Contributions to the Metamathematics of Arithmetic

ACTA PHILOSOPHICA GOTHOBURGENSIA 30

Contributions to the Metamathematics of Arithmetic

Fixed Points, Independence, and Flexibility

Rasmus Blanck

UNIVERSITY OF GOTHENBURG
ACTA UNIVERSITATIS GOTHOBURGENSIS

Thesis submitted for the Degree of Doctor of Philosophy in Logic
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg

ISBN 978-91-7346-917-3 (PRINT)
ISBN 978-91-7346-918-0 (PDF)
ISSN 0283-2380

The publication is also available in full text at:
http://hdl.handle.net/2077/52271

Distribution:
ACTA UNIVERSITATIS GOTHOBURGENSIS
Box 222, 405 30 Göteborg, Sweden
acta@ub.gu.se

Typeset in Adobe Garamond Pro using XeLaTeX
Cover design by Peter Johnsen

Printed by Ineko, Kållered 2017

Abstract

Title:	Contributions to the Metamathematics of Arithmetic: Fixed Points, Independence, and Flexibility
Author:	Rasmus Blanck
Language:	English (with a summary in Swedish)
Department:	Philosophy, Linguistics and Theory of Science
Series:	Acta Philosophica Gothoburgensia 30
ISBN:	978-91-7346-917-3 (print)
ISBN:	978-91-7346-918-0 (pdf)
ISSN:	0283-2380
Keywords:	arithmetic, incompleteness, flexibility, independence, non-standard models, partial conservativity, interpretability

This thesis concerns the incompleteness phenomenon of first-order arithmetic: no consistent, r.e. theory T can prove every true arithmetical sentence. The first incompleteness result is due to Gödel; classic generalisations are due to Rosser, Feferman, Mostowski, and Kripke. All these results can be proved using self-referential statements in the form of provable fixed points. Chapter 3 studies sets of fixed points; the main result is that disjoint such sets are creative. Hierarchical generalisations are considered, as well as the algebraic properties of a certain collection of bounded sets of fixed points. Chapter 4 is a systematic study of independent and flexible formulae, and variations thereof, with a focus on gauging the amount of induction needed to prove their existence. Hierarchical generalisations of classic results are given by adapting a method of Kripke's. Chapter 5 deals with end-extensions of models of fragments of arithmetic, and their relation to flexible formulae. Chapter 6 gives Orey-Hájek-like characterisations of partial conservativity over different kinds of theories. Of particular note is a characterisation of partial conservativity over $\mathrm{I}\Sigma_1$. Chapter 7 investigates the possibility to generalise the notion of flexibility in the spirit of Feferman's theorem on the 'interpretability of inconsistency'. Partial results are given by using Solovay functions to extend a recent theorem of Woodin.

Acknowledgements

Writing a thesis throws you from joy to despair, and hopefully back again. This is to express my gratitude to all of you who have contributed to the joyous side of the process: colleagues, friends, family, and students.

I have benefited from having many thesis advisors over the years: Ali Enayat, Christian Bennet, Dag Westerståhl, and Fredrik Engström. This would never have been possible without you. Thank you for the effort, time, and belief you put in me.

Ali, when you first came to the department, I was suffering from a motivational dip and had almost given up on logic. You remedied this by inviting me to work with you, even before you formally became my advisor. Thank you for being such an inspiration to me, and for your overwhelming generosity and patience. Christian, thank you for starting all this when I first set foot in the old Philosophy department years ago; the path has not been straight, but I hope the apple hasn't fallen too far from the tree. Dag, thank you for making it possible to start my graduate studies in Göteborg. Fredrik, thank you for steady guidance, and for your ability to ask exactly the right questions at the right time.

A more collective thank-you goes to the members of the logic group, and to the participants of the logic seminar at the department of Philosophy, Linguistics and Theory of Science. I'd also like to mention the reading group on models of arithmetic that Saeideh Bahrami and Zach McKenzie organised during their visit to the department.

Martin Kaså, your friendship is invaluable to me; I have no idea how I could ever have endured these years without your regular knocks on my door. Peter Johnsen, thank you for the beautiful cover design of this book.

It's been a pleasure sharing an office with a number of other graduate students, in rough order of appearance: Martin Filin Karlsson, Stellan Petersson, Pia Nordgren, Erik Joelsson, and Alla Choifer. Thank you for making office hours (and evenings, and weekends) much more enjoyable.

Costas Dimitracopoulos, thank you for agreeing to be the external reviewer of this thesis, and for your help with spotting a number of misprints in an earlier version of the manuscript.

Among international colleagues, I am grateful to Albert Visser, Taishi Kurahashi, Volker Halbach, and Volodya Shavrukov for expressing interest in and commenting on my work. Volodya has also gracefully allowed me to include one of his unpublished results in Chapter 7.

I have been dependent on scholarships to fund my graduate studies, and therefore I wish to acknowledge generous financial support from the following foundations:

Stiftelsen Anna Ahrenbergs fond för vetenskapliga m.fl. ändamål, *Kungliga och Hvitfeldtska stiftelsen*, *Adlerbertska stipendiestiftelsen*, *Stiftelsen Paul och Marie Berghaus donationsfond*, *Stiftelsen Henrik Ahrenbergs studiefond*, and *Bertil Settergrens fond*.

There is a life outside the department too. Without my friends in the band Räfven I might have finished this thesis on time, or perhaps not at all. You've brought me to far more places around the world that I could ever expect and given me much energy and inspiration.

I would also very much like to thank Niklas Rudbäck and Per Malm, for our writing retreats at Näs and Grönskhult, and for your continuous reminder of the elm/beech distinction; Erik Börjeson, for our hiking trips and for many other distractions; my parents Eva and Hans, for supporting me in oh so many ways.

Jonna, my dear. I believe that you have suffered most during my periods of hard work, head in the clouds. Your support and understanding seem endless. Therefore, my most heartfelt thanks go to you.

Göteborg, April 2017
Rasmus Blanck

Contents

1 Introduction 1
1.1 Scope, theme, and topics 2
1.2 About this thesis 6

2 Background 9
2.1 Notation and conventions 9
2.2 Arithmetised meta-arithmetic 13
2.3 Model theory of arithmetic 20
2.4 Recursion theory 25

3 Sets of fixed points 31
3.1 Recursion theoretic complexity 32
3.2 Counting the number of fixed points 34
3.3 Hierarchical generalisations 35
3.4 Algebraic properties 37

4 Flexibility in fragments 41
4.1 Definitions and motivation 41
4.2 Mostowski's and Kripke's theorems 44
4.3 Flexibility and independence in Robinson's arithmetic 47
4.4 Refinements 49
4.5 Scott's lemma and Lindström's proof 52
4.6 Chaitin's incompleteness theorem 55

5 Formalisation and end-extensions 57
5.1 Formalisation of Kripke's theorem 57
5.2 Formalisation of the GRMMKV theorem 60
5.3 Hierarchical generalisations 63

6 Characterisations of partial conservativity 65
6.1 The Orey-Hájek characterisation and its extensions 66
6.2 A characterisation of partial conservativity over $I\Sigma_1$ 68
6.3 Language extensions 70
6.4 Theories that are not recursively enumerable 72

7 Uniformly flexible formulae and Solovay functions . . 75
7.1 Woodin's theorem and its extensions 76
7.2 Digression: On coding schemes 81
7.3 Uniformly flexible formulae 84
7.4 Partial results on uniformly flexible Σ_1 formulae 87
7.5 Hierarchical generalisations: Asking the right question 92

8 Concluding remarks 95

References 97

Sammanfattning på svenska 105

1 Introduction

A major insight of mathematical logic is that truth and provability are complicated concepts. This thesis aims to contribute to the study of the intricate relationship between truth and provability in formal theories suitable for describing the natural numbers 0, 1, 2, 3, ...

The single most influential technical result describing this relationship is Gödel's first incompleteness theorem. Pick any formal system that is free of contradiction, and for which there is an effective procedure to decide whether a given sentence is an axiom of the system or not. If it is possible to carry out a certain amount of elementary arithmetic within this system, then it is also possible to construct a sentence pertaining to natural numbers that is true but impossible to prove in the system.[1] The proof of the first incompleteness theorem can be paraphrased by appealing to the classic liar paradox. Consider the sentence

> This sentence isn't true.

If that sentence were true, it would truthfully claim its own falsehood – but then the sentence would be false. If the sentence is false, then it falsely asserts its own falsehood, and must therefore be true. Hence, no truth value can be ascribed to the sentence without giving rise to a contradiction.

In formal theories of arithmetic, this observation amounts to a proof that the concept of arithmetical truth is not definable in arithmetic. On the other hand the concept of provability within a fixed system T *is* definable in arithmetic, which allows for the construction of an arithmetical counterpart of

> This sentence isn't provable within T.

[1] (Gödel, 1931). The reader familiar with Gödel's incompleteness theorem is already uneasy with the way that this famous result is paraphrased here. Fear not: technical details abound in Chapter 2.

To actually construct such a sentence in elementary arithmetic is an impressive technical feat, also due to Gödel. If the sentence above is false, then it falsely claims its own unprovability in T. Therefore the sentence must be provable in T. If T only proves true sentences, then the sentence must be true. But then the sentence truthfully claims its own unprovability in T, and is therefore true and unprovable in T.

The argument leading up to the true-but-unprovable sentence is different from that of the liar paradox, in the respect that it does not lead to a contradictory statement. It simply exhibits one aspect of the complicated relationship between truth and provability in formal theories of arithmetic. In effect: no arithmetically definable formal theory of arithmetic can be complete in the sense that it proves all and precisely all true arithmetical sentences.

The study of incompleteness phenomena is no longer in the mainstream of mathematical logic. (And logic is still not in the mainstream of neither mathematics nor philosophy.) This does not, however, mean that all the important problems of the field have been settled. The central parts of this thesis study incompleteness phenomena for their own sake, in an attempt to further the knowledge in the field. The question guiding the research reported in this thesis has been:

> [W]hat more can we say about systems of arithmetic than that they are all incomplete? (Hájek and Pudlák, 1993, p. 3)

A more philosophically inclined researcher may perhaps want to investigate *why* formal arithmetical theories must fail in describing what we expect them to describe. Yet another researcher, with concrete applications in mind, may instead want to ask *when* incompleteness phenomena matter. While these are interesting questions (and perhaps even *more* so than the guiding question stated above), they do not fall within the scope of this thesis.

1.1 Scope, theme, and topics

This thesis concerns the incompleteness phenomena of formal, first-order theories of arithmetic, and the following paragraphs delineate the scope,

theme, and topics treated. The point of departure for this thesis is Gödel (1931), where the incompleteness theorems are presented for the first time. The first incompleteness theorem states that for any ω-consistent, r.e. extension T of formal number theory, there is a proposition undecidable in that theory, in the sense that this proposition is neither provable nor refutable in T, while the second incompleteness theorem states that the formalised consistency statement of T, $\mathrm{Con_T}$, is an example of such a proposition. Rosser's (1936) generalisation of the first incompleteness theorem weakens the assumption of ω-consistency to that of mere consistency.

An important method used in the proofs of the aforementioned incompleteness results, and in many proofs in this thesis, is that of constructing self-referential sentences. The existence of such sentences is guaranteed by the diagonal lemma, stating that for every arithmetical formula $\phi(x)$, and every theory T satisfying some reasonable assumptions, there is a sentence δ which is provably equivalent with $\phi(\ulcorner\delta\urcorner)$ in T. Hence every formula is guaranteed to have at least one *provable fixed point* in this sense. Here, this method is studied from slightly different perspective than usual, by considering the collection of fixed points of a given formula. It is an easy corollary to the proof of the diagonal lemma that every formula has infinitely many syntactically distinct fixed points, inspiring the question:

> What more can be said about the collection of syntactically distinct fixed points of a formula than that it is infinite?

This question is recently treated by Halbach and Visser (2014). The main result presented in this thesis is that every such collection of provable fixed points is creative, in the recursion theoretic sense. Hierarchical generalisations are also considered.

In his 1961 paper, Mostowski introduced the class of *independent formulae*: such a formula has exactly one free variable, and the property that the only propositional combinations of its instances that are provable in T are the tautologies.[2] The existence of such a formula is a generalisation of the first incompleteness theorem. Almost simultaneously, Kripke, in his

[2]Mostowski calls these formulae *free*, but this terminology seems to have fallen out of style almost immediately.

1962 paper (which was submitted some weeks before Mostowski's paper) defined the concept of *flexible* formulae: in Kripke's words, formulae such that 'their extensions as sets are left undetermined by the formal system'. He showed that a flexible formula exists, and that every flexible formula is also independent. Hence, the existence of a flexible formula is in turn a generalisation of Mostowski's generalisation of the first incompleteness theorem.

Feferman (1960) obtained a generalisation of the second incompleteness theorem, showing that not only is $\mathrm{Con}_{\mathrm{T}}$ undecidable in T under reasonable assumptions on T, $\mathrm{T} + \neg\mathrm{Con}_{\mathrm{T}}$ is even *interpretable* in T under the same assumptions. Here, an interpretation is taken as a means of redefining the notions of the former theory in such a way that every theorem of the former theory becomes provable in the latter.

The research reported in this thesis attempts to generalise these generalisations of the incompleteness theorems in a number of ways. One kind of generalisation is to scrutinise what the 'reasonable assumptions' on the formal theories are, and one way of obtaining such generalisations is to consider how much mathematical induction is needed to prove the existence of independent and flexible formulae. Many results in the literature on flexible formulae are stated only for extensions of PA, while it is evident that the assumption that T extends PA is unnecessarily strong. Fine-tuning the amount of induction needed for the existence proofs forms a part of the study of *fragments of arithmetic*, and this line of generalisation is initiated in Chapter 4, and continued in part in Chapter 5.

Another way to generalise the incompleteness theorems is to consider not only r.e. extensions of formal arithmetic, but theories defined by formulae of higher complexity. The first published result of this kind is due to Jeroslow (1975), who showed that every consistent extension of arithmetic whose set of theorems is Δ_2-definable is still Π_1-incomplete, even though there are such extensions that prove their own consistency. For an investigation of such self-supporting theories, see Kaså (2012). In two recent papers (Kikuchi and Kurahashi, 20xx; Salehi and Serahi, 2016) the first incompleteness theorem is generalised to show that for every Σ_{n+1}-definable, Σ_n-sound extension of arithmetic T, there is a true Π_{n+1} sentence that is

undecidable in T. A further generalisation is obtained here, by showing that similar results hold for independent and flexible formulae as well.

A grand generalisation along the lines of Feferman's result would be to show that not only are there independent and flexible formulae, but also that the independence and flexibility is somehow interpretable in arithmetic. Put formally, a formula $\gamma(x)$ is Σ_n-flexible over T iff, for every Σ_n formula $\sigma(x)$, the theory $\mathrm{T} + \forall x(\gamma(x) \leftrightarrow \sigma(x))$ is consistent. This means that the extension of a flexible formula can consistently be claimed to coincide with the extension of any Σ_n formula. The goal would then be to show the 'interpretability of flexibility' in the sense that, with $\gamma(x)$ and $\sigma(x)$ as above, $\mathrm{T} + \forall x(\gamma(x) \leftrightarrow \sigma(x))$ is interpretable in T. Partial results of this kind are given.

To fully appreciate the nature of the partial results alluded to above, it is necessary to take non-standard models of arithmetic into consideration. Recall that a formula is flexible if its extension as a set is left undetermined by the formal system at hand. This means that the theory obtained by adding to T the sentence $\forall x(\gamma(x) \leftrightarrow \sigma(x))$ is consistent. By the completeness theorem for first-order logic, there is then a model of this augmented theory. By the nature of models of first-order arithmetic, every such model is an *end-extension* of the standard model of arithmetic $\mathbb{N}$. The syntactical notion of interpretability can be characterised by the semantical notion of end-extendability: for any two consistent r.e. theories T, S extending PA, S is interpretable in T iff every model of T can be end-extended to a model of S. This characterisation is discussed in some detail in Chapter 6, where also a number of extensions of the Orey-Hájek-Guaspari-Lindström characterisation are established. Of particular note is a version of the OHGL characterisation for extensions of $\mathrm{I}\Sigma_1$.

In light of the characterisation of interpretability, it makes sense to ask: even if not every model of T can be end-extended to a model of S, can there be *some* models of T having such end-extensions? In Chapter 5 it is shown that this is indeed the case, in particular, there is a Σ_{n+1} formula $\gamma(x)$ such that for every $\sigma(x) \in \Sigma_{n+1}$, every model of $\mathrm{T} + \mathrm{Con}_{\mathrm{T}}$ can be end-extended to a model of $\mathrm{T} + \forall x(\gamma(x) \leftrightarrow \sigma(x))$. This result is in turn extended to encompass many of the refinements given in Chapter 4.

Woodin (2011) establishes the existence of an r.e. set W_e with the following properties: W_e is empty in the standard model, and if $\mathcal{M}$ is a countable model of PA, and if s is an $\mathcal{M}$-finite set that extends W_e, then there is an end-extension of $\mathcal{M}$ in which $W_e = s$. This result has a flavour of independence in Mostowski's sense, flexibility in Kripke's sense, as well as of interpretability as discussed in the preceding paragraphs. In Chapter 7, it is shown that the countability assumption on $\mathcal{M}$ can be removed, hence establishing an interpretability result in the spirit of Feferman, but only for these 'finitely flexible' formulae. Moreover, it is shown that if the restriction to countable models is kept, then Woodin's result holds true for extensions of $\mathrm{I}\Sigma_1$, by using the extended version of the OHGL characterisation.

Partial results on 'the interpretability of flexibility' are given. In particular, it is shown that Σ_2-flexibility is indeed interpretable, in the sense that there is a Σ_2 formula $\gamma(x)$ such that for every $\sigma(x) \in \Sigma_2$, every model of T can be end-extended to a model of $\mathrm{T} + \forall x(\gamma(x) \leftrightarrow \sigma(x))$. This result can in turn be generalised to show that for every n, there is a Σ_{n+2} formula as above, such that the extension can be taken to be Σ_n-elementary. The problem of obtaining Σ_n-elementary extensions for Σ_{n+1} formulae seems to be much more difficult.

1.2 About this thesis

This thesis reports on work done within two different projects under two different sets of thesis advisors. Chapter 3 reports on work done under supervision of Christian Bennet, Fredrik Engström, and Dag Westerståhl (main advisor), and concerns properties of sets of provable fixed points in arithmetical theories. Chapters 4 through 7 results from work done under supervision of Christian Bennet, Ali Enayat (main advisor), and Fredrik Engström. These chapters share the common theme of studying independent and flexible formulae of arithmetic, their relationship, and generalisations of those notions.

One ambition in writing this thesis is to give an as complete as possible description of the studied areas. In doing so, given that earlier results in this specific field date between 1930 and 2016, it is often necessary to include a number of theorems that are not original. When this is the case, the origin

of these theorems is clearly stated. Results with no explicit attribution are due to the author.

After this introductory chapter, the second chapter introduces the necessary background and notations that are used in the substantial chapters 3 through 7. Since the technical results in those chapters draws from many different sources, such as the metamathematics of first- and second-order arithmetic, recursion theory and model theory, the background chapter is rather extensive.

Chapter 3 is based on Blanck (2011). The objects of study in this chapter are sets of provable fixed points in arithmetical theories. The main result is that each such set is creative. Hierarchical generalisations are considered, as well as some preliminary results on the algebraic structure of certain collections of sets of fixed points.

Chapter 4 introduces the central notions of independent and flexible formulae, and investigates their relationship. It also acts as a literature review by going through a number of previously published results, but also adding a handful of new generalisations. This chapter is an expanded version of Blanck (2016).

