

($\Leftarrow$) Suppose W_e is infinite. We know that for all $i \in W_e$, $i \leq |s_e(i)|$. So the set $L_{F(e)} = \{\ulcorner\varphi(\overline{s_e(i)})\urcorner \mid i \in W_e\}$ is an infinite set, and by lemma 7.4.3 we do *not* have $PA \rhd_f PA + \{\varphi(\overline{s_e(i)}) \mid i \in W_e\}$.

QED

Corollary 7.4.4 *There is a total recursive function F such that for all e:*

> *W_e is infinite*
>
> *$\Leftrightarrow$ $F(e)$ codes a non-deterministic polynomial time Turing machine such that PA interprets PA + the set of formulas whose codes are in $L_{F(e)}$, but does not feasibly interpret this set.*

Proof. We can take the reduction F as in the proof of theorem 7.4.1. The $\leftarrow$-direction follows immediately from the old proof. For the $\rightarrow$-direction, we only have to remember the additional fact that $PA \vdash \varphi(\overline{n})$ for every n. QED

In section 7.6, we will replace the possibly infinite sets $L_{F(e)}$ by a single sentence $\xi(\overline{e})$ which is just as strong as far as feasible interpretability over PA is concerned. In order to do this we make use of two properties of $L_{F(e)}$ which make it suitable for replacement:

1. it is easy to compute whether $y \in L_{F(e)}$, thus, because we know already that PA is Δ_1^b-axiomatized, $PA + L_{F(e)}$ is Δ_1^b-axiomatized as well; and

2. $L_{F(e)}$ is only sparsely populated (see definition 7.2.6).

We now proceed to make these properties more precise and to prove them for our $L_{F(e)}$.

Remark 7.4.5 Note that, if we take some standard coding of Turing machines (see e.g. [BDG 87]), then to compute whether $t \in L_{F(e)}$ takes only time polynomial in $|t| + |e|$, for:

- to compute whether t is of the form $\ulcorner\varphi(\overline{s})\urcorner$ and, if it is, to find this s, takes time polynomial in $|t|$;

- to simulate the behavior of e on inputs $1, \dots, |s|$, for at most $|s|$ steps each takes time polynomial in $|s| + |e| \leq |t| + |e|$;

- to see whether $|s| > |s-1|$ takes time linear in $|s| \leq |t|$;

- to see whether e halts on some $i \leq |s|$ within $|s|$ but not within $|s| - 1$ steps takes time polynomial in $|s| + |e| \leq |t| + |e|$.

Remark 7.4.6 Note that for every e, the set $L_{F(e)}$ is sparse. More precisely, all members of $L_{F(e)}$ are of the form $\ulcorner\varphi(\overline{s})\urcorner$ for some s such that $|s| > |s-1|$ (i.e. there is a k such that $s = 2^k$).

Generalizations of theorem 7.4.1 By some slight adaptations, the proof of theorem 7.4.1 immediately gives rise to other results. We mention two directions of generalization.

- Sam Buss suggested the following restricted definition of feasible interpretability to us:

  $$\begin{aligned} U \rhd_{fm} V \leftrightarrow \exists K \exists M (\text{“}&K \textit{is an interpretation and } M \textit{is a determ.}\\ &\textit{pol. time Turing Machine''} \wedge \forall a(\alpha_V(a) \rightarrow \mathit{Prf}_U(M(a), a^K))). \end{aligned}$$

  This definition is more in line with the conventional use of the word “feasible” in the context of polynomial time computability. The clause $\mathit{Prf}_U(M(a), a^K)$ is a P-like formula, while the clause $\exists p(\text{“}|p| \leq P(|a|)\text{”} \wedge \mathit{Prf}_U(p, a^K))$ in the definition of feasible interpretability used in this dissertation is an NP-like formula.

  It is easily seen that under the new definition, $E' := \{e \mid e$ codes a deterministic Turing machine such that PA “feasibly” interprets PA + the set of formulas whose codes are in $L_e\} = \{e \mid \exists K\ \exists P$-time polynomial Turing machine $M\ \forall x \forall y(x$ is an axiom of PA or y is an accepting computation of e on $x \rightarrow \mathit{Prf}(M(x), x^K))\}$ is in Σ_2. Moreover, by inspection of the proofs of lemma 7.4.3 and theorem 7.4.1, we see that E' is in fact Σ_2-complete.

- We could also define feasible interpretability by bounding the length of proofs used in terms of other standard function classes than the polynomials; e.g. the linear functions or the exponential functions would be a good choice. In the latter case we have to adapt the fixed point of lemma 7.4.3 in order to diagonalize out of the function class, but we still get Σ_2-completeness of the set of theories “feasibly” interpretable over PA.

7.5 Lindström's general lemmas polynomialized

Let A be an extension of PA by a provably sparse set of axioms in the language of PA, where the set of axioms of A is given by the Σ_1^b-formula α. We give a definition from [Li 84], and we adapt some of the lemmas from that paper.

Definition 7.5.1 Let Γ be either Π or Σ. For every $i \geq 1$, we define the following:

$$[\Gamma_i]_\alpha(x, y) \ := \ \forall u, v \leq y(\mathit{Fmla}_{\Gamma_i}(u) \wedge \mathit{Prf}_{\alpha(z) \vee z = x}(v, u) \rightarrow \mathit{True}_{\Gamma_i}(u))$$

The following lemma corresponds to [Li 84, Lemma 1].