Chapter 5 shifts attention from the syntactic study in Chapter 4, to instead focus on models of arithmetical theories. It is shown that most of the results of Chapter 4 can be formalised, giving rise to particular end-extensions of models of arithmetic. The contents of Chapter 5 is again based on Blanck (2016) but the hierarchical generalisations in Section 5.3 appear here for the first time.

Chapter 6 gives an overview of the famed Orey-Hájek characterisation of interpretability and some of its extensions. For use in some applications in Chapter 7, some other essentially well-known results are included. A new characterisation of partial conservativity over $\mathrm{I}\Sigma_1$ is given. The original results appearing in this chapter have been previously published in Blanck and Enayat (2017).

Chapter 7 focuses on stronger generalisations of theorems from chapters 4 and 5. Partial results on the interpretability aspect of flexible and independent formulae are given. The original results of Section 7.1 and the coding schemes of Section 7.2 have appeared in Blanck and Enayat (2017).

2 Background

The purpose of this chapter is to provide the necessary background material for the rest of this thesis. The results presented in this chapter are all listed as Facts; some of these are rather obvious, while other are substantial, more or less well known, theorems. No proofs of the Facts are given, except in the rare cases where it is difficult to find a proof in the literature. The terminology is chosen to emphasise that these results are the foundation upon which this thesis rests.

The reader is assumed to be acquainted with first-order logic, the first-order theories Q (Robinson's arithmetic) and PA (Peano arithmetic), naive set theory, and the basic theory of recursive functions. More details on the material presented below can be found in the more or less standard textbooks Hájek and Pudlák (1993); Kaye (1991); Lindström (2003); Rogers (1967); Smoryński (1985). Another source, relevant for many of the hierarchical generalisations, is Beklemishev (2005).

2.1 Notation and conventions

The objects of study in this thesis are formal, first-order theories, formulated in (finite extensions of) the language of arithmetic $\mathscr{L}_A$, which contains the non-logical symbols 0, S, $+$, $\times$, $<$. Theories are regarded as sets of sentences: the set of non-logical axioms of the theory. Each theory denoted T, S, ..., possibly with subscripts or other decorations, is assumed to be a consistent extension of Robinson's arithmetic Q. If T is a theory, $\mathrm{Th}(\mathrm{T})$ is the set of theorems of T, i.e., the sentences provable from T.

The terms, formulae and sentences of $\mathscr{L}_A$ are defined as usual. The numerals are written 0, 1, 2, ..., without bars or other devices otherwise used to indicate numerals. Generally, the symbols used for formal variables are x, y, z, u, and v, while the symbols used for numerals are e, i, j, k, m, n. Both kinds of symbols may appear with subscripts or other decorations.

Sentences and formulae of $\mathscr{L}_A$ are denoted by lower case Greek letters, while upper case Greek letters are used for sets of sentences or formulae. The variables displayed are almost always exactly the free variables of a formula, and $\bar{x}$ is sometimes used to denote any finite sequence of free variables.

Fix a Gödel numbering of terms and formulae. $\ulcorner \phi \urcorner$ denotes the numeral representing the Gödel number of ϕ. $\ulcorner \phi(\dot{x}) \urcorner$ denotes the numeral representing the Gödel number of the sentence obtained by replacing x with the value of x. Hence x is free in $\ulcorner \phi(\dot{x}) \urcorner$ but not in $\ulcorner \phi(x) \urcorner$. The symbol $:=$ is used to denote equality between formulae. Let $\top := 0 = 0$, and $\bot := \neg\top$.

Models are structures for a first-order language; this language is always (a finite extension of) the language of arithmetic $\mathscr{L}_A$. A model consists of a non-empty set (called the domain), together with interpretations of the non-logical symbols for functions, relations and constants. Models are denoted $\mathcal{M}$, $\mathcal{N}$, $\mathcal{K}$, $\mathcal{M}'$, $\mathcal{M}_0$, and similarly. The domain of a model $\mathcal{M}$ is denoted by M, while elements of the domain are generally denoted a, b, c. There is one privileged model, the standard model of arithmetic (denoted $\mathbb{N}$), consisting of the set ω of natural numbers, together with the symbols of $\mathscr{L}_A$ under their intuitive interpretations. A sentence is *true*, if it is true in $\mathbb{N}$. Any model that is not isomorphic to the standard model is called non-standard.

If $\phi(x)$ is an $\mathscr{L}_A$-formula and $\mathcal{M}$ an $\mathscr{L}_A$-structure with $a \in M$, the notation $\mathcal{M} \models \phi(a)$ is shorthand for '$\phi(x)$ is true in $\mathcal{M}$ when x is interpreted as a'. It is also possible to treat $\phi(a)$ as shorthand for a formula $\phi(c)$ in an expanded language $\mathscr{L} = \mathscr{L}_A + \{c\}$. If $\mathcal{M}$ is an $\mathscr{L}_A$-structure, and $a \in M$, then $(\mathcal{M}, a)$ denotes the $\mathscr{L}$-structure where c is interpreted as a.

The set of finite binary strings is denoted by ${}^{<\omega}2$, and when s and t are finite binary strings, $s^\frown t$ denotes their concatenation. The set of functions from a set X to $\{0,1\}$ is denoted by ${}^{X}2$, and $\omega^{<\omega}$ denotes the set of non-empty finite subsets of ω.

The notation $\exists x \leq t \phi(x)$ is used as shorthand for $\exists x(x \leq t \wedge \phi(x))$, and similarly $\forall x \leq t \phi(x)$ denotes $\forall x(x \leq t \to \phi(x))$, where t is some $\mathscr{L}_A$-term. The initial quantifiers of these formulae are *bounded*, and a formula containing only bounded quantifiers is a *bounded formula*.

Definition 2.1 (The arithmetical hierarchy)**.**

1. $\Delta_0 = \Sigma_0 = \Pi_0$ is the class of bounded formulae of $\mathscr{L}_A$.
2. A formula is Σ_{n+1} iff it is of the form $\exists x_1 \ldots \exists x_k \pi(x_1, \ldots, x_k, \bar{y})$, where $\pi(x_1, \ldots, x_k, \bar{y})$ is Π_n.
3. A formula is Π_{n+1} iff it is of the form $\forall x_1 \ldots \forall x_k \sigma(x_1, \ldots, x_k, \bar{y})$, where $\sigma(x_1, \ldots, x_k, \bar{y})$ is Σ_n.

The notation introduced above, along with the definition of the arithmetical hierarchy is standard in many textbooks on models of arithmetic such as Kaye (1991), Hájek and Pudlák (1993), and Kossak and Schmerl (2006). It is, however, *not* as standard in the literature on arithmetised metamathematics, e.g. Feferman (1960), Bennet (1986), Lindström (2003). As this thesis lies in the intersection of these two fields, an intermediate class PR of *primitive recursive* formulae is therefore introduced, with the following properties:

Fact 2.2 (Cf. Lindström, 2003, Chapter 1)**.**

1. The class PR contains Δ_0, and is primitive recursive.
2. PR is closed under propositional connectives and bounded quantification.
3. If $\phi(x_1, \ldots, x_n)$ is PR, then $\mathrm{Q} \vdash \phi(k_1, \ldots, k_n)$ iff $\phi(k_1, \ldots, k_n)$ is true.
4. If $\phi(x_1, \ldots, x_n, \bar{y})$ is PR, then $\exists x_1 \ldots \exists x_n \phi(x_1, \ldots, x_n, \bar{y})$ is Σ_1, and $\forall x_1 \ldots \forall x_n \phi(x_1, \ldots, x_n, \bar{y})$ is Π_1.

In what follows, Γ is either Σ_{n+1} or Π_{n+1} and Γ^+ is either Σ_n or Π_n or PR. Γ^d is Σ_n if Γ is Π_n, and vice versa. A Σ_n formula is Δ_n^{T} (or $\Delta_n^{\mathcal{M}}$) if it is equivalent in T (or $\mathcal{M}$) to a Π_n formula, and $\Delta_n = \Delta_n^{\mathbb{N}}$. Note that $\mathrm{PR} \subset \Delta_1$. B_n is the class of Boolean combinations of Σ_n formulae.

For some applications below, it is necessary to consider finite extensions $\mathscr{L}$ of $\mathscr{L}_A$. It is possible to relativise the definition of the arithmetical hierarchy to the extended language, and the resulting classes are denoted

$\Sigma_n(\mathscr{L})$, $\Pi_n(\mathscr{L})$, $\Gamma(\mathscr{L})$, and so on. In those cases that $\mathscr{L} = \mathscr{L}_A \cup \{c\}$, where c is a single constant, the notation $\Sigma_n(c)$ etc. is used for brevity. Whenever $\mathscr{L} = \mathscr{L}_A$, the reference to $\mathscr{L}$ is omitted.

Definition 2.3. For every n, $\mathrm{I}\Sigma_n(\mathscr{L})$ is the theory obtained by augmenting Robinson's Q with an induction axiom for every formula in $\Sigma_n(\mathscr{L})$. Since $\Sigma_0 = \Delta_0$, $\mathrm{I}\Delta_0(\mathscr{L})$ is identical to $\mathrm{I}\Sigma_0(\mathscr{L})$.

Definition 2.4. The theory $\mathrm{I}\Delta_0(\mathscr{L}) + \exp$ is obtained from $\mathrm{I}\Delta_0(\mathscr{L})$ by adding an axiom asserting that the exponentiation function is total.

The induction axioms referred to above are assumed to be formulated *with parameters*. This has the convenient consequence that if $\mathscr{L}$ is an expansion of $\mathscr{L}_A$ obtained by adding finitely many constants, then for all n, $\mathrm{I}\Sigma_n \vdash \mathrm{I}\Sigma_n(\mathscr{L})$.

The strength of the theories $\mathrm{Q}, \mathrm{I}\Delta_0, \mathrm{I}\Delta_0 + \exp, \mathrm{I}\Sigma_1, \mathrm{I}\Sigma_2, \ldots, \mathrm{PA}$ is strictly increasing; each theory in the list proves all the consequences of the previous ones, and no theory proves all the consequences of a later theory.

The theories $\mathrm{I}\Sigma_{n+1}$ and $\mathrm{I}\Delta_0 + \exp$ are the *strong fragments* of PA, while the weak fragments of arithmetic are the ones occurring strictly between $\mathrm{I}\Delta_0 + \exp$ and Q. Q, $\mathrm{I}\Delta_0 + \exp$ and $\mathrm{I}\Sigma_n$ are finitely axiomatisable, but PA is *not*. It is not known if $\mathrm{I}\Delta_0$ is finitely axiomatisable.

In these theories, it is possible to prove some useful closure properties of the classes in the arithmetical hierarchy. The first two items below are presumably folklore, while the last is explicitly stated in Hájek and Pudlák (1993, p. 63–64).

Fact 2.5.

1. Every finite conjunction (or disjunction) of $\Gamma(\mathscr{L})$ formulae is, provably in first-order logic, equivalent to a $\Gamma(\mathscr{L})$ formula.

2. Every $\Gamma(\mathscr{L})$ formula is, provably in $\mathrm{I}\Delta_0(\mathscr{L})$, equivalent to a $\Gamma(\mathscr{L})$ formula with only one quantifier in each block.

3. In $\mathrm{I}\Sigma_{n+1}(\mathscr{L})$ the classes $\Sigma_{n+1}(\mathscr{L})$ and $\Pi_n(\mathscr{L})$ are closed under bounded quantification.

2.2 Arithmetised meta-arithmetic

This section is devoted to introducing the necessary concepts and results from the field here called arithmetised meta-arithmetic. Most, if not all, definitions and facts stated here can be found in Hájek and Pudlák (1993) or Lindström (2003).

The formula $\rho(x_0, \ldots, x_n)$ *numerates* the relation $R(k_0, \ldots, k_n)$ in T if, for every $k_0, \ldots, k_n$,

$$R(k_0, \ldots, k_n) \text{ iff } \mathrm{T} \vdash \rho(k_0, \ldots, k_n).$$

Hence, $\xi(x)$ numerates the set X in T if, for every k,

$$k \in X \text{ iff } \mathrm{T} \vdash \xi(k).$$

Moreover, $\rho(x_0, \ldots, x_n)$ *binumerates* the relation $R(k_0, \ldots, k_n)$ in T if, for every $k_0, \ldots, k_n$,

$$\begin{aligned} &R(k_0, \ldots, k_n) \text{ iff } \mathrm{T} \vdash \rho(k_0, \ldots, k_n), \text{ and} \\ &\text{not } R(k_0, \ldots, k_n) \text{ iff } \mathrm{T} \vdash \neg\rho(k_0, \ldots, k_n). \end{aligned}$$

Hence, $\xi(x)$ binumerates the set X in T if, for every k,

$$\begin{aligned} &k \in X \text{ iff } \mathrm{T} \vdash \xi(k), \text{ and} \\ &k \notin X \text{ iff } \mathrm{T} \vdash \neg\xi(k). \end{aligned}$$

The existence of well-behaved (bi)numerations is guaranteed by the following four results.

Fact 2.6 (Feferman, 1960)**.** A set X is primitive recursive iff there is a PR formula that binumerates X in Q.

Fact 2.7 (Ehrenfeucht and Feferman, 1960)**.** Let T be a consistent, r.e. extension of Q, and let X be any r.e. set. There is then a Σ_1 formula (and also a Π_1 formula) that numerates X in T.

Fact 2.8 (Putnam and Smullyan, 1960)**.** Let T be a consistent, r.e. extension of Q, and let X_0 and X_1 be disjoint r.e. sets. There is then a Σ_1 formula $\xi(x)$ such that $\xi(x)$ numerates X_0 in T and $\neg\xi(x)$ numerates X_1 in T.

In particular, if X is a recursive set, then there is a Σ_1 formula (and therefore also a Π_1 formula) that binumerates X in T.

Definition 2.9. A set X of sentences is monoconsistent with a theory T iff $\mathrm{T} + \phi$ is consistent for all $\phi \in X$.

Fact 2.10 (Lindström, 1979, Lemma 4)**.** Let T be a consistent, r.e. extension of Q, and let X and Y be r.e. sets, Y monoconsistent with T. There is then a Σ_1 formula (and also a Π_1 formula) $\xi(x)$ such that for every k, if $k \in X$, then $\mathrm{T} \vdash \xi(k)$, and if $k \notin X$, then $\xi(k) \notin Y$.

Given a formula $\tau(z)$, let $\mathrm{Prf}_\tau(x, y)$ be a formula expressing the relation 'y is a proof of the sentence x from the set of sentences satisfying $\tau(z)$'. Then $\mathrm{Prf}_\tau(x, y)$ is Γ^+ whenever $\tau(z)$ is. Let $\mathrm{Pr}_\tau(x) := \exists y \mathrm{Prf}_\tau(x, y)$ and $\mathrm{Con}_\tau := \neg \mathrm{Pr}_\tau(\bot)$. Whenever $\tau(z)$ is Σ_{n+1}, $\mathrm{Pr}_\tau(x)$ is Σ_{n+1} and Con_τ is Π_{n+1}. For any formula $\tau(z)$, let $(\tau|y)(z) := \tau(z) \wedge z \leq y$, and $(\tau + y)(z) := \tau(z) \vee z = y$.

If T is an r.e. theory, $\mathrm{Prf}_\mathrm{T}(x, y)$, $\mathrm{Pr}_\mathrm{T}(x)$, $\mathrm{Pr}_{\mathrm{T}+y}(x)$, Con_T, etc. denotes ambiguously $\mathrm{Prf}_\tau(x, y)$, $\mathrm{Pr}_\tau(x)$, $\mathrm{Pr}_{\tau+y}(x)$, Con_τ, etc., where $\tau(z)$ is any PR binumeration of T. A theory T is Γ-definable if there is a $\tau(z) \in \Gamma$ such that $\mathrm{T} = \{k \in \omega : \mathbb{N} \models \tau(k)\}$. If T is Γ-definable but not r.e., $\tau(z)$ is instead assumed to be any Γ formula defining T in $\mathbb{N}$.

The first part of the following useful fact is due to Craig (1953), and the latter part to Grzegorczyk et al. (1958). The generalisation to extended languages is immediate.

Fact 2.11 (Craig's trick)**.**

1. For every $\Sigma_1(\mathscr{L})$-definable theory there is a deductively equivalent PR-definable theory.

2. For every $\Sigma_{n+2}(\mathscr{L})$-definable theory there is a deductively equivalent $\Pi_{n+1}(\mathscr{L})$-definable theory.

Hence, by Fact 2.6, every r.e. (that is, Σ_1-definable) theory has a deductively equivalent axiomatisation that is binumerated by a PR formula in Q.

Many properties of the proof and provability predicates are provable in $\mathrm{I}\Delta_0 + \mathrm{exp}$. The following observations are sometimes useful:

1. $\mathrm{I}\Delta_0 + \exp \vdash \forall x(\tau(x) \to \tau'(x)) \to \forall y(\mathrm{Pr}_\tau(y) \to \mathrm{Pr}_{\tau'}(y))$
2. $\mathrm{I}\Delta_0 + \exp \vdash \forall x(\tau(x) \to \tau'(x)) \to (\mathrm{Con}_{\tau'} \to \mathrm{Con}_\tau)$
3. $\mathrm{I}\Delta_0 + \exp \vdash \forall x \forall y(\mathrm{Pr}_{\tau+x}(y) \leftrightarrow \mathrm{Pr}_\tau(x \to y))$
4. $\mathrm{I}\Delta_0 + \exp \vdash \forall x(\mathrm{Pr}_\tau(x) \wedge \mathrm{Pr}_\tau(\neg x) \to \neg \mathrm{Con}_\tau)$
5. $\mathrm{I}\Delta_0 + \exp \vdash \forall x(\mathrm{Pr}_\tau(\neg x) \leftrightarrow \neg \mathrm{Con}_{\tau+x})$
6. $\mathrm{I}\Delta_0 + \exp \vdash \forall x(\mathrm{Pr}_\tau(x) \leftrightarrow \neg \mathrm{Con}_{\tau+\neg x})$.

A number of constructions of this thesis make use of some kind of self-referential statements. The existence of such statements follows from the following facts. The first is essentially due to Gödel (1931); it is stated in full generality in Carnap (1937).

Fact 2.12 (Diagonal lemma)**.** For every Γ^+ formula $\gamma(x)$, a Γ^+ sentence ξ can be effectively found, such that

$$\mathrm{Q} \vdash \xi \leftrightarrow \gamma(\ulcorner \xi \urcorner).$$

The next two generalisations of the diagonal lemma are due to Ehrenfeucht and Feferman (1960) and Montague (1962), respectively.

Fact 2.13 (Parametric diagonal lemma)**.** For every Γ^+ formula $\gamma(x, y)$, a Γ^+ formula $\xi(x)$ can be effectively found, such that for every $k \in \omega$,

$$\mathrm{Q} \vdash \xi(k) \leftrightarrow \gamma(k, \ulcorner \xi(k) \urcorner).$$

Fact 2.14 (Uniform diagonal lemma)**.** For every Γ^+ formula $\gamma(x, y)$, a Γ^+ formula $\xi(x)$ can be effectively found, such that

$$\mathrm{I}\Delta_0 + \exp \vdash \forall x(\xi(x) \leftrightarrow \gamma(x, \ulcorner \xi(\dot{x}) \urcorner)).$$

The following three results show the incompleteness of sufficiently strong axiomatisable formal theories of arithmetic. The first two are due to Gödel (1931), and the third is due to Rosser (1936). The original results were stated for much stronger theories; see, e.g., Tarski et al. (1953) and Hájek and Pudlák (1993) for the subsequent refinements.

Fact 2.15 (The first incompleteness theorem)**.** Let T be any consistent, r.e. theory extending Q. Then there is a true Π_1 sentence γ such that $\mathrm{T} \nvdash \gamma$.

Fact 2.16 (The second incompleteness theorem)**.** Let T be any consistent, r.e. theory extending $\mathrm{I}\Delta_0 + \mathrm{exp}$. Then $\mathrm{T} + \neg\mathrm{Con}_\mathrm{T}$ is consistent.

Fact 2.17 (Rosser's incompleteness theorem)**.** Let T be any consistent, r.e. theory extending Q. Then there is a Π_1 sentence ρ such that $\mathrm{T} \nvdash \rho$ and $\mathrm{T} \nvdash \neg\rho$.

A related limitative result is Tarski's theorem on the undefinability of truth. A truth-definition for T is a formula $\mathrm{Tr}(x)$ such that for every sentence ϕ, $\mathrm{T} \vdash \phi \leftrightarrow \mathrm{Tr}(\ulcorner\phi\urcorner)$.

Fact 2.18 (Tarski, 1933)**.** Let T be any consistent extension of Q. There is no truth-definition for T.

On the other hand, there are partial truth-definitions, and partial satisfaction predicates, for extensions of $\mathrm{I}\Delta_0 + \mathrm{exp}$; these go back to Hilbert and Bernays (1939).

Fact 2.19. Let $\mathscr{L}$ be a finite extension of $\mathscr{L}_A$. For every k and every $\Gamma(\mathscr{L})$, there is a $k+1$-ary $\Gamma(\mathscr{L})$-formula $\mathrm{Sat}_{\Gamma(\mathscr{L})}(x, x_1, \ldots, x_k)$, such that for every $\Gamma(\mathscr{L})$-formula $\phi(x_1, \ldots, x_k)$ with exactly the variables $x_1, \ldots, x_k$ free, $\mathrm{I}\Delta_0(\mathscr{L}) + \mathrm{exp}$ proves

$$\forall x_1 \ldots \forall x_k(\phi(x_1, \ldots, x_k) \leftrightarrow \mathrm{Sat}_{\Gamma(\mathscr{L})}(\ulcorner\phi\urcorner, x_1, \ldots, x_k)).$$

Such a formula is called a *partial satisfaction predicate* for $\Gamma(\mathscr{L})$.

It follows from the above that for every $\Gamma(\mathscr{L})$, there is a $\Gamma(\mathscr{L})$-formula $\mathrm{Tr}_{\Gamma(\mathscr{L})}(x)$, such that for every $\Gamma(\mathscr{L})$-formula $\phi(x)$,

$$\mathrm{I}\Delta_0(\mathscr{L}) + \mathrm{exp} \vdash \forall x(\phi(x) \leftrightarrow \mathrm{Tr}_{\Gamma(\mathscr{L})}(\ulcorner\phi(\dot{x})\urcorner)),$$

and consequently for every $\Gamma(\mathscr{L})$-sentence ϕ,

$$\mathrm{I}\Delta_0(\mathscr{L}) + \mathrm{exp} \vdash \phi \leftrightarrow \mathrm{Tr}_{\Gamma(\mathscr{L})}(\ulcorner\phi\urcorner).$$

Such a formula is called a *partial truth predicate* for $\Gamma(\mathscr{L})$. These formulae can be used to construct hierarchical provability predicates. Let, for each $\Gamma(\mathscr{L})$,

$$\mathrm{Pr}_{\mathrm{T},\Gamma(\mathscr{L})}(x) := \exists z(z \in \Gamma(\mathscr{L}) \wedge \mathrm{Tr}_{\Gamma(\mathscr{L})}(z) \wedge \mathrm{Pr}_{\mathrm{T}}(\ulcorner \dot{z} \to \dot{x} \urcorner)).$$

Hence $\mathrm{Pr}_{\mathrm{T},\Gamma(\mathscr{L})}(x)$ is the provability predicate corresponding to the theory defined by $\tau(x) \vee \mathrm{Tr}_{\Gamma(\mathscr{L})}(x)$, where $\tau(x)$ is any PR binumeration of T. Let $\mathrm{Con}_{\mathrm{T},\Gamma(\mathscr{L})}$ be the sentence $\neg\mathrm{Pr}_{\mathrm{T},\Gamma(\mathscr{L})}(\bot)$.

The next fact, provable Γ-completeness, has its roots with Hilbert and Bernays (1939). A detailed proof of the Σ_1 case is found in Feferman (1960), see also Beklemishev (2005).

Fact 2.20. Let $\sigma(x_1, \ldots, x_n)$ be any Γ formula. Then

$$\mathrm{I}\Delta_0 + \exp \vdash \forall x_1, \ldots, x_n(\sigma(x_1, \ldots, x_n) \to \mathrm{Pr}_{\mathrm{T},\Gamma}(\ulcorner \sigma(\dot{x}_1, \ldots, \dot{x}_n) \urcorner)).$$

In particular, if $\sigma(x_1, \ldots, x_n)$ is a Σ_1 formula,

$$\mathrm{I}\Delta_0 + \exp \vdash \forall x_1, \ldots, x_n(\sigma(x_1, \ldots, x_n) \to \mathrm{Pr}_{\mathrm{T}}(\ulcorner \sigma(\dot{x}_1, \ldots, \dot{x}_n) \urcorner)),$$

and this can be verified in $\mathrm{I}\Sigma_1$ (Hájek and Pudlák, 1993, Theorem 1.4.32).

The provability predicates are subject to the following very useful conditions; they originate with Hilbert and Bernays (1939), subsequently refined by Löb (1955). For the hierarchical versions presented here, see Smoryński (1985); Beklemishev (2005).

Fact 2.21 (The Hilbert-Bernays-Löb derivability conditions)**.** Let X be any set of Γ sentences such that $\mathrm{T} + X$ is consistent. Then for all sentences ϕ, ψ,

1. if $\mathrm{T} + X \vdash \phi$, then $\mathrm{I}\Delta_0 + \exp + X \vdash \mathrm{Pr}_{\mathrm{T},\Gamma}(\ulcorner \phi \urcorner)$
2. $\mathrm{I}\Delta_0 + \exp \vdash \mathrm{Pr}_{\mathrm{T},\Gamma}(\ulcorner \phi \urcorner) \wedge \mathrm{Pr}_{\mathrm{T},\Gamma}(\ulcorner \phi \to \psi \urcorner) \to \mathrm{Pr}_{\mathrm{T},\Gamma}(\ulcorner \psi \urcorner)$
3. $\mathrm{I}\Delta_0 + \exp \vdash \mathrm{Pr}_{\mathrm{T},\Gamma}(\ulcorner \phi \urcorner) \to \mathrm{Pr}_{\mathrm{T},\Gamma}(\ulcorner \mathrm{Pr}_{\mathrm{T},\Gamma}(\phi) \urcorner)$.

Similar statements also hold for formulae with free variables.

T is Γ-sound if every Γ sentence provable in T is true, and T is sound iff T is Γ-sound for all Γ. A theory is Γ-complete iff all true Γ-sentences are provable in T. The theories Q, $\mathrm{I}\Delta_0+\exp$, PA, $\mathrm{I}\Sigma_1$ etc. are all Σ_1-complete, and are assumed to be sound. An extension of such a theory need not be sound; for example, if T is r.e. and consistent, and even if T is sound, the consistent theory $\mathrm{T} + \neg\mathrm{Con}_\mathrm{T}$ is not Σ_1-sound. The following fact exhibits some essentially well-known properties of these notions, cf., e.g., Lindström (2003), Beklemishev (2005), Kikuchi and Kurahashi (20XX), and Salehi and Seraji (2016). A model-theoretic characterisation of Σ_n-soundness appears in Section 2.3 below.

Fact 2.22.

1. If T is Σ_n-sound, then T is Π_{n+1}-sound.

2. If T is Π_n-complete, then T is Σ_{n+1}-complete.

If X is any set, then $X|k = \{n \in X : n \leq k\}$. A theory T is *reflexive* if $\mathrm{T} \vdash \mathrm{Con}_{\mathrm{T}|k}$ for every k. T is *essentially reflexive* if every extension of T in the same language is reflexive.

Fact 2.23 (Mostowski, 1952)**.** PA is essentially reflexive.

If T is essentially reflexive, then $\mathrm{T} \vdash \phi \rightarrow \mathrm{Con}_\phi$ for all $\phi \in \mathscr{L}_A$. Reflexivity does not imply essential reflexivity: the theory PRA of primitive recursive arithmetic is reflexive but not essentially reflexive.

It follows from the first incompleteness theorem that no finitely axiomatisable theory can be reflexive. There is, however, a notion of small reflection that holds even for finitely axiomatisable theories. This notion is based on that of 'restricted provability from true Γ sentences' and has indispensable use in this thesis.

Let $\mathrm{Pr}^n_\mathrm{T}(x)$ be a formula expressing that there is a proof of x, whose Gödel number is less than n. For most reasonable Gödel numberings, this also restricts the length of all formulae occurring in the proof, as well as the number of quantifier alternations in those formulae, to be less than n. A proof of ϕ from T with these properties is called an n-proof of ϕ. Similarly, $\mathrm{Con}^n_\mathrm{T}$ means that there is no proof of contradiction from T, if only considering proofs whose Gödel numbers are less than n.

Let $\mathrm{Pr}^n_{\mathrm{T},\Gamma(\mathscr{L})}(\ulcorner\phi(\dot{x})\urcorner)$ denote the formula

$$\exists\psi \leq n(\psi \in \Gamma(\mathscr{L}) \wedge \mathrm{Tr}_{\Gamma(\mathscr{L})}(\psi) \wedge \mathrm{Pr}^n_{\mathrm{T}}(\ulcorner\psi \to \phi(\dot{x})\urcorner)),$$

i.e., 'there is an n-proof of $\phi(x)$ from a true $\Gamma(\mathscr{L})$ sentence whose Gödel number is less than n'. If T is r.e. and $\Gamma(\mathscr{L}) = \Sigma_{k+1}(\mathscr{L})$, then $\mathrm{Pr}^n_{\mathrm{T},\Gamma(\mathscr{L})}$ is $\Sigma_{k+1}(\mathscr{L})$. Let $\mathrm{Con}^n_{\mathrm{T},\Gamma(\mathscr{L})}$ denote $\neg\mathrm{Pr}^n_{\mathrm{T},\Gamma(\mathscr{L})}(\bot)$. The desired reflection principle for this provability notion follows from the next fact, which is due to Feferman (1962, Lemma 2.18).[3]

Fact 2.24. Let T be an r.e. theory formulated in a finite language $\mathscr{L} \supseteq \mathscr{L}_A$ and let $\phi(x) \in \mathscr{L}$. Then for all $n \in \omega$,

$$\mathrm{I}\Delta_0(\mathscr{L}) + \exp \vdash \forall x(\mathrm{Pr}^n_{\mathrm{T}}(\ulcorner\phi(\dot{x})\urcorner) \to \phi(x)).$$

Fact 2.25 (Small reflection)**.** Let T be an r.e. theory formulated in a finite language $\mathscr{L} \supseteq \mathscr{L}_A$, and let $\phi(x) \in \mathscr{L}$. For each $n \in \omega$,

$$\mathrm{I}\Delta_0(\mathscr{L}) + \exp \vdash \forall x(\mathrm{Pr}^n_{\mathrm{T},\Gamma(\mathscr{L})}(\ulcorner\phi(\dot{x})\urcorner) \to \phi(x)).$$

Proof. Pick $\phi(x) \in \mathscr{L}$, fix $n \in \omega$ and reason in $\mathrm{I}\Delta_0(\mathscr{L}) + \exp$:

> Pick x. If $\neg\mathrm{Pr}^n_{\mathrm{T},\Gamma(\mathscr{L})}(\ulcorner\phi(\dot{x})\urcorner)$, then the implication is vacuously true. Hence suppose that $\mathrm{Pr}^n_{\mathrm{T},\Gamma(\mathscr{L})}(\ulcorner\phi(\dot{x})\urcorner)$. Then
>
> $$\exists\psi \leq n(\psi \in \Gamma(\mathscr{L}) \wedge \mathrm{Tr}_{\Gamma(\mathscr{L})}(\psi) \wedge \mathrm{Pr}^n_{\mathrm{T}}(\ulcorner\psi \to \phi(\dot{x})\urcorner).$$

Now recall Fact 2.24, and continue reasoning in $\mathrm{I}\Delta_0(\mathscr{L}) + \exp$:

> It follows that $\psi \to \phi(x)$, and ψ follows from $\mathrm{Tr}_{\Gamma(\mathscr{L})}(\psi)$. Hence $\phi(x)$, so $\forall x(\mathrm{Pr}^n_{\mathrm{T},\Gamma(\mathscr{L})}(\ulcorner\phi(\dot{x})\urcorner) \to \phi(x))$. □

This principle can also be formalised (Verbrugge and Visser, 1994), which yields $\mathrm{I}\Delta_0(\mathscr{L}) + \exp \vdash \forall z\mathrm{Pr}_{\mathrm{T}}(\ulcorner\forall x(\mathrm{Pr}^z_{\mathrm{T},\Gamma(\mathscr{L})}(\phi(x)) \to \phi(x))\urcorner)$.

An important concept is that of *partial conservativity*, which in its general form appears in Guaspari (1979). Earlier examples of partial conservative sentences can be found in, e.g., Kreisel (1962).

[3] Feferman's result is stated for extensions of PA, but is well known to hold for extensions of $\mathrm{I}\Delta_0 + \exp$ (Hájek and Pudlák, 1993, Lemma III.4.40; Beklemishev, 2005, Lemma 2.2).

Definition 2.26. A sentence θ is $\Gamma(\mathscr{L})$-conservative over a theory T iff for all $\gamma \in \Gamma(\mathscr{L})$, whenever $T + \theta \vdash \gamma$, then $T \vdash \gamma$. In other words, $T + \theta$ and T have the same $\Gamma(\mathscr{L})$-consequences.

The notion of partial conservativity is occasionally used in an extended sense, saying that (for theories $S \vdash T$) S is $\Gamma(\mathscr{L})$-conservative over T iff for all $\gamma \in \Gamma(\mathscr{L})$, whenever $S \vdash \gamma$, then $T \vdash \gamma$.

A related notion is that of an interpretation of one theory in another. Roughly speaking, S is interpretable in T if the primitive concepts and variables of S are definable in T in a way that turns every theorem of S into a theorem of T. It is easy to see that if $T \vdash S$, then S is interpretable in T. Further properties of interpretations are discussed in Chapter 6.

Fact 2.27 (Feferman, 1960)**.** Let T be any consistent, r.e. extension of $I\Delta_0 + \exp$. Then $T + \neg \mathrm{Con}_T$ is interpretable in T.

2.3 Model theory of arithmetic

A model of arithmetic is a first-order structure that is adequate for the language $\mathscr{L}_A$. A detailed introduction to models of arithmetic, containing most of the material covered here, is Kaye (1991). The first three facts are standard tools from the general theory of first-order models. The first is due to Gödel (1930), as is the countable case of the second; the uncountable case is due to Maltsev (1936).

Fact 2.28 (The completeness theorem)**.** Let X be a set of sentences. Then X has a model iff X is consistent.

Fact 2.29 (The compactness theorem)**.** Let X be a set of sentences. Then X has a model iff every finite subset of X has a model.

Fact 2.30 (The Löwenheim-Skolem theorem (1915), (1920))**.** Let T be a theory formulated in a countable language $\mathscr{L}$. If T has an infinite model, then T has a countable model.

Non-standard models of arithmetic contain infinitely many 'infinitely large' non-standard elements. A useful feature of these elements is that

some properties holding of arbitrarily large standard numbers spill over to the non-standard elements. The notion of overspill is originally due to Robinson (1963), while the hierarchical version stated here is from Hájek and Pudlák (1993).

Fact 2.31 (Overspill)**.** Let $\mathcal{M} \models \mathrm{I}\Sigma_n(\mathscr{L})$. Suppose that $a \in M$, and that $\phi(x, y)$ is a $\Sigma_n(\mathscr{L})$ (or $\Pi_n(\mathscr{L})$) formula such that

$$\mathcal{M} \models \phi(n, a) \text{ for all } n \in \omega.$$

Then there is a $b \in M \setminus \omega$ such that $\mathcal{M} \models \forall x \leq b\phi(x, a)$.

If $\mathcal{M}$ is a submodel of $\mathcal{N}$, and for all $a \in M$ and all $\gamma(x) \in \Gamma$,

$$\mathcal{M} \models \gamma(a) \Leftrightarrow \mathcal{N} \models \gamma(a),$$

then $\mathcal{M}$ is a Γ-elementary submodel of $\mathcal{N}$, in symbols $\mathcal{M} \prec_\Gamma \mathcal{N}$. Equivalently, $\mathcal{N}$ is said to be a Γ-elementary extension of $\mathcal{M}$.

Let $\mathcal{M}$ and $\mathcal{N}$ be models of arithmetic, and suppose that $\mathcal{M}$ is a submodel of $\mathcal{N}$. Then $\mathcal{M}$ is an *initial segment* of $\mathcal{N}$, or equivalently, $\mathcal{N}$ is an *end-extension* of $\mathcal{M}$, in symbols $\mathcal{M} \subseteq_e \mathcal{N}$, iff

$$\text{for each } a \in N \setminus M, \mathcal{N} \models b < a \text{ for all } b \in M.$$

If $\mathcal{M} \subseteq_e \mathcal{N}$, then $\mathcal{N}$ is a Δ_0-elementary extension of $\mathcal{M}$: hence Δ_0 formulae are absolute between end-extensions. An important consequence of this is that Σ_1-sentences are preserved when passing to an end-extension: if σ is a Σ_1-sentence, $\mathcal{M} \models \sigma$ and $\mathcal{M} \subseteq_e \mathcal{N}$, then $\mathcal{N} \models \sigma$. Conversely, Π_1-sentences are preserved when passing to an initial submodel.

Recall that $\mathrm{Th}(\mathrm{T})$ denotes the set of theorems of T. Analogously, the notation $\mathrm{Th}(\mathcal{M})$, where $\mathcal{M}$ is an $\mathscr{L}$-structure, is used for the set of sentences that hold in $\mathcal{M}$, i.e. $\{\phi \in \mathscr{L} : \mathcal{M} \models \phi\}$. Furthermore, for each Γ, the set $\mathrm{Th}_\Gamma(\mathcal{M})$ is defined as $\{\phi \in \Gamma : \mathcal{M} \models \phi\}$. The model-theoretic characterisation of soundness can now be given:

Fact 2.32. T is Σ_n-sound iff $\mathrm{T} + \mathrm{Th}_{\Sigma_{n+1}}(\mathbb{N})$ is consistent.

Let $\mathcal{M}$ be any model of arithmetic, and let R be a relation in M^n. Then R is $\mathcal{M}$-definable if there is a formula $\phi(x_1, \ldots, x_n)$ such that $R = \{\langle a_1, \ldots, a_1\rangle : \mathcal{M} \models \phi(a_1, \ldots, a_n)\}$. If this ϕ can be chosen as a Γ formula, then R is Γ-definable in $\mathcal{M}$.

Due to the existence of partial truth definitions $\mathrm{Tr}_\Gamma(x)$ of complexity Γ, the set

$$\mathrm{True}_\Gamma(\mathcal{M}) = \{m \in M : \mathcal{M} \models \mathrm{Tr}_\Gamma(m)\}$$

of 'true Γ-sentences' as calculated within $\mathcal{M}$ is Γ-definable in $\mathcal{M}$. Hence $\mathrm{True}_\Gamma(\mathcal{M}) \cap \omega = \mathrm{Th}_\Gamma(\mathcal{M})$. Using the hierarchical consistency statement $\mathrm{Con}_{\mathrm{T},\Sigma_n}$ introduced in the previous section, the assertion that '$\mathcal{M}$ thinks that $\mathrm{T} + \mathrm{True}_{\Sigma_n}(\mathcal{M})$ is consistent', which would otherwise require triple subscripts, can now be conveniently expressed as $\mathcal{M} \models \mathrm{Con}_{\mathrm{T},\Sigma_n}$.