Lemma 7.5.2 *Let Γ be either Π or Σ. Then $[\Gamma_i]_\alpha(x, y)$ is a Γ_i-formula such that*

1. *$PA \vdash [\Gamma_i]_\alpha(x, y) \wedge z \leq y \rightarrow [\Gamma_i]_\alpha(x, z)$;*

2. *For every e and every φ, $A + \varphi(\bar{e}) \overset{|m|}{\underset{*}{\vdash}} [\Gamma_i]_\alpha(\ulcorner \varphi(\dot{e}) \urcorner, \bar{m})$*

3. *For every e and every φ, there is a polynomial P such that if $\psi \in \Gamma_i$ and $A + \varphi(\bar{e}) \vdash \psi$ via a proof coded by q, then $A + [\Gamma_i]_\alpha(\ulcorner \varphi(\bar{e}) \urcorner, \bar{q}) \overset{P(|q|)}{\longmapsto} \psi$.*

Proof.

1. See [Li 84, Lemma 1].

2. By lemma 7.2.8, we have

$$A + \varphi(\bar{e}) \overset{|m|}{\underset{*}{\vdash}} \forall u, v \leq \overline{m}(Prf_{\alpha(z) \vee z = \ulcorner \varphi(\bar{e}) \urcorner}(v, u) \rightarrow True_{|m|}(u)).$$

   Moreover, we have by lemma 7.2.10:

$$A \overset{|m|}{\underset{*}{\vdash}} \forall u(Fmla_{\Gamma_i}(u) \wedge True_{|m|}(u) \rightarrow True_{\Gamma_i}(u)),$$

   so indeed

$$A + \varphi(\bar{e}) \overset{|m|}{\underset{*}{\vdash}} [\Gamma_i]_\alpha(\ulcorner \varphi(\bar{e}) \urcorner, \overline{m}).$$

3. Suppose $\psi \in \Gamma_i$ and $A + \varphi(\bar{e}) \vdash \psi$, via a proof coded by q. Then there is a polynomial P_1 given in advance such that

$$A \overset{P_1(|q|)}{\vdash} Fmla_{\Gamma_i}(\ulcorner \psi \urcorner) \wedge Prf_{\alpha(z) \vee z = \ulcorner \varphi(\bar{e}) \urcorner}(\bar{q}, \ulcorner \psi \urcorner),$$

   thus by definition of $[\Gamma_i]_\alpha$, there is a polynomial P_2 given in advance such that

$$A + [\Gamma_i]_\alpha(\ulcorner \varphi(\bar{e}) \urcorner, \bar{q}) \overset{P_2(|q|)}{\vdash} True_{\Gamma_i}(\ulcorner \psi \urcorner).$$

   So by lemma 7.2.9, there is a polynomial P given in advance such that

$$A + [\Gamma_i]_\alpha(\ulcorner \varphi(\bar{e}) \urcorner, \bar{q}) \overset{P(|q|)}{\vdash} \psi.$$

QED

The fixed points that we define below in definition 7.5.3 were introduced by Lindström in his paper [Li 84]. Our lemmas 7.5.4 and 7.5.5 are analogous to [Li 84, lemma 2]. The difference is that we keep track of the lengths of the proofs.

Definition 7.5.3 Let $i \geq 1$ and $\chi \in \Sigma_i$ be given. Define ξ by diagonalization such that

$$PA \vdash \xi(\bar{e}) \leftrightarrow \exists y(\neg[\Pi_i]_\alpha(\ulcorner \xi(\bar{e}) \urcorner, y) \wedge \forall z \leq y\ \chi(\bar{e}, z)).$$

Dually, let $i \geq 1$ and $\chi \in \Pi_i$ be given. Define θ by diagonalization such that

$$PA \vdash \theta(\bar{e}) \leftrightarrow \forall y([\Sigma_i]_\alpha(\ulcorner \theta(\bar{e}) \urcorner, y) \rightarrow \chi(\bar{e}, y)).$$

Lemma 7.5.4 *If $\chi(x, y)$ is a Σ_i-formula, then $\xi(x)$ is also Σ_i and the following holds:*

1. *For all e, $A + \xi(\bar{e}) \overset{|m|}{\underset{*}{\vdash}} \forall z \leq \overline{m} \chi(\bar{e}, z)$.*

2. *For every e, there is a polynomial P such that if $\pi \in \Pi_i$ and $A + \xi(\bar{e}) \vdash \pi$ via a proof coded by q, then $A + \forall z \leq \bar{q}\ \chi(\bar{e}, z) \overset{P(|q|)}{\vdash} \pi$.*

Proof.