Let $n\varepsilon a$ be Ackermann's epsilon notation, meaning that the n'th position of the binary expansion of a is 1. Then a can be understood as a code for the set consisting of all the n's such that $n\varepsilon a$. Let $\mathcal{M}$ be a non-standard model of PA. Then $\mathrm{SSy}(\mathcal{M})$, the standard system of $\mathcal{M}$, is the collection of sets $X \subseteq \omega$ such that for some $a \in M$, $X = \{n \in \omega : \mathcal{M} \models n\varepsilon a\}$. Then a is said to be a code for X, and X is coded in $\mathcal{M}$. Moreover, every coded set has arbitrarily small non-standard codes.

Fact 2.33. Let $\mathcal{M}$ be a non-standard model of $\mathrm{I}\Sigma_n$, and let $\phi(x)$ be any Σ_n formula. Then $\{k \in \omega : \mathcal{M} \models \phi(k)\}$ is coded in $\mathcal{M}$.

Let T be a complete, consistent theory. $\mathrm{Rep}(\mathrm{T})$ is the collection of sets $X \subseteq \omega$ representable in T, i.e. the sets for which there exists a $\xi(x)$ that binumerates X in T.

Fact 2.34 (Wilkie, 1977)**.** If $\mathcal{M} \models \mathrm{PA}$, and T is a complete, consistent extension of PA, then there is an $\mathcal{N} \models \mathrm{T}$ end-extending $\mathcal{M}$ iff

1. $\mathrm{Rep}(\mathrm{T}) \subseteq \mathrm{SSy}(\mathcal{M})$;
2. $\mathrm{T} \cap \Pi_1 \subseteq \mathrm{Th}(\mathcal{M})$.

The next result is a refinement of Friedman's embedding theorem, due to Ressayre (1987), and Dimitracopoulos and Paris (1988), independently.

Fact 2.35. If $\mathcal{M}$ and $\mathcal{N}$ are two countable models of $\mathrm{I}\Sigma_1$ with $a \in M$ and $b \in N$, then the following are equivalent:

1. $\mathcal{M}$ is embeddable as an initial segment of $\mathcal{N}$ via an embedding f with $f(a) = b$;
2. $\mathrm{SSy}(\mathcal{M}) = \mathrm{SSy}(\mathcal{N})$, and $\mathrm{Th}_{\Sigma_1}(\mathcal{M}, a) \subseteq \mathrm{Th}_{\Sigma_1}(\mathcal{N}, b)$.

The arithmetised completeness theorem (ACT) states that the canonical proof of the completeness theorem can be carried out within a suitable arithmetic theory, such as PA, and this theorem a useful tool for producing end-extensions of models of arithmetic. The version that is sufficient for all applications in this thesis (Fact 2.38) is an easy corollary to the next two facts. First, it is necessary to introduce some new notation: the closure of a set X under propositional connectives and bounded quantification is denoted $\Sigma_0(X)$. By Lemma I.2.14 of Hájek and Pudlák (1993), $\mathrm{I}\Sigma_n$ proves induction for all $\Sigma_0(\Sigma_n)$ formulae.

The following 'mild refinement' of the arithmetised completeness theorem is essentially due to Paris (1981); the version stated here is from Cornaros and Dimitracopoulos (2000).

Fact 2.36 (The arithmetised completeness theorem)**.** Let $\mathcal{M} \models \mathrm{I}\Sigma_{n+1}$. Let $\mathscr{L}$ be a language in $\mathcal{M}$, extending $\mathscr{L}_A$, which is $\Delta_1^{\mathcal{M}}$, and let S denote the set of sentences of $\mathscr{L}$ in the sense of $\mathcal{M}$. Let $A_1 \subseteq S$ be Σ_n in $\mathcal{M}$ and let $A_2 \subseteq S$ be Π_n in $\mathcal{M}$ such that $\mathcal{M} \models \mathrm{Con}_{A_1 \cup A_2}$. Then there is a set B with $A_1 \cup A_2 \subseteq B \subseteq S$ such that:

1. for every $\phi \in S$, $\phi \in B$ or $\neg\phi \in B$;
2. B is $\Sigma_0(\Sigma_{n+1})$ in $\mathcal{M}$;
3. $\mathcal{M} \models \mathrm{Con}_B$.

For the intended applications where $n = 0$ it is not always possible to guarantee that A_1 and A_2 are Σ_0 and Π_0, respectively, only that their union is PR. In these cases the following version of the ACT comes in handy. It appears in Hájek and Pudlák (1993, Theorem I.4.27), the present wording is taken from Wong (2016).

Fact 2.37. $\mathrm{I}\Sigma_1$ proves: If T is a Δ_1-definable consistent theory, then T has a definable model all of whose Σ_1 properties are $\Sigma_0(\Sigma_1)$-definable.

Together, these results have the following useful consequence, which is exactly what is used in the applications in thesis.

Fact 2.38. Let $\mathscr{L} = \mathscr{L}_A \cup \{c\}$. Suppose that $\mathcal{M} \models \mathrm{I}\Sigma_{n+1}$, and that T is a theory formulated in $\mathscr{L}$, such that T is Σ_{n+1}-definable in $\mathcal{M}$, and that $\mathcal{M} \models \mathrm{Con}_{\mathrm{T}}$. Then there is an $\mathscr{L}$-structure $(\mathcal{N}, c)$ such that:

1. $\mathcal{N}$ end-extends $\mathcal{M}$;
2. $(\mathcal{N}, c)$ satisfies the standard sentences of T.

Proof. Suppose first that $n > 0$, and suppose that T is Σ_{n+1}-definable, and that $\mathcal{M} \models \mathrm{Con}_{\mathrm{T}}$. Then by Craig's trick (Fact 2.11), T has a deductively equivalent Π_n definition A_2, and $\mathcal{M} \models \mathrm{Con}_{A_2}$. By Fact 2.36, there is a complete, consistent extension B of A_2 which is $\Sigma_0(\Sigma_{n+1})$-definable in $\mathcal{M}$, so B is the elementary diagram of some model $\mathcal{N}$ of A_2 (and therefore, of T). Since $\mathcal{M} \models \mathrm{I}\Sigma_0(\Sigma_{n+1})$, it is possible to define an embedding from $\mathcal{M}$ onto an initial submodel of $\mathcal{N}$.

Now, suppose that $n = 0$, and that T is Σ_1-definable. Then T has a deductively equivalent PR definition, whence this definition is Δ_1. By Fact 2.37, there is a definable model $\mathcal{N}$ of T, all of whose Σ_1 properties are $\Sigma_0(\Sigma_1)$-definable. Since $\mathcal{M} \models \mathrm{I}\Sigma_0(\Sigma_1)$, it is again possible to embed $\mathcal{M}$ onto an initial submodel of $\mathcal{N}$. □

The next result concerns the existence of non-standard initial segments satisfying stronger theories. It is due to McAloon (1982); cf. D'Aquino (1993).

Fact 2.39. If $\mathcal{M}$ is a countable non-standard model of $\mathrm{I}\Delta_0$, and T is a Σ_1-sound $\mathscr{L}_A$-theory, then there is some non-standard initial segment of $\mathcal{M}$ that is a model of T. In particular, for every non-standard $a \in M$, $\mathcal{M}$ has a non-standard initial segment below a that is a model of PA.

Although this thesis deals mainly with first-order arithmetic, models of second-order arithmetic appear occasionally; hence the need to introduce

some terminology. More details, and proofs of the results below, can be found in Simpson (1999). The language of second-order arithmetic, $\mathscr{L}_2$, is two-sorted; it has a number sort, and a set sort. There is also a symbol for set membership, $\in$. WKL_0 is the subsystem of second-order arithmetic consisting of Q plus induction for Σ_1-formulae in $\mathscr{L}_2$, plus weak König's lemma, i.e. the statement that every infinite subtree of the full binary tree has an infinite path.

A model of second-order arithmetic is a tuple $(\mathcal{M}, \mathcal{A})$, where $\mathcal{M}$ is a first-order structure and $\mathcal{A}$ is a collection of subsets of M. If $(\mathcal{M}, \mathcal{A})$ is a model of second-order arithmetic, $\mathcal{M}$ is referred to as its first-order part.

The following result is due to Harrington for countable $\mathcal{M}$ (unpublished, see Simpson, 1999, Lemma IX.1.8 and Theorem IX.2.1); the uncountable case is due to Hájek (1993).

Fact 2.40. Every model $\mathcal{M}$ of $\mathrm{I}\Sigma_1$ can be expanded to a model $(\mathcal{M}, \mathcal{A})$ of WKL_0.

Fact 2.41 (Simpson, 1999, Theorem IV.3.3)**.** WKL_0 proves the compactness theorem and the completeness theorem.

The final result of this section follows immediately from Fact 2.40 and the completeness theorem.

Fact 2.42. WKL_0 proves the same first-order sentences as $\mathrm{I}\Sigma_1$.

2.4 Recursion theory

The reader is assumed to be familiar with the concepts of (partial) recursive functions and Turing machines. The material presented here can be found in any of the two detailed introductions to recursion theory Kleene (1952) and Rogers (1967). For a careful development of formalised recursion theory of the kind introduced below, the reader is directed to Smoryński (1985, Chapter 0).[4]

[4]Some of the results below could equally well be thought to belong to Section 2.2, in that they concern the formalisation of some parts of recursion theory within arithmetic. In that case, Section 2.2 would have to appear after the present section, with a similar comment added to the introduction of that section.

Definition 2.43. A set $Y \subset \omega$ is recursive in X (or X-recursive) if Y is solvable under the assumption that X is solvable; in symbols $Y \leq_T X$. The notation $X \equiv_T Y$ is shorthand for $Y \leq_T X$ and $X \leq_T Y$.

Fact 2.44 (The enumeration theorem, Kleene, 1952, Theorem XXII)**.** Let X be any set. For each k, there is a function that is universal for k-ary X-recursive functions: that is, a function $\Phi_k^X(x, y_1, \ldots, y_k)$ that is partial recursive in X, and is such that $\Phi_k^X(n, y_1, \ldots, y_k)$ for $n \in \omega$ is an enumeration, with repetitions, of all the partial X-recursive functions of k variables.

This enumeration is *acceptable* in the sense of Rogers (1967): there is an effective correspondence between the partial X-recursive functions and Turing machines with an oracle for X. For each partial X-recursive function φ_e^X, let the e'th X-r.e. set, W_e^X, be the domain of φ_e^X. If f is a function such that $f \simeq \varphi_e^X$ for some e, then e is an X-index for f. The reference to X is omitted whenever X is a recursive set.

Say that a relation (set, function) is Σ_n iff it is definable in $\mathbb{N}$ by a Σ_n formula, and similarly for Π_n. A relation is $\Delta_n(\mathbb{N})$ iff it is definable in $\mathbb{N}$ by both a Σ_n and a Π_n formula. Formulae in these classes are subject to a strong normal form theorem, as shown by Kleene (1952, Theorem IV).

Fact 2.45 (The normal form theorem)**.** Let $n > 0$. Every k-ary Σ_n relation can be defined in $\mathbb{N}$ by a formula of the form

$$\exists y_1 \forall y_2 \ldots \mathsf{Q} y_n T(e, x_1, \ldots, x_k, y_1, \ldots, y_n)$$

for a suitable choice of e. In the above formula, T is Kleene's primitive recursive T-predicate, Q is $\exists$ or $\forall$ depending on whether n is odd or even, and in this latter case, T is prefixed with a negation symbol.

Similarly, every k-ary Π_n relation can be defined in $\mathbb{N}$ by a formula of the form

$$\forall y_1 \exists y_2 \ldots \mathsf{Q} y_n T(e, x_1, \ldots, x_k, y_1, \ldots, y_n)$$

for a suitable choice of e. In the above formula, T is Kleene's primitive recursive T-predicate, Q is $\forall$ or $\exists$ depending on whether n is odd or even, and in the former case, T is prefixed with a negation symbol.

The next few definitions single out particularly interesting classes of sets: the productive, creative, complete Σ_n, and recursively inseparable sets.

Definition 2.46. A set X is productive if there is a total recursive function $f(x)$ such that for all $n \in \omega$, if $W_n \subseteq X$, then $f(n) \in X \setminus W_n$. A set X is creative if it is r.e. and its complement is productive.

A set $Y \subset \omega$ is 1-reducible to X ($Y \leq_1 X$) if there is a recursive 1:1 function f such that $k \in Y$ iff $f(k) \in X$.

Definition 2.47. A set X is complete Σ_n if X is Σ_n, and $Y \leq_1 X$ for each Σ_n set Y.

Definition 2.48. Two sets X, Y are effectively inseparable if, for every disjoint, r.e. sets $X' \supseteq X$, $Y' \supseteq Y$, it is possible to effectively find an element of $(X' \cup Y')^c$.

Fact 2.49 (Myhill, 1955)**.** X is creative iff X is complete Σ_1.

If X is any r.e. set other than ω, then X is creative iff for every r.e. set Y that is disjoint from X, $X \equiv_T X \cup Y$. If X and Y are disjoint, r.e., effectively inseparable sets, then both X and Y are creative.

Let X be any set, and let $\{\varphi_i^X : i \in \omega\}$ be an enumeration of the X-recursive functions. Let the Turing jump of X, denoted X', be the set $\{x : \varphi_x^X(x) \text{ is defined}\}$. The nth Turing jump of X is inductively defined by

$$X^{(0)} = X$$
$$X^{(n+1)} = (X^{(n)})'$$

Of particular interest are the jumps of the empty set $\emptyset$, as they are closely connected to the arithmetical hierarchy. This relationship is clarified in the following fact, which is due to Post (1948).

Fact 2.50 (Post's theorem)**.**

1. A set X is Σ_{n+1} iff X is r.e. in $\emptyset^{(n)}$.
2. The set $\emptyset^{(n)}$ is complete Σ_n.

Whenever $X = \emptyset^{(n)}$ for some $n \in \omega$, the notation $\Phi^n_k(x, y_1, \ldots, y_k)$ is used in place of $\Phi^X_k(x, y_1, \ldots, y_k)$. The superscript is completely dropped when $n = 0$. By Kleene's normal form theorem and Post's theorem, the relation $\Phi^n_k(x, y_1, \ldots, y_k) = z$ can be defined in $\mathbb{N}$ by a Σ_{n+1} formula $R^n_k(x, y_1, \ldots, y_k, z)$. Since the arity of these formulae is always obvious from the context, the subscript k is henceforth dropped.

A k-ary function f is strongly defined by a formula $\phi(x_1, \ldots, x_k, y)$ in T iff

1. if $f(n_1, \ldots, n_k) = m$, then $\mathrm{T} \vdash \phi(n_1, \ldots, n_k, m)$ and $\mathrm{T} \vdash \forall y(\phi(n_1, \ldots, n_k, y) \to y = m)$;

2. if $f(n_1, \ldots, n_k) \neq m$, then $\mathrm{T} \vdash \neg\phi(n_1, \ldots, n_k, m)$.

The function $\Phi(x, y_1, \ldots, y_k)$ is recursive, and since Q is Σ_1-complete, the relation $\Phi(x, y_1, \ldots, y_k) = z$ can be strongly represented in Q by a Σ_1 formula $R(x, y_1, \ldots, y_k, z)$. As pointed out by Ali Enayat, this is a special case of a more general phenomenon.

Fact 2.51. For each n, a function f is recursive in $\emptyset^{(n)}$ (or, equivalently, f is Δ_{n+1}) iff f is strongly representable in $\mathrm{Q} + \mathrm{Th}_{\Pi_n}(\mathbb{N})$.

Proof sketch. Let f be a function recursive in $\emptyset^{(n)}$. By Post's theorem, the graph and the co-graph of f can both be defined in $\mathbb{N}$ by a Σ_{n+1}-formula. Since the theory $\mathrm{Q} + \mathrm{Th}_{\Pi_n}(\mathbb{N})$ is Σ_{n+1}-complete and Π_{n+1}-sound, f is strongly representable in $\mathrm{Q} + \mathrm{Th}_{\Pi_n}(\mathbb{N})$. For the other direction, note that $\mathrm{Th}_{\Pi_n}(\mathbb{N})$ is recursive in $\emptyset^{(n)}$. □

The following result is taken from Smoryński (1985, Theorem 0.6.9) for the case $n = 0$; the hierarchical generalisation given here is supposedly folklore. A slogan for this fact is: there is a Σ_{n+1} function hidden within every Σ_{n+1} relation, and this can be verified in $\mathrm{I}\Sigma_n$.

Fact 2.52 (The selection theorem)**.** For each Σ_{n+1}-formula ϕ with exactly the variables $x_1, \ldots, x_k$ free, there is a Σ_{n+1}-formula $\mathrm{Sel}\{\phi\}$ with exactly the same free variables, such that:

1. $\mathrm{I}\Sigma_n \vdash \forall x_1, \ldots, x_k(\mathrm{Sel}\{\phi\}(x_1, \ldots, x_k) \to \phi(x_1, \ldots, x_k))$;

2. $\mathrm{I}\Sigma_n \vdash \forall x_1, \ldots, x_k, y(\mathrm{Sel}\{\phi\}(x_1, \ldots, x_k) \wedge \mathrm{Sel}\{\phi\}(x_1, \ldots, x_{k-1}, y) \rightarrow x_k = y)$;

3. $\mathrm{I}\Sigma_n \vdash \forall x_1, \ldots, x_{k-1}(\exists x_k \phi(x_1, \ldots, x_k) \rightarrow \exists x_k \mathrm{Sel}\{\phi\}(x_1, \ldots, x_k))$.

These formulae are useful in that they can be used in combination with partial satisfaction predicates to strongly represent $\emptyset^{(n)}$-recursive functions in extensions of $\mathrm{I}\Sigma_n + \mathrm{Th}_{\mathrm{B}_n}(\mathbb{N})$, by letting φ_e be the $\emptyset^{(n)}$-recursive function whose graph is defined by $\mathrm{Sel}\{\mathrm{Sat}_{\Sigma_{n+1}}\}(e, y_1, \ldots, y_k, z)$ in $\mathbb{N}$. The resulting enumeration is acceptable in Rogers's sense, so whenever convenient, it can without loss of generality be assumed that

$$R^n(x, y_1, \ldots, y_k, z) := \mathrm{Sel}\{\mathrm{Sat}_{\Sigma_{n+1}}\}(x, y_1, \ldots, y_k, z).$$

Fact 2.53 (The recursion theorem)**.** Let $f(z, x_1, \ldots, x_n)$ be any partial X-recursive function. There is an X-index e such that

$$\varphi_e^X(x_1, \ldots, x_n) \simeq f(e, x_1, \ldots, x_n).$$

This theorem is due to Kleene (1952, Theorem XXVII), and is usually employed in the following manner. Define a recursive function $f(z, x)$ in stages, using z as a parameter. The resulting function may differ depending on the choice of z. By the recursion theorem, there is an index e such that $\varphi_e \simeq f(e, x)$. Hence the function $\varphi_e(x)$ computes the same function as $f(z, x)$ does when fed its own index as the first parameter. This legitimates self-referential constructions where an index of f is being used in the actual construction of f. The recursion theorem can be formalised in $\mathrm{I}\Delta_0 + \exp$ using the diagonal lemma (Smoryński, 1985, Theorem 0.6.12; Lindström and Shavrukov, 2008, Section 1.2). To show that, e.g., a Σ_{n+1} function acquired in this way is provably total, $\mathrm{I}\Sigma_{n+1}$ is required.

3 Sets of fixed points

The diagonal lemma (and its variations) is frequently used to construct 'self-referential' sentences in the form of *provable fixed points*: ϕ is a provable fixed point of $\theta(x)$ in T iff $\mathrm{T} \vdash \phi \leftrightarrow \theta(\ulcorner\phi\urcorner)$. Some classic results established by this means are Gödel's first incompleteness theorem, Tarski's theorem on the undefinability of truth, and Löb's theorem. Lindström (2003) gives many examples of how versatile the technique can be. In this chapter, provable fixed points, or merely *fixed points*, are studied from another perspective: given a formula $\theta(x)$ in $\mathscr{L}_A$, what can be said about the set of fixed points of $\theta(x)$?

It is a well known fact that the set of all provable fixed points of $\mathrm{Pr}_{\mathrm{PA}}$,

$$\{\phi : \mathrm{PA} \vdash \phi \leftrightarrow \mathrm{Pr}_{\mathrm{PA}}(\ulcorner\phi\urcorner)\}$$

is creative. This is an easy corollary of Löb's theorem (1955), together with Smullyan's theorem (1961) showing that the set of theorems of PA and the set of refutable sentences of PA are effectively inseparable.

Say that a formula $\theta(x)$ is *extensional* (or preserves the provable equivalence) if, for each ϕ and ψ, $\mathrm{T} \vdash \phi \leftrightarrow \psi$ implies $\mathrm{T} \vdash \theta(\ulcorner\phi\urcorner) \leftrightarrow \theta(\ulcorner\psi\urcorner)$. An important subclass of the extensional formulae is the formulae that are T-substitutable in the sense that for all ϕ and ψ,

$$\mathrm{T} \vdash \mathrm{Pr}_{\mathrm{T}}(\ulcorner\phi \leftrightarrow \psi\urcorner) \rightarrow \theta(\ulcorner\phi\urcorner) \leftrightarrow \theta(\ulcorner\psi\urcorner).$$

Smoryński (1987) shows that the class of T-substitutable formulae have, up to provable equivalence in T, a unique provable fixed point, and Bernardi (1981) generalises Smullyan's result to show that every two different PA-equivalence classes are effectively inseparable. It is then easy to see that the set of provable fixed points of a substitutable formula is creative.

Bennet shows, in unpublished notes, that the set of Rosser sentences is complete Σ_1, which seems to be the first result of this kind for non-extensional formulae. Halbach and Visser (2014) show, using a method

due to McGee, that each set of fixed points is complete r.e. Their result is mildly strengthened in this chapter to show that each set of fixed points is creative.

3.1 Recursion theoretic complexity

Let, for each $\mathscr{L}_A$-formula $\theta(x)$, $\mathrm{Fix}^{\mathrm{T}}(\theta) = \{\phi : \mathrm{T} \vdash \phi \leftrightarrow \theta(\ulcorner\phi\urcorner)\}$. It is evident that if T is r.e., then each $\mathrm{Fix}^{\mathrm{T}}(\theta)$ is r.e. The equivalence class of ψ over T is the set $[\psi]^{\mathrm{T}} = \{\phi : \mathrm{T} \vdash \phi \leftrightarrow \psi\}$. Where no confusion will arise, the reference to T is omitted. Note that, for each sentence ψ, $[\psi] = \mathrm{Fix}(\psi \wedge x = x)$, which means that the next result is a more general form of Theorem 1 of Bernardi (1981).

Theorem 3.1. Every two disjoint sets of fixed points are effectively inseparable.

Proof. Let $\theta(x)$ and $\chi(x)$ be any formulae, and let X, Y be disjoint r.e. sets containing $\mathrm{Fix}(\theta)$ and $\mathrm{Fix}(\chi)$, respectively. Let, by Fact 2.8, $\xi(x)$ be a Σ_1 formula such that $\xi(x)$ numerates X and $\neg\xi(x)$ numerates Y in T. Let, by the diagonal lemma (Fact 2.12), ϕ be such that

$$\mathrm{T} \vdash \phi \leftrightarrow (\theta(\ulcorner\phi\urcorner) \wedge \neg\xi(\ulcorner\phi\urcorner)) \vee (\chi(\ulcorner\phi\urcorner) \wedge \xi(\ulcorner\phi\urcorner)).$$

Suppose $\phi \in X$. Then $\mathrm{T} \vdash \xi(\ulcorner\phi\urcorner)$, so $\mathrm{T} \vdash \phi \leftrightarrow \chi(\ulcorner\phi\urcorner)$, contradicting the assumption that X and Y are disjoint. Suppose instead that $\phi \in Y$. Then $\mathrm{T} \vdash \neg\xi(\ulcorner\phi\urcorner)$, and $\mathrm{T} \vdash \phi \leftrightarrow \theta(\ulcorner\phi\urcorner)$, again contradicting the assumption that X is disjoint from Y. Hence $\phi \notin X \cup Y$. □

By Myhill's theorem, the concepts of creativeness and Σ_1-completeness coincide. Furthermore, every two disjoint, effectively inseparable sets are both creative. This allows for the following conclusion, using two different proofs, of which the latter constructs the reducing function directly.

Corollary 3.2. Every set of fixed points is creative.

Proof. Let $\theta(x)$ be any formula. It suffices to show that $\mathrm{Fix}(\theta)$ is disjoint from some other set of fixed points. But $\mathrm{Fix}(\theta) \cap \mathrm{Fix}(\neg\theta) = \emptyset$ on pain

of inconsistency of T, and Tarski's theorem rules out the possibility that $\mathrm{Fix}(\theta) = \omega$. Hence $\mathrm{Fix}(\theta)$ and $\mathrm{Fix}(\neg\theta)$ are disjoint and effectively inseparable, and thus both creative. □

Alternative proof. Let $\theta(x)$ be any formula, let X be an arbitrary r.e. set, and let $\xi(x)$ be a numeration of X. Let, by the parametric diagonal lemma (Fact 2.13), $\phi(x)$ be such that, for all k,

$$\mathrm{T} \vdash \phi(k) \leftrightarrow (\theta(\ulcorner \phi(k) \urcorner) \wedge \xi(k)) \vee (\neg\theta(\ulcorner \phi(k) \urcorner) \wedge \neg\xi(k)).$$

It follows that $k \in X$ iff $\phi(k) \in \mathrm{Fix}(\theta)$, so $\mathrm{Fix}(\theta)$ is complete Σ_1. By Myhill's result, $\mathrm{Fix}(\theta)$ is creative. □

There are plenty of complete Σ_1 sets that are not sets of fixed points over a given theory: if S is an r.e. extension of Q such that $\mathrm{Th(S)} \neq \mathrm{Th(T)}$, then Th(S) is not a set of fixed points over T. This observation follows from the following two results, which are due to Christian Bennet.

Theorem 3.3. If X is an r.e., deductively closed, proper subset of $\mathrm{Th(T)}$, there is no $\theta(x)$ such that $X = \mathrm{Fix}^{\mathrm{T}}(\theta)$.

Proof. Let X be an r.e., deductively closed, proper subset of Th(T), and suppose towards a contradiction that $X = \mathrm{Fix}^{\mathrm{T}}(\theta)$ for some $\theta(x)$. Then there is a sentence ψ that is provable in T but not an element of X. By the diagonal lemma, let ϕ be such that

$$\mathrm{T} \vdash \phi \leftrightarrow \theta(\ulcorner \psi \wedge \phi \urcorner).$$

Since ψ is provable, $\psi \wedge \phi \in \mathrm{Fix}^{\mathrm{T}}(\theta) = X$. But since X is deductively closed, $\psi \in X$, which is a contradiction. □

Definition 3.4. A set X is sufficiently closed if $\psi \in X$ implies $\psi \vee \gamma \in X$, for each sentence γ.

Theorem 3.5. If X is an r.e., sufficiently closed, proper superset of $\mathrm{Th(T)}$, there is no $\theta(x)$ such that $X = \mathrm{Fix}^{\mathrm{T}}(\theta)$.

Proof. Let X be an r.e., sufficiently closed, proper superset of $\mathrm{Th}(\mathrm{T})$, and suppose towards a contradiction that $X = \mathrm{Fix}^{\mathrm{T}}(\theta)$ for some $\theta(x)$. Then there is a sentence $\psi \in X$ that is not provable in T. By the diagonal lemma, let ϕ be such that $\mathrm{T} \vdash \phi \leftrightarrow \neg\theta(\ulcorner \psi \vee \phi \urcorner)$.

Since $\psi \in X$ and X is sufficiently closed, $\psi \vee \phi \in X = \mathrm{Fix}^{\mathrm{T}}(\theta)$. Equivalently, $\mathrm{T} \vdash (\psi \vee \phi) \leftrightarrow \theta(\ulcorner \psi \vee \phi \urcorner)$, whence by construction of ϕ, $\mathrm{T} \vdash (\psi \vee \phi) \leftrightarrow \neg\phi$. By propositional logic, $\mathrm{T} \vdash \psi$, which is impossible. □

Question 3.6. Can the collection of sets of fixed points over T be characterised among the creative sets?

3.2 Counting the number of fixed points

There are at least two ways to count the number of provable fixed points of a formula: the first is to count syntactically different sentences as different fixed points; the other is to count only the number of equivalence classes of fixed points. As all sets of provable fixed points are creative and therefore non-recursive, the number of syntactically distinct fixed points must be infinite, and the more interesting question is to consider the number of equivalence classes involved. The results in this section are easily obtained by applying the results of Bernardi (1981).

Theorem 3.7. If $\theta(x)$ is extensional and has finite range, then $\theta(x)$ is substitutable, whence it has a unique fixed point.

Proof. Let $\theta(x)$ be an extensional formula with finite range. By Corollary 5 of Bernardi (1981), $\theta(x)$ is constant. Hence $\mathrm{T} \vdash \theta(\ulcorner \phi \urcorner) \leftrightarrow \theta(\ulcorner \psi \urcorner)$ for all ϕ, ψ, so $\theta(x)$ is substitutable, and has a unique fixed point. □

For extensional formulae in general, it is possible to show a more vague statement, namely that for each such set of fixed points, there is a recursive enumeration of the equivalence classes contained in it.

Theorem 3.8. If $\theta(x)$ is extensional, then there is a recursive set X such that for every ϕ, $\phi \in \mathrm{Fix}(\theta)$ iff there is a $\xi \in X$ such that $\mathrm{T} \vdash \xi \leftrightarrow \phi$.

Proof. This follows directly from Lemma 1 in Bernardi (1981), the proof of which is a version of Craig's trick. □

The following corollary is a parallel to Bernardi's Theorem 6. However, the class of fixed points does not constitute a partition of ω.

Corollary 3.9. If $I \subseteq \omega$ is such that $\mathscr{F} = \{\mathrm{Fix}(\theta_i) : i \in I\}$ constitutes a partition of ω, then $\mathscr{F}$ is not r.e. without repetition.

Proof. Suppose $\mathscr{F}$ is r.e. without repetition. Then $\varphi \in \mathrm{Fix}(\theta_i)$ iff $\varphi \notin \omega \setminus \mathrm{Fix}(\theta_i)$, whence $\mathrm{Fix}(\theta_i)$ has an r.e. complement, which is not the case. □

3.3 Hierarchical generalisations

Let, for each $\theta(x)$, $\mathrm{Fix}_\Gamma(\theta) = \Gamma \cap \mathrm{Fix}(\theta)$, and for each ψ, $[\psi]_\Gamma = \Gamma \cap [\psi]$. Here the precise definition of the class Γ is of interest. If Γ is defined to be closed under provable equivalence in T, then for each $\theta(x)$ and Γ, $\mathrm{Fix}_\Gamma(\theta) = \mathrm{Fix}(\theta)$. This situation does not arise when using the strictly syntactical definition of the classes Γ, as given in the background chapter. With this restriction in mind, it is possible to prove the following characterisation.

Theorem 3.10 (Lindström, 2003, Exercise 2.28c)**.** If X is an r.e. subset of Γ, then there is a B_n formula $\theta(x)$ such that $\mathrm{Fix}_\Gamma(\theta) = X$.

Hence, the interesting cases are the sets $\mathrm{Fix}_\Gamma(\theta)$, where $\theta(x)$ is itself a Γ formula. From this point on, assume that $\theta(x)$ is of the same complexity as the fixed points asked for. The main result of this section is that Theorem 3.1 can be extended to hold for these bounded sets of fixed points as well – the only additional care needed lies in choosing the correct numerations to keep the complexity down.

Theorem 3.11. Any two disjoint sets of Γ-fixed points of Γ formulae are effectively inseparable.

Proof. Let $\theta(x)$ and $\chi(x)$ be any Γ formulae. Suppose that $\mathrm{Fix}_\Gamma(\theta)$ and $\mathrm{Fix}_\Gamma(\chi)$ are disjoint, and let X and Y be disjoint r.e. subsets of Γ containing $\mathrm{Fix}_\Gamma(\theta)$ and $\mathrm{Fix}_\Gamma(\chi)$, respectively. If $\Gamma \neq \Pi_1$, let, by Fact 2.8,

$\xi_0(x)$ be a Π_1 formula and $\xi_1(x)$ a Σ_1 formula such that for $i = 0, 1$, $\xi_i(x)$ numerates X and $\neg\xi_i(x)$ numerates Y in Q. If $\Gamma = \Pi_1$, switch the complexity of $\xi_0(x)$ and $\xi_1(x)$. Let, by the diagonal lemma, ϕ be such that $\mathrm{T} \vdash \phi \leftrightarrow ((\theta(\ulcorner\phi\urcorner) \wedge \neg\xi_0(\ulcorner\phi\urcorner) \vee (\chi(\ulcorner\phi\urcorner) \wedge \xi_1(\ulcorner\phi\urcorner))$.

The rest of the proof is almost identical to the proof of Theorem 3.1. □

It is easy to see that for an extensional formula $\theta(x)$, $\mathrm{Fix}_\Gamma(\theta)$ is a union of equivalence classes. Hence, if $\theta(x)$ is extensional and $\mathrm{Fix}_\Gamma(\theta) \neq \Gamma$, there must be at least one Γ-equivalence class outside $\mathrm{Fix}_\Gamma(\theta)$. By Theorem 3.11 it follows that whenever $\theta(x)$ is extensional, $\mathrm{Fix}_\Gamma(\theta)$ is creative.

The next result gives a sufficient condition for a bounded r.e. set to be a set of fixed points. Note that the condition is not necessary: any equivalence class $\neq \top$ satisfies the consequent, but not the antecedent of theorem. For the statement of the theorem, the following definition is convenient.

Definition 3.12. A set X has a lower bound in T iff there is a non-refutable sentence ϕ such that $\mathrm{T} \vdash \phi \rightarrow \psi$ for all $\psi \in X$.

Theorem 3.13. Suppose that T is a consistent, r.e. extension of $\mathrm{I}\Delta_0 + \exp$. Let $X \subset \Gamma$ be an r.e. set such that X^c has a lower bound. There is then a Γ formula $\theta(x)$ such that $\mathrm{Fix}^{\mathrm{T}}_\Gamma(\theta) = X$.

Proof. Let X be any r.e. subset of Γ such that X^c has a lower bound ϕ; then $\mathrm{T} + \phi$ is consistent. If $\Gamma \neq \Pi_1$, let by Fact 2.10, $\xi(x)$ be a Σ_1 formula such that if $k \in X$, then $\mathrm{T} \vdash \xi(k)$, and if $k \notin X$, then $\xi(k) \notin \mathrm{Th}(\mathrm{T} + \phi)$. If $\Gamma = \Pi_1$, choose $\xi(x)$ as a Π_1 formula instead. Let $\theta(x) := \mathrm{Tr}_\Gamma(x) \wedge \xi(x)$.

Suppose $\psi \in X$. Then $\mathrm{T} \vdash \xi(\ulcorner\psi\urcorner)$ and $\mathrm{T} + \psi \vdash \mathrm{Tr}_\Gamma(\ulcorner\psi\urcorner) \wedge \xi(\ulcorner\psi\urcorner)$. Hence $\mathrm{T} \vdash \psi \rightarrow \theta(\ulcorner\psi\urcorner)$. Moreover, $\mathrm{T} + \theta(\ulcorner\psi\urcorner) \vdash \mathrm{Tr}_\Gamma(\ulcorner\psi\urcorner)$, so it follows that every element of X is a fixed point of $\theta(x)$.

Suppose that ψ is any Γ sentence such that $\psi \notin X$. Then by the choice of $\xi(x)$, $\mathrm{T} + \phi \nvdash \xi(\ulcorner\psi\urcorner)$, so $\mathrm{T} + \phi + \neg\xi(\ulcorner\psi\urcorner)$ is consistent and proves $\neg\theta(\ulcorner\psi\urcorner)$. Suppose now, for a contradiction, that $\psi \in \mathrm{Fix}_\Gamma(\theta)$. Then $\mathrm{T} + \phi + \neg\xi(\ulcorner\psi\urcorner) \vdash \neg\psi$, so by propositional logic $\mathrm{T} + \phi + \psi \vdash \xi(\ulcorner\psi\urcorner)$. But ϕ is a lower bound on X^c, whence $\mathrm{T} \vdash \phi \rightarrow \psi$. It follows that $\mathrm{T} + \phi \vdash \xi(\ulcorner\psi\urcorner)$, contradicting the choice of $\xi(x)$. □

The next result shows that it is possible to successively remove recursive sets from Γ, obtaining infinitely many recursive sets of Γ-fixed points. By inseparability of disjoint sets of fixed points, each recursive set of fixed points intersects every set of Γ-fixed points (and therefore also every Γ-equivalence class) non-recursively. It also follows that any formula with a recursive set of fixed points ($\neq \Gamma$) must be non-extensional.

Theorem 3.14. Let $\theta(x)$ be any Γ formula, and let $A \subset \Gamma$ be a recursive set such that $\mathrm{Fix}_\Gamma(\tau) \cap A = \emptyset$ for some $\tau(x) \in \Gamma$. Then there is a formula $\chi(x) \in \Gamma$ such that $\mathrm{Fix}_\Gamma(\chi) = \mathrm{Fix}_\Gamma(\theta) \setminus A$.

Proof. Let $\theta(x)$ and A be as above. If $\Gamma \neq \Pi_1$, let by Fact 2.8 $\alpha_0(x)$ be a Π_1 formula and $\alpha_1(x)$ a Σ_1 formula such that $\alpha_i(x)$ binumerates A in Q. If $\Gamma = \Pi_1$, switch the complexities of $\alpha_0(x)$ and $\alpha_1(x)$. Let $\chi(x) := (\theta(x) \wedge \neg\alpha_0(x)) \vee (\tau(x) \wedge \alpha_1(x))$, which is then a Γ formula.

Suppose that $\phi \in A$. Then it follows that $\mathrm{T} \vdash \alpha_0(\ulcorner\phi\urcorner) \wedge \alpha_1(\ulcorner\phi\urcorner)$, so $\mathrm{T} \vdash \chi(\ulcorner\phi\urcorner) \leftrightarrow \tau(\ulcorner\phi\urcorner)$. Suppose further that $\phi \in \mathrm{Fix}_\Gamma(\chi)$. Then $\mathrm{T} \vdash \phi \leftrightarrow \tau(\ulcorner\phi\urcorner)$, but A is assumed to be disjoint from $\mathrm{Fix}_\Gamma(\tau)$, a contradiction.

Now suppose $\phi \notin A$. Then it follows that $\mathrm{T} \vdash \neg\alpha_0(\ulcorner\phi\urcorner) \wedge \neg\alpha_1(\ulcorner\phi\urcorner)$, so $\mathrm{T} \vdash \chi(\ulcorner\phi\urcorner) \leftrightarrow \theta(\ulcorner\phi\urcorner)$. Hence $\phi \in \mathrm{Fix}_\Gamma(\chi)$ iff $\phi \in \mathrm{Fix}_\Gamma(\theta)$, so $\mathrm{Fix}_\Gamma(\chi) = \mathrm{Fix}_\Gamma(\theta) \setminus A$. □

3.4 Algebraic properties

This section makes essential use of partial truth predicates, so it is assumed that $\mathrm{T} \vdash \mathrm{I}\Delta_0 + \mathrm{exp}$. For each Γ, there is then, up to provable equivalence in T, precisely one Γ formula whose set of provable Γ-fixed points is equal to Γ, namely $\mathrm{Tr}_\Gamma(x)$. Let $\mathscr{R}_\Gamma$ be the set of recursive subsets of Γ. It is easy to see that $\mathscr{R}_\Gamma$ forms a Boolean algebra when ordered under set inclusion.