1. Take e fixed. By lemma 7.5.2(2) we have

$$A + \xi(\bar{e}) \vdash^{|m|}_{*} [\Pi_i]_\alpha(\ulcorner \xi(\bar{e}) \urcorner, \overline{m}),$$

so by lemma 7.5.2(1), we have

$$A + \xi(\bar{e}) \vdash^{|m|}_{*} \forall z \leq \overline{m}[\Pi_i]_\alpha(\ulcorner \xi(\bar{e}) \urcorner, z). \tag{7.8}$$

Now by definition of ξ,

$$A + \xi(\bar{e}) \vdash^{|m|}_{*} \exists y(\neg[\Pi_i]_\alpha(\xi(\bar{e}), y) \wedge \forall z \leq y\ \chi(\bar{e}, z)). \tag{7.9}$$

From (7.8) and (7.9), we finally conclude that

$$A + \xi(\bar{e}) \vdash^{|m|}_{*} \forall z \leq \overline{m}\chi(\bar{e}, z).$$

2. Suppose $\pi \in \Pi_i$ and

$$A + \xi(\bar{e}) \vdash \pi \text{ via a proof coded by } q, \tag{7.10}$$

then by lemma 7.5.2(3), we have a polynomial P_1 given in advance such that

$$A + [\Pi_i]_\alpha(\xi(\bar{e}), \bar{q}) \vdash^{P_1(|q|)} \pi. \tag{7.11}$$

By definition of ξ, we have a polynomial P_2 fixed in advance such that

$$A + \forall z \leq \bar{q}\ \chi(\bar{e}, z) \vdash^{P_2(|q|)} \xi(\bar{e}) \vee [\Pi_i]_\alpha(\xi(\bar{e}), \bar{q}). \tag{7.12}$$

From (7.10), (7.11) and (7.12) we conclude that there is a polynomial P given in advance such that $A + \forall z \leq \bar{q}\ \chi(\bar{e}, z) \vdash^{P(|q|)} \pi$.

QED

We state the next lemma for reasons of symmetry only: it will not be used in the sequel.

Lemma 7.5.5 *If $\chi(x, y)$ is a Π_i-formula, then $\theta(x)$ is also Π_i and the following holds:*

1. *For all e, $A + \theta(\bar{e}) \vdash^{|m|}_{*} \chi(\bar{e}, \overline{m})$.*

2. *For every e, there is a polynomial P such that if $\sigma \in \Sigma_i$ and $A + \theta(\bar{e}) \vdash^{|q|} \sigma$, then $A + \forall z \leq \bar{q}\ \chi(\bar{e}, z) \vdash^{P(|q|)} \sigma$.*

Proof. We leave the proof, which is similar to the proof of lemma 7.5.4, to the distrustful reader. QED

7.6 Feasible interpretability is Σ_2-complete

The result of section 7.4 is not yet entirely satisfactory: we seek an elegant result that corresponds more neatly to the idea that feasible interpretability is Σ_2-complete. In order to do this, we would like to be able to replace each possibly infinite set $L_{F(e)}$ of codes of axioms accepted by a Turing machine with code e by an instance $\xi(\bar{e})$ of a *single* formula ξ. Moreover, we would like this replacement to be such that $L_{F(e)}$ and $\xi(\bar{e})$ have the same status with respect to feasible interpretability over PA. Luckily, one of the fixed points of section 7.5 can do our job.

Definition 7.6.1 Let α be a Σ_1^b-formula defining the set of axioms of PA.

By remark 7.4.5, the relation $y \in L_{F(e)}$ is polynomial-time computable, so by [Bu 86, Theorem 3.2] it is Δ_1^b-definable in $I\Delta_0 + \Omega_1$. This means that there are a Σ_1^b-formula $\eta(e,y)$ and a Π_1^b-formula $\nu(e,y)$ that both correspond to the relation $y \in L_{F(e)}$ and such that $I\Delta_0 + \Omega_1 \vdash \eta(e,y) \leftrightarrow \nu(e,y)$.

Now we define $\chi(e,y) := \eta(e,y) \rightarrow \mathit{True}_{\Sigma_1}(y)$. Clearly χ is Σ_1^0. We define ξ as given by definition 7.5.3 for $i = 1$ and this χ. To be explicit, ξ is defined by diagonalization such that

$$PA \vdash \xi(\bar{e}) \leftrightarrow \exists y(\neg[\Pi_1]_\alpha(\ulcorner \xi(\bar{e}) \urcorner, y) \wedge \forall z \leq y\ \chi(\eta(\bar{e},z) \rightarrow \mathit{True}_{\Sigma_1}(z))).$$

Definition 7.6.2 For every e, let X_e be the set of $\Delta_0(exp)$-formulas whose Gödel numbers are contained in $L_{F(e)}$.

Lemma 7.6.3 *For all e, the following holds:*

1. $PA + \xi(\bar{e}) \rhd_f PA + X_e$;
2. $PA + X_e \rhd_{\Pi_1 f} PA + \xi(\bar{e})$.

Proof.

1. Suppose that $\varphi \in X_e$. Because $\eta \in \Sigma_1^b$, we have

$$PA \vdash^{|\varphi|}_{*} \eta(\bar{e}, \ulcorner \varphi \urcorner). \tag{7.13}$$

Next, by lemma 7.5.4.1, we have

$$PA + \xi(\bar{e}) \vdash^{|\varphi|}_{*} \eta(\bar{e}, \ulcorner \varphi \urcorner) \rightarrow \mathit{True}_{\Sigma_1}(\ulcorner \varphi \urcorner). \tag{7.14}$$

Combining (7.13) and (7.14), we derive $PA + \xi(\bar{e}) \vdash^{|\varphi|}_{*} \mathit{True}_{\Sigma_1}(\ulcorner \varphi \urcorner)$. Thus by lemma 7.2.9, $PA + \xi(\bar{e}) \vdash^{|\varphi|}_{*} \varphi$. So we certainly have $PA + \xi(\bar{e}) \rhd_f PA + X_e$.