Theorem 3.15. The set

$$\mathscr{F}_\Gamma = \{X \subseteq \Gamma : X \text{ is recursive and } \exists\theta(x) \in \Gamma \text{ s.t. } X = \mathrm{Fix}_\Gamma(\theta)\},$$

that is, the set of recursive sets of Γ-fixed points of Γ formulae, forms a non-principal filter on $\mathscr{R}_\Gamma$.

The theorem gives a negative answer to a question left open in Blanck (2011), and it follows from the following sequence of lemmas.

Lemma 3.16. Let F be any finite subset of Γ. Then $\Gamma \setminus F$ is a recursive set of Γ-fixed points, and F is not a set of Γ-fixed points.

Proof. Let F be any finite subset of Γ. Then by Theorem 3.14, $\Gamma \setminus F$ is a recursive set of Γ-fixed points. If F were a set of Γ-fixed points, then F would be both recursive and creative at the same time. □

Lemma 3.17. If $A \in \mathscr{F}_\Gamma$ and $A \subseteq B \in \mathscr{R}_\Gamma$, then $B \in \mathscr{F}_\Gamma$.

Proof. Suppose $A \in \mathscr{F}_\Gamma$, then there is a $\theta(x) \in \Gamma$ such that $A = \mathrm{Fix}_\Gamma(\theta)$. Let $B \supseteq A$ be a recursive subset of Γ. If $\Gamma \neq \Pi_1$, let $\beta_0(x)$ be a Π_1 formula and $\beta_1(x)$ be a Σ_1 formula such that $\beta_i(x)$ binumerates B in Q. If $\Gamma = \Pi_1$, switch the complexities of $\beta_0(x)$ and $\beta_1(x)$. Let

$$\psi(x) := (\mathrm{Tr}_\Gamma(x) \wedge \beta_0(x)) \vee (\theta(x) \wedge \neg\beta_1(x)),$$

which is then a Γ formula.

Suppose $\phi \in B$. Then $\mathrm{T} \vdash \beta_0(\ulcorner\phi\urcorner) \wedge \beta_1(\ulcorner\phi\urcorner)$, so $\mathrm{T} \vdash \phi \leftrightarrow \psi(\ulcorner\phi\urcorner)$, and $\phi \in \mathrm{Fix}_\Gamma(\psi)$.

Suppose $\phi \notin B$. Then it follows that $\mathrm{T} \vdash \neg\beta_0(\ulcorner\phi\urcorner) \wedge \neg\beta_1(\ulcorner\phi\urcorner)$, so $\mathrm{T} \vdash \psi(\ulcorner\phi\urcorner) \leftrightarrow \theta(\ulcorner\phi\urcorner)$. So if ϕ is a fixed point of $\psi(x)$, it is also a fixed point of $\theta(x)$, which contradicts the fact that B was chosen to contain $\mathrm{Fix}_\Gamma(\theta)$. Hence $\phi \notin \mathrm{Fix}_\Gamma(\psi)$. □

Lemma 3.18. If $A, B \in \mathscr{F}_\Gamma$, then $A \cap B \in \mathscr{F}_\Gamma$.

Proof. Let $A, B \in \mathscr{F}_\Gamma$ be such that $A = \mathrm{Fix}_\Gamma(\theta)$ and $B = \mathrm{Fix}_\Gamma(\psi)$. If $\Gamma \neq \Pi_1$, let $\beta_i(x)$, for $i = 0, 1$ be binumerations of B as above. If $\Gamma = \Pi_1$, switch the complexities of $\beta_0(x)$ and $\beta_1(x)$. Let

$$\chi(x) := (\beta_1(x) \wedge \theta(x)) \vee (\neg\beta_0(x) \wedge \psi(x)).$$

Suppose $\phi \in B$. Then $\mathrm{T} \vdash \theta(\ulcorner\phi\urcorner) \leftrightarrow \chi(\ulcorner\phi\urcorner)$, so $\phi \in \mathrm{Fix}_\Gamma(\chi)$ iff $\phi \in \mathrm{Fix}_\Gamma(\theta) = A$. Hence: the only elements of B that are fixed points of $\chi(x)$ are the elements of A.

Suppose $\phi \notin B$. Then $\mathrm{T} \vdash \chi(\ulcorner\phi\urcorner) \leftrightarrow \psi(\ulcorner\phi\urcorner)$. Since $B = \mathrm{Fix}_\Gamma(\psi)$, ϕ can not be a fixed point of $\chi(x)$. □

Question 3.19. Is $\mathscr{F}_\Gamma$ an ultrafilter on $\mathscr{R}_\Gamma$?

It is easy to see that if A is a recursive subset of Γ, then *at most* one of A and $\Gamma \setminus A$ is in $\mathscr{F}_\Gamma$. If *at least* one of those is in $\mathscr{F}_\Gamma$, then $\mathscr{F}_\Gamma$ is an ultrafilter on $\mathscr{R}_\Gamma$.

4 Flexibility in fragments

This chapter serves multiple purposes. First, it introduces the central notions of independent and flexible formulae, and surveys the literature on the topic, from its inception in the early 1960s until 2016. A second purpose is to introduce a general method, due to Kripke (1962), for constructing independent and flexible formulae, which has a number of applications in coming chapters.

Apart from these purposes, the opportunity is also taken to relate the classic results on flexibility and independence to the modern research in fragments of arithmetic, gauging the amount of induction needed to prove the various results. As is pointed out by Beklemishev (1998), the question of which of these results hold in theories weaker than PA has been largely ignored. Hence, this chapter attempts to partly rectify this situation.

4.1 Definitions and motivation

The central notions of this chapter and the next are those of *independence* and *flexibility*, which are due to Mostowski (1961) and Kripke (1962), respectively. These terms are not used univocally in the literature, as is explained below, but for the purpose of this thesis the following definitions are adopted. Let, for any formula ϕ, $\phi^0 := \phi$ and $\phi^1 := \neg\phi$.

Definition 4.1. A formula $\phi(x)$ is *independent* over a theory T, iff for every function $f \in {}^{\omega}2$, the theory $\mathrm{T} + \{\phi(n)^{f(n)} : n \in \omega\}$ is consistent.

Definition 4.2. A formula $\gamma(x)$ is *flexible* for class of formulae X over a theory T iff for every $\xi(x) \in X$, the theory $\mathrm{T} + \forall x(\gamma(x) \leftrightarrow \xi(x))$ is consistent.

The definition of independence used here is equivalent to the one used by Mostowski, although he calls such formulae *free*. It seems, however,

that the habit of using 'independent', or variations thereof, for these formulae emerged quickly: both Feferman et al. (1962) and Scott (1962) use this terminology. Later sources conforming to this usage are, e.g., Lindström (1984), Smoryński (1984), Sommaruga-Rosolemos (1991), Lindström (2003), and Kikuchi and Kurahashi (2016). A notion of an 'independent set', which is related to that of independent formulae, also appears in Harary (1961), Kripke (1962), and Myhill (1972). Proofs essentially establishing the existence of independent formulae abound in the literature (Mostowski, 1961, Theorem 2; Kripke, 1962, Corollary 1.1; Myhill, 1972, Remark 2). A similar result also appears in Jeroslow (1971, Lemma 3.1), and most recently in Hamkins (2016).

The term 'flexible' has a more complicated history. It is introduced by Kripke (1962), although the meaning is somewhat different (see Section 4.3 below). Formulae of the type described above appear in, e.g., Visser (1980), Montagna (1982), Lindström (1984), and Lindström (2003) without a definite name, while Sommaruga-Rosolemos (1991) uses the term 'flexible' without an explicit definition. The standard work Hájek and Pudlák (1993), uses 'flexible' for what here is called 'independent', which is at odds with the other sources listed above, and a possible source of confusion.

Recall that Gödel's first incompleteness theorem exhibits a true but T-unprovable Π_1 sentence γ, thus instantiating the incompleteness of T at the lowest possible level: since T is Σ_1-complete, truth and provability coincide for Σ_1 sentences, and no incompleteness is possible at that level. It is trivial to construct a Π_1 formula with similar properties, e.g., a formula $\gamma(x)$ such that for each $n \in \omega$, $\gamma(n)$ is true but unprovable in T, and similarly a Π_1 formula $\rho(x)$ such that for every $n \in \omega$, $\mathrm{T} \nvdash \rho(n)$ and $\mathrm{T} \nvdash \neg\rho(n)$.

The condition on $\phi(n)$ in Definition 4.1 is strictly stronger than the trivial generalisations discussed in the preceding paragraph: it is equivalent to the condition that the only propositional combinations of sentences of the form $\phi(n)$ provable in T are the tautologies. For a simple example, whenever $\phi(x)$ is independent over T, and $m \neq n$, then it follows that $\mathrm{T} \nvdash \phi(m) \vee \phi(n)$ and $\mathrm{T} \nvdash \neg\phi(m) \vee \neg\phi(n)$. Elaborating on this example, if $\phi(n)$ is Π_1, then $\phi(n)$ must be true, since otherwise $\neg\phi(n)$ is

a true Σ_1 sentence, and therefore provable in T. With this in mind, it is straightforward to see that the existence of an independent Π_1 formula (or equivalently, an independent Σ_1 formula) is a generalisation of both Gödel's and Rosser's incompleteness theorems, a perspective which is also adopted by Mostowski (1961). Hence it is clear that the study of independent and flexible formulae can be regarded as a study of the important incompleteness phenomenon of formal arithmetical theories.

Although the notions of independent and flexible formulae are similar, to the point that every flexible formula also is independent (Kripke, 1962, Corollary 1.1; see also Theorem 4.5 below), there are still important differences between them. A key distinguishing feature can be articulated as follows. Given an independent formula $\xi(x)$, and any prescribed set $A \subset \omega$, the completeness theorem gives rise to a model $\mathcal{M}$ of arithmetic such that $\{n \in \omega : \mathcal{M} \models \xi(n)\} = A$. No information is, however, gotten about the truth values of any specific non-standard instances of $\xi(x)$. Moreover, supposing that $\xi(x) \in \Sigma_n$ and $\mathcal{M} \models \mathrm{I}\Sigma_n$, if A is cofinal in ω, then $\xi(x)$ holds on a downwards cofinal subset of the non-standard part of $\mathcal{M}$. This follows from the overspill principle, together with the fact that in a model of $\mathrm{I}\Sigma_n$, every bounded, Σ_n-definable subset of the model has a maximum.

For a flexible formula $\gamma(x)$, on the other hand, the completeness theorem ensures the existence of a model $\mathcal{M}$ in which the extension of $\gamma(x)$ agrees with that of any desired formula $\sigma(x)$ (of a suitable complexity class). This does not give any absolute control of the actual content of the extension of $\gamma(x)$ in $\mathcal{M}$; the crucial point being that the extension of $\gamma(x)$ *as calculated within* $\mathcal{M}$ agrees with the extension of $\sigma(x)$, also *as calculated within* $\mathcal{M}$. If any instance of $\sigma(x)$ is undecidable in T, this instance may very well have different truth values in $\mathbb{N}$ and $\mathcal{M}$. For a concrete example, the extension of the Σ_1 formula $\neg\mathrm{Con}_\mathrm{T} \wedge x = x$ is empty as calculated within $\mathbb{N}$, while it can have two radically different extensions in $\mathcal{M}$: either the empty set, or the whole domain, depending on whether or not $\mathcal{M}$ satisfies Con_T.

4.2 Mostowski's and Kripke's theorems

Having discussed in some detail the properties and merits of independent and flexible formulae, it is time to prove that both kinds of formulae actually exist. Below, a detailed proof of Kripke's theorem on the existence of flexible formulae is given.[5] The proof method used by Kripke, and detailed below, is very versatile, and is applied on numerous occasions throughout this thesis.

Theorem 4.3 (Essentially Kripke, 1962, Theorem 1)**.** Suppose that T is a consistent, r.e. extension of $\mathrm{I}\Delta_0 + \mathrm{exp}$. For each $n > 0$, there is a Σ_n formula $\gamma(x)$ such that for each $\sigma(x) \in \Sigma_n$, the theory

$$\mathrm{T} + \forall x(\gamma(x) \leftrightarrow \sigma(x))$$

is consistent.

Note that, in one sense, this theorem is the best possible, since it immediately leads to a contradiction to assume the existence of a formula flexible for a higher complexity than its own. Suppose, e.g., that there is a Σ_1 formula $\gamma(x)$ that is flexible for Σ_2 over T. It is then possible to choose a formula $\sigma(x) \in \Sigma_2$ such that $\sigma(x)$ is provably equivalent to $\neg\gamma(x)$. By assumption, $\mathrm{T} + \forall x(\gamma(x) \leftrightarrow \sigma(x))$ is consistent, which is impossible by the particular choice of $\sigma(x)$.

As in Kripke's original proof, the theorem above is derived from an innocent-looking lemma, establishing the existence of a partial recursive function f with index e that is such that Q does not refute any sentences of the form $f(e) = k$. Strictly speaking, the expression $f(e) = k$ is not a sentence in $\mathscr{L}_A$, but may instead be regarded as shorthand for the $\mathscr{L}_A$-sentence $R(e, e, k) \wedge \exists! z R(e, e, z)$. Here $R(x, y, z)$ is a Σ_1 formula strongly representing the relation $\Phi(x, y) = z$, where $\Phi(x, y)$ is a function that is universal for partial recursive functions.

[5] Kripke states the theorem for extensions of what is essentially Robinson's arithmetic Q, but with a different notion of flexibility. More on this in Section 4.3. In Visser (1980) and Lindström (2003), the theorem is proved for extensions of PA. Sommaruga-Rosolemos (1991) states the theorem for what he calls primitive recursive arithmetic, PRA, but his theory is augmented with induction for Σ_1 formulae, thus making it equivalent to (a conservative extension of) $\mathrm{I}\Sigma_1$.

Lemma 4.4 (Kripke, 1962, Lemma 1)**.** Suppose that T is a consistent, r.e. extension of Q. There is a partial recursive function f with index e (which depends on T), such that for every k, the theory $\mathrm{T}+f(e)=k$ is consistent.

Proof. Define $f(x)$ by the stipulation that $f(n)=k$ iff

$$\mathrm{T}\vdash\neg(R(n,n,k)\wedge\exists!zR(n,n,z)).$$

If more than one sentence of this form is provable in T, pick the one whose proof has the least Gödel number. This assures that $f(x)$ is well-defined, and since T (and therefore also provability in T) is r.e., $f(x)$ is a partial recursive function. Let e be an index for f.

Pick k, and suppose, for a contradiction, that $\mathrm{T}+f(e)=k$ is inconsistent. Then it follows that $\mathrm{T}\vdash f(e)\neq k$, or equivalently, that $\mathrm{T}\vdash\neg(R(e,e,k)\wedge\exists!zR(e,e,z))$. But then $f(e)=k$ by definition, so $\mathrm{T}\vdash R(e,e,k)\wedge\exists!zR(e,e,z)$, contradicting the consistency of T. □

Proof of Theorem 4.3. Let T be a consistent, r.e. extension of $\mathrm{I}\Delta_0+\mathrm{exp}$. Let e be as in the proof of Lemma 4.4, and recall that e depends on the actual choice of T. Pick $n>0$, and let $\gamma(x):=\exists z(R(e,e,z)\wedge\mathrm{Sat}_{\Sigma_n}(z,x))$. Let $\sigma(x)$ be any Σ_n formula.

By Lemma 4.4, the theory $\mathrm{T}+f(e)=\ulcorner\sigma\urcorner$ is consistent, so reason in that theory:

> If $\gamma(x)$, then for some z, $R(e,e,z)\wedge\mathrm{Sat}_{\Sigma_n}(z,x)$. But this z is unique, and $R(e,e,\ulcorner\sigma\urcorner)$ holds, so $\mathrm{Sat}_{\Sigma_n}(\ulcorner\sigma\urcorner,x)$. Hence $\sigma(x)$.
>
> If $\sigma(x)$, then $\mathrm{Sat}_{\Sigma_n}(\ulcorner\sigma\urcorner,x)$. Since $R(e,e,\ulcorner\sigma\urcorner)$, $\gamma(x)$ follows by one application of $\exists$-introduction.

This proves that $\mathrm{T}+f(e)=\ulcorner\sigma\urcorner\vdash\forall x(\gamma(x)\leftrightarrow\sigma(x))$, and since the theory $\mathrm{T}+f(e)=\ulcorner\sigma\urcorner$ is consistent, $\mathrm{T}+\forall x(\gamma(x)\leftrightarrow\sigma(x))$ is also consistent. □

This proof of Kripke's theorem and lemma depends essentially on the recursion theorem for defining $f(x)$, the representability of partial recursive functions in Q, and the existence of partial satisfaction predicates. The

ingenious part of Kripke's proof is to define the flexible Σ_n formula $\gamma(x)$ as expressing 'x satisfies the Σ_n formula whose Gödel number is output by $f(e)$', which is possible to express formally by using the partial satisfaction predicate for Σ_n. Since $f(e)$ can consistently assume any value, it must also be consistent that $\gamma(x)$ coincides with any desired formula of a suitable complexity.

This trick of Kripke's, feeding the output of a function to a partial satisfaction predicate or other normal form theorem, is used on a number of occasions later on. For an immediate example, the existence of a Π_n formula that is flexible for Π_n over T follows by letting e be an index for f as above, and by letting $\gamma(x) := \forall z(R(e, e, z) \to \mathrm{Sat}_{\Pi_n}(z, x))$.

In Kripke (1962), a result is proved (Corollary 1.1) which is, as pointed out by the referee of Kripke's paper, essentially equivalent to Mostowski's theorem on the existence of an independent Σ_1 formula. Mostowski's original proof of this result goes via a quite sophisticated witness comparison argument (Mostowski, 1961, Theorem 2; Lindström, 2003, Theorem 2.9). Kripke instead establishes that the instances of a Σ_1-flexible formula form an *independent set over* T: a set A such that, for every subset B of A, the theory $\mathrm{T} + B + \{\neg\alpha : \alpha \in A \setminus B\}$ is consistent. In fact, a formula $\xi(x)$ is independent iff $\{\xi(n) : n \in \omega\}$ is an independent set. Since the method of showing independence by using a flexible formula has a number of applications in this thesis, the opportunity is taken to give a perspicuous proof.

Theorem 4.5 (Mostowski, 1961)**.** Suppose that T is a consistent, r.e. extension of $\mathrm{I}\Delta_0 + \exp$. Then there is a Σ_1 formula $\xi(x)$ that is independent over T.

Proof. Let T be a consistent r.e. extension of $\mathrm{I}\Delta_0 + \exp$, let f be any function in ${}^{\omega}2$, let $\xi(x)$ be a Σ_1-flexible Σ_1 formula as in Theorem 4.3, and let $X = \{\xi(n)^{f(n)} : n \in \omega\}$.

By compactness, it suffices to show that for each finite subset A of X, the theory $\mathrm{T} + A$ is consistent. Let A be any finite subset of X, let B be the set $\{n : \xi(n) \in A\}$, and let $\beta(x)$ be a Σ_1 binumeration of B in Q. By Theorem 4.3, the theory $\mathrm{T} + \forall x(\xi(x) \leftrightarrow \beta(x))$ is consistent.

For each $n \in \omega$:

1. If $\xi(n) \in A$, then $n \in B$, and since $\beta(x)$ binumerates B in Q, $\mathrm{T} \vdash \beta(n)$. But then $\mathrm{T} + \forall x(\xi(x) \leftrightarrow \beta(x)) \vdash \xi(n)$.

2. If $\neg\xi(n) \in A$, then $n \notin B$, so $\mathrm{T} \vdash \neg\beta(n)$. It then follows that $\mathrm{T} + \forall x(\xi(x) \leftrightarrow \beta(x)) \vdash \neg\xi(n)$.

Since the consistent theory $\mathrm{T} + \forall x(\xi(x) \leftrightarrow \beta(x))$ proves all the sentences in A, $\mathrm{T} + A$ is consistent. Hence $\mathrm{T} + X$ is consistent. □

The upcoming section elaborates on the modifications required to construct the analogue of flexible formulae in theories not allowing for partial satisfaction predicates.

4.3 Flexibility and independence in Robinson's arithmetic

The attentive reader reacts to the fact that the proof of Mostowski's theorem, as presented above, is not stated for extensions of Q, while on the other hand, it is well known that his theorem *does* apply to such extensions. Mostowski's original proof uses a sophisticated witness comparison argument rather than partial satisfaction predicates, and the use of the former generates no additional difficulties in Q. It it still, however, possible to prove Mostowski's theorem for Q using Kripke's method, by reverting to Kripke's own, nowadays somewhat peculiar-looking, notion of flexibility. The modifications needed to obtain this rederivation in Q is the topic of this subsection. Kripke's original definition of flexibility can be expressed as follows.

Definition 4.6. A formula $\gamma(x)$ is flexible in Kripke's sense for Σ_n over T if every Σ_n relation can be defined by a Σ_n formula $\sigma(x)$ that is such that $\mathrm{T} + \forall x(\gamma(x) \leftrightarrow \sigma(x))$ is consistent.

By Kleene's normal form theorem, every Σ_n-definable relation has a definition on Kleene normal form. Although these formulae are similar to partial satisfaction predicates, their satisfaction-like properties can not

be proved in Q, which is what motivates Kripke's formulation above. By replacing the partial satisfaction predicate in the definition of $\gamma(x)$ with a formula on a suitable Kleene normal form, Kripke's original theorem on flexible formulae is obtained. E.g., if $n = 1$, $\gamma(x)$ can be defined as $\exists z(R(e, e, z) \wedge \exists y T(z, x, y))$.

Theorem 4.7 (Kripke, 1962, Theorem 1)**.** Suppose that T is a consistent, r.e. extension of Q. For each $n > 0$, there is a Σ_n formula $\gamma(x)$ that is flexible in Kripke's sense for Σ_n over T.

Mostowski's theorem now follows from the theorem above, by a proof similar to the proof of Theorem 4.5. The crucial point is that the Σ_1 binumeration $\beta(x)$ used in that proof can here be replaced by a formula $\exists y T(k, x, y)$ for a suitable choice of k.

Another method for constructing a set independent over a given extension of Q is exhibited in Remark 2 of Kripke (1962), and similar constructions appear in Jensen and Ehrenfeucht (1976), Lindström (1979), Kikuchi and Kurahashi (2016), and Hamkins (2016). Let, by Rosser's theorem, ρ_ϵ be a Π_1 sentence that is undecidable in T. Continue by letting ρ_0 be a Π_1 sentence undecidable in $\mathrm{T} + \rho_\epsilon$ and ρ_1 a Π_1 sentence undecidable in $\mathrm{T} + \neg\rho_\epsilon$. Let ρ_{00}, ρ_{01}, ρ_{10}, ρ_{11} be Π_1 sentences undecidable in $\mathrm{T} + \rho_\epsilon \wedge \rho_0$, $\mathrm{T} + \rho_\epsilon \wedge \neg\rho_0$, $\mathrm{T} + \neg\rho_\epsilon \wedge \rho_1$, and $\mathrm{T} + \neg\rho_\epsilon \wedge \neg\rho_1$, respectively. In general, if s is a binary string, the Π_1 sentence ρ_s is undecidable in T plus the appropriate Rosser sentences or their negations, picked out by the pattern prescribed by s, where, as usual, $\phi^0 = \phi$ and $\phi^1 = \neg\phi$. Continue adding Rosser sentences in this fashion, to obtain the binary branching *Rosser tree*, in which every branch can be picked out by a binary sequence, and the nodes along each such branch together with T constitute a consistent theory. Then the set R, defined as

$$\begin{aligned}
&\{\rho_\epsilon,\\
&\ (\rho_\epsilon \wedge \rho_0) \vee (\neg\rho_\epsilon \wedge \rho_1),\\
&\ (\rho_\epsilon \wedge \rho_0 \wedge \rho_{00}) \vee (\rho_\epsilon \wedge \neg\rho_0 \wedge \rho_{01}) \vee\\
&\ (\neg\rho_\epsilon \wedge \rho_1 \wedge \rho_{10}) \vee (\neg\rho_\epsilon \wedge \neg\rho_1 \wedge \rho_{11}), \dots\}
\end{aligned}$$

is independent in Kripke's sense.

It is straightforward but tedious to prove this, due to the mere length of the sentences in R. The idea is that for any combination of positive and negative instances of the elements of R, there is a corresponding path through the Rosser tree that proves all of these sentences. As an illustration, let r_0, r_1 be the first two sentences of R; then $\mathrm{T} + \rho_\epsilon + \rho_0 \vdash r_0 \wedge r_1$, while $\mathrm{T} + \rho_\epsilon + \neg\rho_0 \vdash r_0 \wedge \neg r_1$. Similarly, $\mathrm{T} + \neg\rho_\epsilon + \rho_1 \vdash \neg r_0 \wedge r_1$ and $\mathrm{T} + \neg\rho_\epsilon + \rho_0 \vdash \neg r_0 \wedge \neg r_1$. Since all of these extended theories are consistent by construction, $\{r_0, r_1\}$ is independent over T. This argument is easily extended to show that any finite subset of R is consistent, which by compactness is sufficient to show that R is independent over T.

Let $r_0, r_1, r_2, \ldots$ be an enumeration of all the sentences in R. Since all of these sentences are B_1, let $\phi(x)$ be a Δ_2 formula such that

$$\mathrm{I}\Delta_0 + \exp \vdash \phi(k) \leftrightarrow \mathrm{Tr}_{\mathrm{B}_1}(r_k)$$

for all k. Then $\phi(x)$ is independent over T; see also Theorem 4.11 below for an improvement of this result.

4.4 Refinements

The proof of Theorem 4.3 can be modified in a number of ways to yield similar, and stronger, conclusions. For example, it is possible to construct a recursive function f_0 such that for any choice of $k \in \omega$, the theory $\mathrm{T} + f_0(0) = k$ is consistent. More generally, for each $n \in \omega$, there is a recursive function f_n such that for any choice of $k \in \omega$, the theory $\mathrm{T} + f_n(n) = k$ is consistent. One such function comes to use in Section 4.6.

Another generalisation, suggested by Ali Enayat, is of the hierarchical kind. Salehi and Seraji (2016) and Kikuchi and Kurahashi (20xx) discuss generalisations of the incompleteness theorems to Σ_{n+1}-definable theories. In particular, both papers show that for every Σ_{n+1}-definable, Σ_n-sound theory there is a true Π_{n+1} sentence undecidable in the theory (Kikuchi and Kurahashi, 20xx, Theorem 4.21; Salehi and Seraji, 2016, Theorem 2.5). Theorems 4.8 and 4.9 below are in turn generalisations of these results, in that they provide Σ_{n+1} formulae that are flexible or independent over the Σ_{n+1}-definable theory T, rather than just exhibiting a true undecidable Π_{n+1} sentence. Cf. the discussion of independent formulae in Section 4.1.

Theorem 4.8. Suppose that T is a Σ_{n+1}-definable, Σ_n-sound extension of $\mathrm{I}\Delta_0 + \exp$. There is a Σ_{n+1} formula $\gamma(x)$ such that for each $\sigma(x) \in \Sigma_{n+1}$, the theory $\mathrm{T} + \forall x(\gamma(x) \leftrightarrow \sigma(x))$ is consistent.

Proof. Let T be a Σ_{n+1}-definable, Σ_n-sound extension of $\mathrm{I}\Delta_0 + \exp$. By Craig's trick (Fact 2.11), T can without loss of generality be assumed to be Π_n-definable.[6]

By Kleene's enumeration theorem, there is a function $\Phi^n(x, y)$, recursive in $\emptyset^{(n)}$, that is universal for functions recursive in $\emptyset^{(n)}$. By Fact 2.51, let $R^n(x, y, z)$ be a Σ_{n+1} formula strongly representing $\Phi^n(x, y) = z$ in $\mathrm{Q} + \mathrm{Th}_{\Sigma_{n+1}}(\mathbb{N})$.

Let $f(m) = k$ iff $\mathrm{T} \vdash \neg(R^n(m, m, k) \wedge \exists! z R^n(m, m, z))$. If more than one sentence of this form is provable in T, pick the one whose proof has the least Gödel number. Since T can be assumed to be Π_n-definable, provability in T, and therefore also the relation $f(x) = y$, is Δ_{n+1} and recursive in $\emptyset^{(n)}$ by Post's theorem.

Let e be an index for f, and suppose, for a contradiction, that for some $k \in \omega$, the theory $\mathrm{T} + R^n(e, e, k) \wedge \exists! z R^n(e, e, z)$ is inconsistent. Then $\mathrm{T} \vdash \neg(R^n(e, e, k) \wedge \exists! z R^n(e, e, z))$, so by definition of f, $f(e) = k$. It then follows that $\mathrm{T} + \mathrm{Th}_{\Sigma_{n+1}}(\mathbb{N}) \vdash R^n(e, e, k) \wedge \exists! z R^n(e, e, z)$, so $\mathrm{T} + \mathrm{Th}_{\Sigma_{n+1}}(\mathbb{N})$ must be inconsistent. But T is Σ_n-sound, so by Fact 2.32, $\mathrm{T} + \mathrm{Th}_{\Sigma_{n+1}}(\mathbb{N})$ is consistent. This is the desired contradiction, whence $\mathrm{T} + R^n(e, e, k) \wedge \exists! z R^n(e, e, z)$ is consistent for each $k \in \omega$.

Let $\sigma(x)$ be any Σ_{n+1} formula. By letting $\gamma(x)$ be the Σ_{n+1} formula $\exists z(R^n(e, e, z) \wedge \mathrm{Sat}_{\Sigma_{n+1}}(z, x))$, it follows that $\mathrm{T} + \forall x(\gamma(x) \leftrightarrow \sigma(x))$ is consistent. □

A hierarchical generalisation of Theorem 4.5 follows by an easy modification of the proof of that theorem.

Theorem 4.9. Suppose that T is a Σ_{n+1}-definable, Σ_n-sound extension of $\mathrm{I}\Delta_0 + \exp$. There is then a Σ_{n+1} formula $\xi(x)$ such that for any $f \in {}^{\omega}2$, the theory

$$\mathrm{T} + \{\xi(k)^{f(k)} : k \in \omega\}$$

is consistent.

[6] If $n = 0$, the proof of Theorem 4.3 goes through as it stands.

Remark 4.10. By a further modification of the proof of this theorem, in the spirit of the discussion in Section 4.3, the assumption can be weakened to $\mathrm{T} \vdash \mathrm{Q}$.

Another kind of generalisation is due to Montagna (1982, Corollary 1).

Theorem 4.11. Let T be a consistent, r.e. extension of $\mathrm{I}\Delta_0 + \mathrm{exp}$. There is a Δ_{n+1} formula $\delta(x)$ such that for every B_n formula $\beta(x)$, the theory $\mathrm{T} + \forall x(\delta(x) \leftrightarrow \beta(x))$ is consistent.

This can be proved by a proof almost identical to that of Theorem 4.3, using the fact that there is a Δ_{n+1} satisfaction predicate for B_n formulae. Montagna's original proof is stated for extensions of PA, and imitates the quite different method used by Visser (1980) to prove the following general theorem.[7]

Theorem 4.12 (The Gödel-Rosser-Mostowski-Myhill-Kripke-Visser Theorem)**.** Suppose that $\{\mathrm{T}_i : i \in \omega\}$ is an r.e. family of consistent, r.e. extensions of $\mathrm{I}\Delta_0 + \mathrm{exp}$, and that $\{X_i : i \in \omega\}$ is an r.e. family of r.e. sets of sentences such that $\mathrm{T}_i \nvdash \xi$ for all $\xi \in X_i$.

For each $n > 0$, there is then a Σ_n formula $\gamma(x)$, such that for every $\sigma(x) \in \Sigma_n$, every $i \in \omega$, and every $\xi \in X_i$,

$$\mathrm{T}_i + \forall x(\gamma(x) \leftrightarrow \sigma(x)) \nvdash \xi.$$

Proof. The goal is to use Kripke's trick to prove the theorem, hence the proof is similar to the proof of Theorem 4.3 via Lemma 4.4.

Let $\{\mathrm{T}_i : i \in \omega\}$ and $\{X_i : i \in \omega\}$ be as in the statement of the theorem. Define a partial recursive function $f(x)$ by the stipulation that $f(m) = k$ iff $i \in \omega$ and (the Gödel number of) $\xi \in X_i$ are the least numbers such that

$$\mathrm{T}_i \vdash (R(m, m, k) \wedge \exists! z R(m, m, z)) \rightarrow \xi.$$

[7] The result also appears in Sommaruga-Rosolemos (1991), stated for extensions of $\mathrm{I}\Sigma_1$, with a proof reminiscent of the ones in Visser (1980) and Montagna (1982). Visser's result is stated for extensions of PA.

If more than one sentence of this form is provable in T_i, choose the one whose proof has the least Gödel number. Let e be an index for f, and suppose that for some $i, k \in \omega$, and $\xi \in X_i$,

$$\mathrm{T}_i + R(e, e, k) \wedge \exists! z R(e, e, z) \vdash \xi.$$

Then $f(e) = k$, so $\mathrm{I}\Delta_0 + \exp \vdash R(e, e, k) \wedge \exists! z R(e, e, z)$, but then $\mathrm{T}_i \vdash \xi$, contradicting the assumptions on T_i and X. Let $\gamma(x)$ be the Σ_n formula $\exists z(R(e, e, z) \wedge \mathrm{Sat}_{\Sigma_n}(z, x))$.

Let $\sigma(x)$ be any Σ_n formula, and suppose, for a contradiction, that for some $i \in \omega$ and $\xi \in X_i$,

$$\mathrm{T}_i + \forall x(\gamma(x) \leftrightarrow \sigma(x)) \vdash \xi.$$

By Kripke's trick,

$$\mathrm{T}_i + R(e, e, \ulcorner \sigma \urcorner) \wedge \exists! z R(e, e, z) \vdash \forall x(\gamma(x) \leftrightarrow \sigma(x))$$

so

$$\mathrm{T}_i + R(e, e, \ulcorner \sigma \urcorner) \wedge \exists! z R(e, e, z) \vdash \xi,$$

which is a contradiction. Hence for every $i \in \omega$ and every $\xi \in X_i$,

$$\mathrm{T}_i + \forall x(\gamma(x) \leftrightarrow \sigma(x)) \nvdash \xi. \qquad \square$$

4.5 Scott's lemma and Lindström's proof

One of the most successful applications of independent formulae is due to Scott (1962). To show that every countable Scott set can be realised as the standard system of some prime model of PA, he generalises Mostowski's proof to obtain:

Theorem 4.13 (Scott's lemma)**.** Suppose that T is a consistent, r.e. extension of Q. For any Σ_n formula $\phi(x)$, there is then a Σ_{n+1} formula $\xi(x)$ such that for any $g, h \in {}^{\omega}2$, if $\mathrm{T}_g = \mathrm{T} + \{\phi(k)^{g(k)} : k \in \omega\}$ is consistent, so is $\mathrm{T}_g + \{\xi(k)^{h(k)} : k \in \omega\}$.

Lindström (1984) gives a generalisation of Scott's lemma, similar to the way in which Kripke's theorem generalises Mostowski's theorem. Since Lindström (1984) is part of a departmental report series with limited circulation, the result is not widely known, and the opportunity is taken to repeat it and its proof below.

Lindström's proof uses a slight modification of Kripke's method: instead of first constructing a suitable flexible function by appealing to the recursion theorem and then feeding its output to a partial truth definition, the flexible formula is constructed directly via the diagonal lemma. To give an example of how this method can be applied, the following theorem is proved.

Theorem 4.14 (Lindström, 1984, Proposition 1)**.** Suppose that T is a consistent, r.e. extension of $\mathrm{I}\Delta_0 + \exp$, and that X is an r.e. set that is monoconsistent with T. For each $n > 0$, there is then a Σ_n formula $\gamma(x)$ such that for every $\sigma(x) \in \Sigma_n$,

$$\neg\forall x(\gamma(x) \leftrightarrow \sigma(x)) \notin X.$$

It is straightforward to construct a partial recursive function f with index e such that for every k, $f(e) \neq k \notin X$, but in order for this to imply that $\neg\forall x(\gamma(x) \leftrightarrow \sigma(x)) \notin X$ (for a suitable choice of γ), it is required that X is a deductively closed extension of $\mathrm{I}\Delta_0 + \exp$. With this additional assumption the theorem is no stronger than Theorem 4.3, and this makes it clear that Kripke's trick depends on the deductive closure of the set of sentences (i.e., the theory) used to define the function f. Here, the partial recursive function f is instead defined in terms of Gödel numbers of formulae, and the diagonalisation implemented by using the uniform diagonal lemma rather than the recursion theorem.

Proof. Let T be a consistent, r.e. extension of $\mathrm{I}\Delta_0 + \exp$, and let X be an r.e. set monoconsistent with T. Fix $n > 0$, and let a partial recursive function f be defined by the stipulation that $f(\ulcorner\eta\urcorner) = \ulcorner\sigma\urcorner$ iff

$$\sigma(x) \in \Sigma_n \text{ and } \neg\forall x(\eta(x) \leftrightarrow \sigma(x)) \in X.$$

If there are many formulae of this kind, pick the one with the least Gödel number. Let e be an index for f, and let by the uniform diagonal lemma (Fact 2.14), $\gamma(x)$ be a Σ_n formula such that

$$\mathrm{I}\Delta_0 + \exp \vdash \forall x(\gamma(x) \leftrightarrow \exists z(R(e, \ulcorner\gamma\urcorner, z) \wedge \mathrm{Sat}_{\Sigma_n}(z, x))).$$

Let $\sigma(x) \in \Sigma_n$ and suppose, for a contradiction, that

$$\neg\forall x(\gamma(x) \leftrightarrow \sigma(x)) \in X.$$

Then $f(\ulcorner\gamma\urcorner) = \ulcorner\sigma\urcorner$, so $\mathrm{I}\Delta_0 + \exp \vdash R(e, \ulcorner\gamma\urcorner, \ulcorner\sigma\urcorner) \wedge \exists! z R(e, \ulcorner\gamma\urcorner, z)$. Using Kripke's trick, it follows that $\mathrm{I}\Delta_0 + \exp \vdash \forall x(\gamma(x) \leftrightarrow \sigma(x))$. But then $\mathrm{T} + \neg\forall x(\gamma(x)) \leftrightarrow \sigma(x))$ is inconsistent, contradicting the assumption that X is monoconsistent with T. □

This method is now applied to prove Lindström's generalisation of Scott's lemma.

Theorem 4.15 (Lindström, 1984, Proposition 2)**.** Suppose that T is a consistent, r.e. extension of $\mathrm{I}\Delta_0 + \exp$. For any Σ_n formula $\phi(x)$, there is a Σ_{n+1} formula $\gamma(x)$ such that for any $g \in {}^{\omega}2$, if

$$\mathrm{T}_g = \mathrm{T} + \{\phi(k)^{g(k)} : k \in \omega\}$$

is consistent, then for every $\sigma(x) \in \Sigma_{n+1}$, $\mathrm{T}_g + \forall x(\gamma(x) \leftrightarrow \sigma(x))$ is also consistent.