2. Suppose $\pi \in \Pi_1$ and $PA + \xi(\bar{e}) \vdash \pi$ by a proof with Gödel number q. Then by lemma 7.5.4(2), we have a polynomial P_1 fixed in advance such that

$$PA + \forall z \leq \bar{q}\ \chi(\bar{e}, z) \vdash^{P_1(|q|)} \pi. \tag{7.15}$$

Also we have a polynomial P_2 given in advance such that

$$PA + X_e \vdash^{P_2(|q|)} \forall z \leq \bar{q}\ \chi(\bar{e}, z) \tag{7.16}$$

The reason for this is as follows. By a formalized version of remark 7.4.6, X_e provably contains only formulas $\varphi(\bar{s})$ where s is a power of 2, i.e.

$$PA \vdash \eta(\bar{e}, \bar{z}) \rightarrow \bigvee_{k \leq |z|} z = \ulcorner \varphi(2^k) \urcorner.$$

Moreover by Δ_1^b-completeness (see [Bu 86]) we have a polynomial P_3 such that for all n ($n \notin X_e \Rightarrow PA + X_e \overset{P_3(|n|)}{\vdash\!\!\!-\!\!\!-} \neg\eta(\bar{e}, \bar{n})$), so there is a polynomial P_4 given in advance such that

$$PA \overset{P_4(|q|)}{\vdash\!\!\!-\!\!\!-} \forall z(\eta(\bar{e}, z) \wedge z \leq \bar{q} \rightarrow \bigvee_{x \leq q \wedge x \in X_e} z = \bar{x}).$$

Also, $PA \overset{|n|}{\underset{*}{\vdash}} \varphi(\bar{n}) \rightarrow True_{\Sigma_1}(\ulcorner \varphi(\bar{n}) \urcorner)$, thus $PA + X_e \overset{|q|}{\underset{*}{\vdash}} \forall z \leq \bar{q}(\eta(\bar{e}, z) \rightarrow True_{\Sigma_1}(z))$.

From (7.15) and (7.16), we conclude that there is a polynomial P given in advance such that $PA + X_e \overset{P(|q|)}{\vdash\!\!\!-\!\!\!-} \pi$, as desired.

QED

Corollary 7.6.4 *For all e, the following holds:*

1. $PA + \xi(\bar{e}) \rhd PA + X_e$;

2. $PA + X_e \rhd PA + \xi(\bar{e})$.

Proof. Directly from lemma 7.6.3. Alternatively, see Lindström's original nonfeasible argument in [Li 93, lemma 5]. QED

Theorem 7.6.5 $\{e \mid PA \rhd_f PA + \xi(\bar{e})\}$ *is* Σ_2*-complete.*

Proof. By theorem 7.4.1, we have for all e: W_e is finite $\Longleftrightarrow PA \rhd_f PA + X_e$. By inspection of lemma 7.2.8 we see that for all e, $PA + X_e$ is feasibly essentially reflexive. Therefore corollary 7.3.3 gives for all e:

$$PA \rhd_f PA + X_e \Longleftrightarrow PA \rhd_{\Pi_1 f} PA + X_e.$$

Combining this with lemma 7.6.3 finally gives us for all e:

$$W_e \text{ is finite } \Longleftrightarrow PA \rhd_{\Pi_1 f} PA + \xi(\bar{e}),$$

thereby giving a reduction from the Σ_2-complete set $\{e \mid W_e \text{ is finite}\}$ to the set $\{e \mid PA \rhd_f PA + \xi(\bar{e})\}$. QED

Theorem 7.6.6 $\{e \mid PA \rhd PA + \xi(\bar{e}) \wedge \neg(PA \rhd_f PA + \xi(\bar{e}))\}$ *is* Π_2*-complete.*

Proof. It is easy to see that $PA \rhd PA + \xi(\bar{e}) \wedge \neg(PA \rhd_f PA + \xi(\bar{e}))$, being a conjunction of two Π_2-formulas, is again a Π_2-formula.

From corollary 7.4.4, we conclude that for all e, W_e is infinite if and only if

$$PA \rhd PA + X_e \wedge \neg(PA \rhd_f PA + X_e).$$

But by lemma 7.6.3 and corollary 7.6.4, we find that W_e is infinite if and only if $PA \rhd PA + \xi(\bar{e}) \wedge \neg(PA \rhd_f PA + \xi(\bar{e}))$. Thus we have reduced the Π_2-complete set $\{e \mid W_e \mathit{is\ infinite}\}$ to $\{e \mid PA \rhd PA + \xi(\bar{e}) \wedge \neg(PA \rhd_f PA + \xi(\bar{e}))\}$. QED

Generalizations of theorem 7.6.5 We can generalize theorem 7.6.5 on lines suggested at the end of section 7.4. However in this case it would not be a good choice to define feasible interpretability by bounding the length of proofs by linear functions.

Bibliography

[Ad 90] Z. Adamowicz. End-extending models of $I\Delta_0+EXP+B\Sigma_1$. *Fundamenta Mathematicae* 136 (1990) 133-145.

[Ad 93] Z. Adamowicz. A contribution to the end-extension problem and the Π_1 conservativeness problem. *Annals of Pure and Applied Logic* 61 (1993) 3-48.

[BDG 87] J.L. Balcázar, J. Díaz and J. Gabarró. *Structural Complexity I.* Springer Verlag, Berlin, 1987.

[Bek 89] L.D. Beklemishev. On the classification of propositional provability logics. *Math. USSR Izvestiya* 35 (1990) 247-275.

[Bek 91] L.D. Beklemishev. On bimodal provability logics for Π_1-axiomatized extensions of arithmetical theories. *ITLI Prepublication Series*, X-91-09, University of Amsterdam, Amsterdam (1991).

[Ben 62] J.H. Bennett. *On Spectra.* Ph.D. thesis, Princeton University, 1962.

[Ber 90] A. Berarducci. The interpretability logic of Peano Arithmetic. *The Journal of Symbolic Logic* 55 (1990) 1059-1089.

[BV 91] A. Berarducci and L.C. Verbrugge. On the metamathematics of weak theories. *ITLI Prepublication Series for Mathematical Logic and Foundations*, ML-91-02, University of Amsterdam, 1991.

[BV 93] A. Berarducci and R. Verbrugge. On the Provability Logic of Bounded Arithmetic. *Annals of Pure and Applied Logic* 61 (1993) 75-93.

[Bu 86] S.R. Buss. *Bounded Arithmetic.* Bibliopolis, edizioni di filosofia e scienze, Napoli, 1986.

[BS 90] S.R. Buss and P.J. Scott (editors). *Feasible Mathematics.* Birkhäuser, Boston, 1990.

[Co 64] A. Cobham. The intrinsic computational difficulty of functions. In: Y. Bar-Hillel (editor), *Logic, Methodology and Philosophy of Science II.* North-Holland, Amsterdam, 1964, pp. 24-30.

[Co 66] P.J. Cohen. *Set Theory and the Continuum Hypothesis.* Benjamin, New York, 1966.

[DPR 61] M. Davis, H. Putnam, and J. Robinson. The decision problem for exponential diophantine equations. *Annals of Mathematics* 74 (1961) 425-436.

[Di 80] C. Dimitracopoulos. *Matijasevič's Theorem and Fragments of Arithmetic.* Ph.D. thesis, Manchester University, 1980.

[Fe 60] S. Feferman. Arithmetization of metamathematics in a general setting. *Fund. Math.* 49 (1960) 35-92.

[FR 79] J. Ferrante and C.W. Rackoff. *The Computational Complexity of Logical Theories.* Springer-Verlag, Berlin, 1979.

[Gö 31] K. Gödel. Über formal unentscheidbare Sätze der *Principia Mathematica* und verwandter Systeme I. *Monatshefte Mathematik und Physik* 38 (1931) 173-198.

[GS 79] D. Guaspari and R.M. Solovay. Rosser sentences. *Annals of Mathematical Logic* 16 (1979) 81-99.

[Há 71] P. Hájek. On interpretability in set theories. *Commentationes Mathematicae Universitatis Carolinae* 12 (1971) 73-79.

[Há 72] P. Hájek. On interpretability in set theories II. *Commentationes Mathematicae Universitatis Carolinae* 13 (1972) 445-455.

[Há 79a] P. Hájek, Arithmetical hierarchy and complexity of computation. *Theoretical Computer Science* 8 (1979) 227-237.

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

[Há 83] P. Hájek. On a new notion of partial conservativity. In: E. Börger et al.(editors), *Logic Colloquium 83* vol. 2, Springer-Verlag, Berlin, 1983, pp. 217-232.

[HP 93] P. Hájek and P. Pudlák. *Metamathematics of First-order Arithmetic.* Springer-Verlag, Berlin, 1993.

[HS 84] A.L. Hayward and J.J. Sparkes. *The Concise English Dictionary.* Omega Books, Hertfordshire, 1984.

[Hi 1899] D. Hilbert. *Grundlagen der Geometrie.* Teubner, Leipzig, 1899.

[Ho 79] D. Hofstadter. *Gödel, Escher and Bach: An Eternal Golden Braid.* Harvester Press, Sussex, 1979.

[Jo 87] D.H.J. de Jongh. A simplification of a completeness proof of Guaspari and Solovay. *Studia Logica* 46 (1987) 187-192.

[JM 87] D.H.J. de Jongh and F. Montagna. Generic generalized Rosser fixed points. *Studia Logica* 46 (1987) 193-203.

[JM 88] D.H.J. de Jongh and F. Montagna. Provable fixed points. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik* 34 (1988) 229-250.

[JV 88] D.H.J. de Jongh and F. Veltman. *Intensional Logic.* Lecture notes, Philosophy Department, University of Amsterdam, Amsterdam, 1988.

[JV 90] D.H.J. de Jongh and F. Veltman. Provability logics for relative interpretability. In: P.P. Petkov (editor), *Mathematical Logic (Proceedings, Chaika, Bulgaria, 1988)*. Plenum Press, New York, 1990, pp. 31-42.

[JMM 91] D.H.J. de Jongh, M. Jumelet and F. Montagna. On the proof of Solovay's theorem. *Studia Logica* 50 (1991) 51-70.

[JM 91] D.H.J. de Jongh and F. Montagna. Rosser orderings and free variables. *Studia Logica* 50 (1991), pp.71-80.

[Ka 89] M. B. Kalsbeek. An Orey Sentence for Predicative Arithmetic. *ITLI Prepublication Series*, X-89-01, University of Amsterdam, Amsterdam, 1989.

[KH 82] C.F. Kent and B.R. Hodgson. An arithmetical characterization of NP. *Theoretical Computer Science* 21 (1982) 255-267.

[KP 89] J. Krajíček and P. Pudlák. On the structure of initial segments of models of arithmetic. *Archives of Mathematical Logic* 28 (1989) 91-98.

[KPT 89] J. Krajíček, P. Pudlák and G. Takeuti. Bounded arithmetic and the polynomial hierarchy. *Annals of Pure and Applied Logic* 52 (1991) 143-153.

[Kr 51] G. Kreisel. On the interpretation of non-finitist proofs I. *Journal of Symbolic Logic* 16 (1951) 241-267.

[Kr 52] G. Kreisel. On the interpretation of non-finitist proofs II. *Journal of Symbolic Logic* 17 (1952) 43-58.

[Kr 58] G. Kreisel. Mathematical significance of consistency proofs. *Journal of Symbolic Logic* 23 (1958) 155-182.

[Ku 80] K. Kunen. *Set Theory: An Introduction to Independence Proofs.* North-Holland, Amsterdam, 1980.

[Li 79] P. Lindström. Some results on interpretability. In: F.V. Jensen et al. (editors), *Proceedings from the 5th Scandinavian Logic Symposium.* Aalborg University Press, Aalborg, 1979, pp. 329-361.

[Li 84] P. Lindström. On partially conservative sentences and interpretability. *Proceedings of the American Mathematical Society*, vol. 91 (1984), pp. 436-443.

[Li 93] P. Lindström. On Σ_1 and Π_1 sentences and degrees of interpretability. *Annals of Pure and Applied Logic* 61 (1993) 175-193.

[Lö 55] M.H. Löb. Solution of a problem of Leon Henkin. *Journal of Symbolic Logic* 20 (1955) 115-118.

[MS 73] A. Macintyre and H. Simmons. Gödel's diagonalization technique and related properties of theories. *Colloquium Mathematicum* 28 (1973) 165-180.

[MA 78] K. Manders and L. Adleman. NP-complete decision problems for binary quadratics. *Journal of Computer System Sciences* 16 (1978) 168-184.

[Ma 70] Yu. V. Matijasevich. Enumerable sets are Diophantine. *Doklady Akad. Nauk SSSR* 191 (1970) 279-282 (in Russian. English translation: *Soviet Math. Doklady* 1970, 354-357.)

[Mi 71] G.E. Mints. Quantifier-free and one-quantifier systems. In: Yu. V. Matijasevich and A.O. Slisenko (editors), *Zapiski Nauchnykh Seminarov* 20 (1971) 115-133 (in Russian. English translation: *Journal of Soviet Mathematics* 1 (1973) 211-226.)

[Mo 65] R. Montague. Interpretability in terms of models. *Indagationes Mathematicae* 27 (1965) 467-476.

[Ne 86] E. Nelson. *Predicative Arithmetic.* Math. Notes 32, Princeton University Press, Princeton, 1986.

[Od 89] P. Odifreddi. *Classical Recursion Theory.* North-Holland, Amsterdam, 1989.

[Or 61] S. Orey. Relative interpretations. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik* 7 (1961) 146-153.

[Pa 71] R. Parikh. Existence and feasibility in arithmetic. *Journal of Symbolic Logic* 36 (1971) 494-508.

[PD 82] J.B. Paris and C. Dimitracopoulos. Truth definition for Δ_0 formulae. In: *Logic and Algorithmic*, Monographie No. 30 de L'Enseignement Mathématique. Université de Genève, Genève, 1982, pp. 317-329.

[Pa 72] C. Parsons. On n-quantifier induction. *Journal of Symbolic Logic* 37 (1972) 466-482.

[Pu 83] P. Pudlák. A definition of exponentiation by a bounded arithmetical formula. *Commentationes Mathematicae Universitatatis Carolinae* 24 (1983) 667-671.

[Pu 85] P. Pudlák. Cuts, consistency statements and interpretations. *Journal of Symbolic Logic* 50 (1985) 423-441.

[Pu 86] P. Pudlák. On the length of proofs of finitistic consistency statements in first order theories. In: J.B. Paris et al. (editors), *Logic Colloquium '84*, North Holland, Amsterdam, 1986, pp. 165-196.

[Pu 87] P. Pudlák. Improved bounds to the length of proofs of finitistic consistency statements. In: S.G. Simpson (editor), *Logic and Combinatorics*, Contemporary Mathematics 35. AMS, Providence, 1987, pp. 309-332.

[Rog 67] H. Rogers Jr. *Theory of Recursive Functions and Effective Computability.* McGraw-Hill, New York, 1967.

[Ros 36] J.B. Rosser. Extensions of some theorems of Gödel and Church. *Journal of Symbolic Logic* 1 (1936) 87-91.

[Sh 88] V. Yu. Shavrukov. The logic of relative interpretability over Peano Arithmetic (in Russian). Preprint No. 5, Steklov Mathematical Institute, Moscow, 1988.

[Sm 77] C. Smoryński. The incompleteness theorems. In: J. Barwise (editor), *Handbook of Mathematical Logic.* North Holland, Amsterdam, 1977, pp. 821-865.

[Sm 85] C. Smoryński. *Self-reference and Modal Logic.* Springer-Verlag, New York (1985).

[So 87] R.I. Soare. *Recursively Enumerable Sets and Degrees: a Study of Computable Functions and Computably Generated Sets.* Springer-Verlag, Berlin, 1987.

[So 76] R.M. Solovay. Provability interpretations of modal logic. *Israel Journal of Mathematics* 25 (1976) 287-304.

[So 76b] R. M. Solovay. *On interpretability in set theories.* Manuscript, 1976.

[So 89] R.M. Solovay. Injecting inconsistencies into models of *PA*. *Annals of Pure and Applied Logic* 44 (1989) 261-302.

[St 76] L.J. Stockmeyer. The polynomial-time hierarchy. *Theoretical Computer Science* 3 (1976) 1-22.

[Šv 83] V. Švejdar. Modal analysis of generalized Rosser sentences. *Journal of Symbolic Logic* 48 (1983) 986-999.

[Ta 75] G. Takeuti. *Proof Theory.* North Holland, Amsterdam, 1975.

[Ta 88] G. Takeuti. Bounded arithmetic and truth definition. *Annals of Pure and Applied Logic* 39 (1988) 75-104.

[TMR 53] A. Tarski, A. Mostowski and R. Robinson. *Undecidable Theories.* North-Holland, Amsterdam, 1953.

[TvD 88] A.S. Troelstra and D. van Dalen. Constructivism in Mathematics: An Introduction. North-Holland, Amsterdam, 1988.

[Ve 88] L.C. Verbrugge. *Does Solovay's Completeness Theorem Extend to Bounded Arithmetic?.* Master's thesis, University of Amsterdam, Amsterdam, 1988.

[Ve 89] L.C. Verbrugge. Σ-completeness and bounded arithmetic. *ITLI Prepublication Series for Mathematical Logic and Foundations*, ML-89-05, University of Amsterdam, Amsterdam, 1989.

[Ve 93] L.C. Verbrugge. Feasible interpretability. In: P. Clote and J. Krajíček (editors), *Arithmetic, Proof Theory and Computational Complexity.* Oxford University Press, Oxford, 1993.

[VV] L.C. Verbrugge and A. Visser. A small reflection principle for bounded arithmetic. To appear in *Journal of Symbolic Logic.*

[Vi 81] A. Visser. *Aspects of Diagonalization & Provability.* Ph.D. thesis, University of Utrecht, 1981.

[Vi 82] A. Visser. On the completeness principle. *Annals of Mathematical Logic* 22 (1982) 263-295.

[Vi 85] A. Visser. Evaluation, provably deductive equivalence in Heyting's Arithmetic of substitution instances of propositional formulas. *Logic Group Preprint Series* No. 4, University of Utrecht, Utrecht, 1985.

[Vi 89] A. Visser. Peano's smart children: a provability logical study of systems with built-in consistency. *Notre Dame Journal of Formal Logic* 30 (1989) 161-196.

[Vi 90a] A. Visser. Interpretability Logic. In: P.P. Petkov (editor), *Mathematical Logic (Proceedings, Chaika, Bulgaria, 1988)*. Plenum Press, New York, 1990, 1990.

[Vi 90b] A. Visser. *Proofs of* Π_2*-completeness by Per Lindström and Robert Solovay, as told by Albert Visser*. Unpublished manuscript, 1990.

[Vi 91a] A. Visser. The formalization of interpretability. *Studia Logica* 50 (1991) 81-105.

[Vi 91b] A. Visser. On the Σ_1^0-conservativity of Σ_1^0- completeness. *Notre Dame Journal of Formal Logic* 32 (1991) 554-561.

[Vi 92] A. Visser. An inside view of EXP; or, The closed fragment of the provability logic of $I\Delta_0 + \Omega_1$ with a propositional constant for EXP. *Journal of Symbolic Logic* 57 (1992) 130-165.

[Vi 93] A. Visser. The unprovability of small consistency. *Archives of Mathematical Logic* 32 (1993) 275-298.

[Vi b] A. Visser. *Questiones Longae et Breves*. Unpublished cumulative manuscript.

[VH 72] P. Vopĕnka and P. Hájek. *The Theory of Semisets*. North-Holland, Amsterdam, 1972.

[WP 87] A.J. Wilkie and J.B. Paris. On the scheme of induction for bounded arithmetic formulas. *Annals of Pure and Applied Logic*, vol. 35 (1987) 261-302.

[WP 89] A.J. Wilkie and J.B. Paris. On the existence of end extensions of models of bounded induction. In: J.E. Fenstad et al. (editors), *Logic, Methodology and Philosophy of Science VIII*, North-Holland, Amsterdam, 1989.

[Wr 76] C. Wrathall. Complete sets and the polynomial time hierarchy. *Theoretical Computer Science* 3 (1976) 23-33.

[Za 93] D. Zambella. On second order bounded arithmetic. Manuscript, 1993.

Samenvatting

Dit proefschrift bevat een aantal resultaten over de metamathematica van eerste orde rekenkunden. Het zwaartepunt ligt bij het bestuderen van bewijsbaarheid voor begrensde rekenkunde en een alternatieve definitie van interpreteerbaarheid.

Deel I gaat vooraf aan de eigenlijke resultaten.

In het inleidende hoofdstuk 1 geven we een informele beschrijving van de rol van het begrip "efficiëntie" in de wiskunde en de metamathematica. We introduceren ook de belangrijkste begrippen die in het proefschrift aan de orde komen: complexiteitstheorie, begrensde rekenkunde, bewijsbaarheidslogica en interpreteerbaarheidslogica.

Hoofdstuk 2 bevat de technische beschrijving van de in hoofdstuk 1 geïntroduceerde begrippen. Daarnaast geven we een beknopte opsomming van de stellingen uit de literatuur die we bij het bewijzen van onze resultaten gebruikt hebben. Zo geven we enkele stellingen over definieerbare sneden en hun toepassingen in de rekenkunde. Het hoofdstuk eindigt met een paragraaf waarin resultaten uit de literatuur besproken worden die de verschillen en overeenkomsten tussen enkele zwakke rekenkundige theorieën belichten. Interpreteerbaarheid en conservativiteit voor bepaalde klassen formules worden hier gebruikt om de kracht van de theorieën te vergelijken.

Deel II is gewijd aan de begrensde rekenkunde.

In hoofdstuk 3 bewijzen we eerst onder de complexiteitstheoretische aanname $NP \neq co\text{-}NP$ dat de begrensde rekenkunde geen volledigheid bewijst voor alle formules voor het vergelijken van getuigen. In de Peano Rekenkunde speelt bewijsbare volledigheid voor zulke formules een belangrijke rol bij het bewijzen van de geformaliseerde versie van Rossers Stelling en Solovay's Volledigheidsstelling.

Om toch ook in de begrensde rekenkunde de geformaliseerde versie van Rossers Stelling te kunnen afleiden, bewijzen we een reflectieprincipe voor "kleine" bewijzen. Het bewijs daarvan maakt gebruik van partiële waarheidspredicaten en definieerbare sneden.

Als toepassing van dit principe geven we een bewijs van een stelling van Bernardi en Montagna voor de begrensde rekenkunde. Bovendien gebruiken we het "kleine" reflectieprincipe voor een simpele versterking van een bekende stelling over het injecteren van kleine bewijzen van inconsistentie. Tenslotte gebruiken we het principe, op een meer verrassende manier, in het bewijs van een stelling over het bestaan van echte eindextensies van modellen van de begrensde rekenkunde die aan een zware extra eis voldoen.

In hoofdstuk 4 keren we terug naar het probleem van bewijsbare volledigheid. We bewijzen dat de complexiteitstheoretische aanname $P \neq NP \cap co\text{-}NP$ impliceert dat Buss' begrensde rekenkunde S_2^1 niet voor alle Σ_1^0-zinnen volledigheid bewijst.

In hoofdstuk 5 presenteren we partiële antwoorden op de vraag: wat is de bewijsbaarheidslogica van de begrensde rekenkunde? Omdat bewijsbare volledigheid voor zinnen voor het vergelijken van getuigen op grond van resultaten uit hoofdstuk 3 en 4 dubieus

is, kunnen we niet rechtstreeks Solovay's methode gebruiken.

Met behulp van het kleine reflectieprincipe uit hoofdstuk 3 en definieerbare sneden passen we voor een geschikte klasse van Kripkeframes de methode van Solovay aan. We geven een inbedding van modellen op zulke eenvoudige frames in de begrensde rekenkunde.

Ook bewijzen we dat de bewijsbaarheidslogica van de begrensde rekenkunde in ieder geval niet de modale theorie van een klasse Kripkebomen kan zijn. De vraag wat de bewijsbaarheidslogica van de begrensde rekenkunde dan wel is, is op het moment van schrijven voor zover bekend nog open.

Deel III behandelt een alternatieve definitie van interpreteerbaarheid.

In hoofdstuk 6 definiëren we "uitvoerbare interpreteerbaarheid," waarbij de lengte van bewijzen van vertaalde axioma's begrensd is door een polynoom in de lengte van die axioma's zelf. We laten zien dat een aantal bekende interpretaties, zoals die van $ZF + \mathbf{V} = \mathbf{L}$ in ZF, uitvoerbaar zijn. Aan de andere kant zijn niet alle interpretaties te vervangen door uitvoerbare interpretaties. Met behulp van diagonalisatie construeren we een theorie die weliswaar in de Peano Rekenkunde interpreteerbaar is, maar er niet op uitvoerbare wijze in geïnterpreteerd kan worden.

Verder laten we zien dat de interpreteerbaarheidslogica ILM arithmetisch correct en volledig is voor uitvoerbare interpreteerbaarheid over de Peano Rekenkunde.

Hoofdstuk 7 behandelt de definitionele complexiteit van uitvoerbare interpreteerbaarheid over de Peano Rekenkunde. We bewijzen, door een recursie-theoretische reductie te combineren met een aangepaste versie van een methode van Lindström waarin partiële waarheidsdefinities een belangrijke rol spelen, dat uitvoerbare interpreteerbaarheid over de Peano Rekenkunde Σ^0_2-volledig is. En passant geven we een karakterisering van uitvoerbare interpreteerbaarheid in de stijl van Orey en Hájek.

De Σ^0_2-volledigheid van uitvoerbare interpreteerbaarheid over de Peano Rekenkunde staat in contrast met de Π^0_2-volledigheid van standaard interpreteerbaarheid over de Peano Rekenkunde. Het blijkt dat standaard interpreteerbaarheid en uitvoerbare interpreteerbaarheid substantieel verschillende extensies hebben. We bewijzen dat de verzameling zinnen die wel gewoon maar niet uitvoerbaar interpreteerbaar is over de Peano Rekenkunde, zelfs Π^0_2-volledig is.