Proof. Fix n, and $\phi(x) \in \Sigma_n$. Let $f(\ulcorner s\urcorner, \ulcorner\eta\urcorner) = \ulcorner\sigma\urcorner$ iff s is a binary sequence of length m such that

$$\sigma(x) \in \Sigma_n \text{ and } \mathrm{T} + \phi(0)^{(s)_0} + \cdots + \phi(m)^{(s)_m} \vdash \neg\forall x(\eta(x) \leftrightarrow \sigma(x)).$$

If there are many such σ's, choose the one with the least Gödel number, and if there are more than one such s for that particular choice of σ, choose the shortest sequence. Let $\mathrm{Seq}_\phi(x)$ be the formula

$$\forall y < l(x)((\phi(y) \to (x)_y = 0) \wedge (\neg\phi(y) \to (x)_y = 1)),$$

where $l(x)$ denotes the length of the sequence x. Hence Seq_ϕ expresses that x is a sequence agreeing with the pattern of negations applied to ϕ. If ϕ is Σ_n, then Seq_ϕ is Δ_{n+1}.

Let e be an index for f, and let $\gamma(x)$ be a Σ_{n+1} formula such that

$$\text{T} \vdash \forall x(\gamma(x) \leftrightarrow \exists s \exists z(\text{Seq}_\phi(s) \wedge R(e, s, \ulcorner\gamma\urcorner, z) \wedge \text{Sat}_{\Sigma_{n+1}}(z, x))).$$

Suppose that there is a $g \in {}^\omega 2$ such that T_g is consistent but

$$\text{T}_g + \forall x(\gamma(x) \leftrightarrow \sigma(x))$$

is inconsistent. Then there is a shortest finite initial subsequence s of g such that

$$\text{T} + \phi(0)^{(s)_0} + \cdots + \phi(m)^{(s)_m} \vdash \neg\forall x(\gamma(x) \leftrightarrow \sigma(x)).$$

By construction of f, this implies $f(\ulcorner s\urcorner, \ulcorner\gamma\urcorner) = \ulcorner\sigma\urcorner$, so

$$\text{T} \vdash R(e, \ulcorner s\urcorner, \ulcorner\gamma\urcorner, \ulcorner\sigma\urcorner) \wedge \exists! z R(e, \ulcorner s\urcorner, \ulcorner\gamma\urcorner, z).$$

With s as above, $\text{T}_g \vdash \text{Seq}_\phi(\ulcorner s\urcorner)$. Hence $\text{T}_g \vdash \forall x(\gamma(x) \leftrightarrow \sigma(x))$, contradicting the assumption that T_g is consistent. □

Theorem 4.13 follows from Theorem 4.15 by using the method from the proof of Theorem 4.5.

4.6 Chaitin's incompleteness theorem

Chaitin's incompleteness theorem is a much-discussed incompleteness result, see e.g., van Lambalgen (1989); Raatikainen (1998); Franzén (2005); Sjögren (2008). It is presented here for two reasons: first, because it can be obtained as an immediate consequence of a minor modification of Kripke's lemma. Secondly, Chaitin's theorem serves as a partial motivation for the research reported in Woodin (2011), a paper that inspired much of the work reported in Chapter 7 below.

The result can be stated in many different forms, each of which uses some kind of complexity measure, or measure of 'information content', for finite

binary strings (or the natural numbers used to represent these strings in arithmetic). A binary string can then be defined to have high complexity, or to be random, if the shortest description of an algorithm producing the string (relative to some fixed method of describing algorithms) is at least as long as the string itself.[8] This conception leads Woodin to paraphrase Chaitin's theorem as:

> [T]he property of randomness for finite binary sequences based on information content is undecidable. (Woodin, 2011, p. 119)

The argument presented below appears, in a slightly different form, in van Lambalgen (1989), where it is credited to Albert Visser and Dick de Jongh. Assuming some fixed enumeration of recursive functions, the *algorithmic complexity of* k, $C(k)$, is defined as the least index e such that $\varphi_e(0) = k$ (Raatikainen, 1998).

Theorem 4.16 (E.g., Chaitin, 1974)**.** For each sound arithmetic theory T, there is a constant c such that T does not prove any true statements of the form $C(k) > c$.

It is easy to find a constant witnessing this theorem (and more) by modifying the proof of Kripke's lemma as suggested in Section 4.4: let c be a number such that for each $k \in \omega$, the theory $\mathrm{T} + R(c,0,k) \wedge \exists! z R(c,0,z)$ is consistent. Observe that the expression $C(k) = c$ can be formalised as $R(c,0,k) \wedge \forall y{<}c \neg R(y,0,k)$. Hence the expression $C(k) > c$ can be formalised as

$$\exists x(x > c \wedge R(x,0,k) \wedge \forall y{<}x \neg R(y,0,k)).$$

Theorem 4.17. Suppose that T is a consistent, r.e. extension of Q, and let c be as above. Then no sentence of the form $C(k) > c$ is provable in T.

Proof. Pick $k \in \omega$, and suppose for a contradiction that $\mathrm{T} \vdash C(k) > c$, i.e.

$$\mathrm{T} \vdash \exists x(x > c \wedge R(x,0,k) \wedge \forall y{<}x \neg R(y,0,k)).$$

In particular, $\mathrm{T} \vdash \neg R(c,0,k)$, but by the modification of Kripke's lemma, the theory $\mathrm{T} + R(c,0,k)$ is consistent. Hence $\mathrm{T} \nvdash C(k) > c$. □

[8] See, e.g., Li and Vitányi (1993).

5 Formalisation and end-extensions

By the completeness theorem for first-order logic, Lemma 4.4 establishes the existence of models satisfying $\mathrm{T} + f(e) = k$ for any $k \in \omega$. Trivially, any such model is an end-extension of the standard model $\mathbb{N}$. The question now arises whether this is a particular feature of the standard model, or if there are other models of arithmetic that can be similarly end-extended. In this chapter it is shown that there indeed are other models that have end-extensions to models of $\mathrm{T} + f(e) = k$ for any k, and that the method for constructing such end-extensions can be generalised to prove stronger results of the same kind.

5.1 Formalisation of Kripke's theorem

By inspection of the proof of Lemma 4.4, it is clear that for every $k \in \omega$, $\mathrm{T} + f(e) = k$ is consistent iff T is consistent, and the main observation of Blanck (2016) is that this statement is formalisable in $\mathrm{I}\Delta_0 + \exp$. In the presence of Σ_1-induction, this formalisation allows for the construction of end-extensions of models of $\mathrm{Con}_{\mathrm{T}}$, thus establishing:

Theorem 5.1 (Blanck, 2016)**.** Suppose that S is a consistent, r.e. extension of $\mathrm{I}\Sigma_1$, and that T is a consistent, r.e. extension of Q. For each $n > 0$, there is then a Σ_n formula $\gamma(x)$, such that:

1. $\mathrm{S} \vdash \mathrm{Con}_{\mathrm{T}} \rightarrow \forall x \neg \gamma(x)$;
2. if $\sigma(x) \in \Sigma_n$, then every model of $\mathrm{S} + \mathrm{Con}_{\mathrm{T}}$ can be end-extended to a model of $\mathrm{T} + \forall x(\gamma(x) \leftrightarrow \sigma(x))$.

The proof of this theorem rests on the next lemma, which is the aforementioned formalisation of Kripke's lemma in $\mathrm{I}\Delta_0 + \exp$.

Lemma 5.2 (Blanck, 2016)**.** If T is a consistent, r.e. extension of Q, then there is a formula with Gödel number e, such that:

1. $\mathrm{I}\Delta_0 + \exp \vdash \forall z(\mathrm{Con}_{\mathrm{T}} \to \neg R(e, e, z))$;
2. $\mathrm{I}\Delta_0 + \exp \vdash \forall z(\mathrm{Con}_{\mathrm{T}} \leftrightarrow \mathrm{Con}_{\mathrm{T}+R(e,e,z)})$.

Proof. Let $\phi(x, z) := \mathrm{Pr}_{\mathrm{T}}(\ulcorner \neg R(\dot{x}, \dot{x}, \dot{z}) \urcorner)$, and let $e = \ulcorner \phi(x, z) \urcorner$. Then e has the desired properties. By definition of $R(x, y, z)$:

$$\mathrm{I}\Delta_0 + \exp \vdash \forall z(R(e, e, z) \leftrightarrow \mathrm{Sel}\{\mathrm{Sat}_{\Sigma_1}\}(e, e, z)) \tag{5.1}$$

which implies

$$\mathrm{I}\Delta_0 + \exp \vdash \forall z(R(e, e, z) \to \mathrm{Sat}_{\Sigma_1}(e, e, z)) \tag{5.2}$$

which in turn gives

$$\mathrm{I}\Delta_0 + \exp \vdash \forall z(R(e, e, z) \to \phi(e, z)). \tag{5.3}$$

By construction of ϕ,

$$\mathrm{I}\Delta_0 + \exp \vdash \forall z(R(e, e, z) \to \mathrm{Pr}_{\mathrm{T}}(\ulcorner \neg R(e, e, \dot{z}) \urcorner)) \tag{5.4}$$

but by provable Σ_1-completeness of Q,

$$\mathrm{I}\Delta_0 + \exp \vdash \forall z(R(e, e, z) \to \mathrm{Pr}_{\mathrm{T}}(\ulcorner R(e, e, \dot{z}) \urcorner)). \tag{5.5}$$

Together with the derivability conditions, 5.4 and 5.5 give

$$\mathrm{I}\Delta_0 + \exp \vdash \forall z(\mathrm{Con}_{\mathrm{T}} \to \neg R(e, e, z)) \tag{5.6}$$

which concludes the proof of part 1. For part 2, observe that

$$\mathrm{I}\Delta_0 + \exp \vdash \exists z \neg \mathrm{Con}_{\mathrm{T}+R(e,e,z)} \leftrightarrow \exists z \mathrm{Pr}_{\mathrm{T}}(\ulcorner \neg R(e, e, \dot{z}) \urcorner) \tag{5.7}$$

which by construction of ϕ implies

$$\mathrm{I}\Delta_0 + \exp \vdash \exists z \neg \mathrm{Con}_{\mathrm{T}+R(e,e,z)} \leftrightarrow \exists z \phi(e, z). \tag{5.8}$$

By the properties of the partial satisfaction predicate,

$$\mathrm{I}\Delta_0 + \exp \vdash \exists z \neg \mathrm{Con}_{\mathrm{T}+R(e,e,z)} \to \exists z \mathrm{Sat}_{\Sigma_1}(e, e, z) \tag{5.9}$$

and

$$\mathrm{I}\Delta_0 + \exp \vdash \exists z \neg \mathrm{Con}_{\mathrm{T}+R(e,e,z)} \to \exists z \mathrm{Sel}\{\mathrm{Sat}_{\Sigma_1}\}(e, e, z). \tag{5.10}$$

By definition of $R(x, y, z)$, this implies

$$\mathrm{I}\Delta_0 + \exp \vdash \exists z \neg \mathrm{Con}_{\mathrm{T}+R(e,e,z)} \to \exists z R(e, e, z) \tag{5.11}$$

so 5.6 and 5.11 together imply

$$\mathrm{I}\Delta_0 + \exp \vdash \forall z(\mathrm{Con}_{\mathrm{T}} \to \mathrm{Con}_{\mathrm{T}+R(e,e,z)}). \tag{5.12}$$

The implication from right to left is immediate. □

Proof of Theorem 5.1. Let S be a consistent, r.e. extension of $\mathrm{I}\Sigma_1$, and T a consistent, r.e. extension of Q. Let $n > 0$, and let e be as in Lemma 5.2; note that e depends on the choice of T. Then

$$\gamma(x) := \exists z(R(e, e, z) \wedge \mathrm{Sat}_{\Sigma_n}(z, x))$$

is as desired.

(1) Let $\mathcal{M}$ be any model of $\mathrm{S} + \mathrm{Con}_{\mathrm{T}}$. By Lemma 5.2(1),

$$\mathcal{M} \models \forall z \neg R(e, e, z).$$

By reasoning within $\mathcal{M}$, conclude that $\forall x \neg \gamma(x)$.

(2) Let $\sigma(x)$ be any Σ_n formula. By Lemma 5.2(2),

$$\mathcal{M} \models \mathrm{Con}_{\mathrm{T}+R(e,e,\ulcorner\sigma\urcorner)}.$$

Since $\mathcal{M} \models \mathrm{I}\Sigma_1$ and $\mathrm{T} + R(e, e, \ulcorner\sigma\urcorner)$ is a Σ_1-definable theory, an application of the arithmetised completeness theorem (as stated in Fact 2.38) now yields an end-extension $\mathcal{N} \models \mathrm{T} + R(e, e, \ulcorner\sigma\urcorner)$ of $\mathcal{M}$. By Kripke's trick, it follows that $\mathcal{N} \models \forall x(\gamma(x) \leftrightarrow \sigma(x))$. □

The following result is due to Hamkins (2016), who quickly obtained it after seeing a draft of Blanck and Enayat (2017). Hamkins's proof makes use of the Rosser tree introduced in Section 4.3, starting from a model of $\mathrm{PA} + \neg\mathrm{Con}_{\mathrm{PA}}$. The result stands to Theorem 5.1 as Mostowski's theorem stands to Kripke's theorem, and the proof presented here is therefore similar to the proof of Theorem 4.5.

Theorem 5.3 (Hamkins, 2016, reformulated)**.** Suppose that S is a consistent, r.e. extension of $\mathrm{I}\Sigma_1$, and that T is a consistent, r.e. extension of Q. There is a Σ_1 formula $\xi(x)$ such that for each $\mathcal{M}$-definable function f in ${}^{M}2$, every model of $\mathrm{S} + \mathrm{Con}_{\mathrm{T}}$ can be end-extended to a model of $\mathrm{T} + \{\xi(m)^{f(m)} : m \in M\}$.[9]

Proof sketch. Let $\mathcal{M} \models \mathrm{S} + \mathrm{Con}_{\mathrm{T}}$. Let f be any $\mathcal{M}$-definable function in ${}^{M}2$. Let $\xi(x)$ be a Σ_1 formula as in Theorem 5.1, and let X be the set $\{\xi(m)^{f(m)} : m \in M\}$.

By Fact 2.40, $\mathcal{M}$ can be expanded to a model $(\mathcal{M}, \mathcal{A}) \models \mathrm{WKL}_0$. By Fact 2.41, the compactness theorem is provable in WKL_0. Reasoning within $(\mathcal{M}, \mathcal{A})$, using the fact that f (and therefore X) is $\mathcal{M}$-definable, carry out the same compactness argument as in the proof of Theorem 4.5, to establish $(\mathcal{M}, \mathcal{A}) \models \mathrm{Con}_{\mathrm{T}+X}$. Since WKL_0 is a conservative extension of $\mathrm{I}\Sigma_1$, the arithmetised completeness theorem (as stated in Fact 2.38) can now be used to construct an end-extension satisfying $\mathrm{T} + X$.[10] □

5.2 Formalisation of the GRMMKV theorem

The proof of the Gödel-Rosser-Mostowski-Myhill-Kripke-Visser theorem can also be formalised within $\mathrm{I}\Sigma_1$ to show the existence of analogous end-extensions.

Theorem 5.4 (Blanck, 2016)**.** Suppose that $\{\mathrm{T}_i : i \in \omega\}$ is an r.e. family of consistent, r.e. theories such that $\mathrm{I}\Delta_0 + \exp \vdash \forall i \forall x(\mathrm{Pr}_{\mathrm{Q}}(x) \to \mathrm{Pr}_{\mathrm{T}_i}(x))$, and that $\{X_i : i \in \omega\}$ is an r.e. family of r.e. sets such that $\mathrm{I}\Delta_0 + \exp \vdash \forall i \forall \xi \in X_i(\mathrm{Con}_{\mathrm{T}_i} \to \mathrm{Con}_{\mathrm{T}_i + \neg\xi})$. For each $n > 0$, there is then a Σ_n formula $\gamma(x)$ such that for each $\sigma(x) \in \Sigma_n$, each $i \in \omega$ and each $\xi \in X_i$,

[9] Hamkins's result is stated for extensions S of PA, with $\mathrm{T} = \mathrm{S}$.

[10] This argument is, modulo minor differences, independently due to Ali Enayat.

1. if $\mathcal{M} \models \mathrm{I}\Sigma_1 + \forall i \mathrm{Con}_{\mathrm{T}_i}$, then $\mathcal{M} \models \forall x \neg\gamma(x)$;
2. every model of $\mathrm{I}\Sigma_1 + \forall i \mathrm{Con}_{\mathrm{T}_i}$ can be end-extended to a model of

$$\mathrm{T}_i + \forall x(\gamma(x) \leftrightarrow \sigma(x)) + \neg\xi.$$

The method is similar to the one used in proving Lemma 5.2.

Lemma 5.5 (Blanck, 2016)**.** Suppose that $\{\mathrm{T}_i : i \in \omega\}$ is an r.e. family of consistent, r.e. theories such that $\mathrm{I}\Delta_0 + \exp \vdash \forall x \forall i(\mathrm{Pr}_{\mathrm{Q}}(x) \to \mathrm{Pr}_{\mathrm{T}_i}(x))$, and that $\{X_i : i \in \omega\}$ is an r.e. family of r.e. sets such that $\mathrm{I}\Delta_0 + \exp \vdash \forall i \forall \xi \in X_i(\mathrm{Con}_{\mathrm{T}_i} \to \mathrm{Con}_{\mathrm{T}_i + \neg\xi})$. For each $n > 0$, there is then a Σ_n formula $\gamma(x)$ such that for each $\sigma(x) \in \Sigma_n$,

1. $\mathrm{I}\Delta_0 + \exp \vdash \forall i \forall \xi \in X_i \forall z \big(\mathrm{Con}_{\mathrm{T}_i + \neg\xi} \to \neg R(e, e, i, \xi, z)\big)$
2. $\mathrm{I}\Delta_0 + \exp \vdash \forall i \forall \xi \in X_i \forall z \big(\mathrm{Con}_{\mathrm{T}_i + \neg\xi} \to \mathrm{Con}_{\mathrm{T}_i + R(e,e,i,\xi,z) + \neg\xi}\big)$.

Proof. Let the T_i's and X_i's be as in the statement of the lemma. Let $\phi(x, u, y, z)$ be the formula $\mathrm{Pr}_{\mathrm{T}_u}(\ulcorner R(\dot{x}, \dot{x}, \dot{u}, \dot{y}, \dot{z}) \to \dot{y} \urcorner)$, and let $e = \ulcorner \phi \urcorner$. To lighten the notation, assume that every formula below is prefixed with $\forall i \forall \xi \in X_i$. By definition of R,

$$\mathrm{I}\Delta_0 + \exp \vdash \forall z \big(R(e, e, i, \xi, z) \to \phi(e, i, \xi, z)\big) \tag{5.13}$$

so by construction of ϕ,

$$\mathrm{I}\Delta_0 + \exp \vdash \forall z \big(R(e, e, i, \xi, z) \to \mathrm{Pr}_{\mathrm{T}_i}(\ulcorner R(e, e, i, \xi, \dot{z}) \to \xi \urcorner)\big). \tag{5.14}$$

By provable Σ_1-completeness of Q,

$$\mathrm{I}\Delta_0 + \exp \vdash \forall z \big(R(e, e, i, \xi, z) \to \mathrm{Pr}_{\mathrm{T}_i}(\ulcorner R(e, e, i, \xi, \dot{z}) \urcorner)\big), \tag{5.15}$$

which together with (5.14) and the derivability conditions give

$$\mathrm{I}\Delta_0 + \exp \vdash \forall z \big(R(e, e, i, \xi, z) \to \mathrm{Pr}_{\mathrm{T}_i}(\ulcorner \xi \urcorner)\big), \tag{5.16}$$

whence

$$\mathrm{I}\Delta_0 + \exp \vdash \forall z \big(\mathrm{Con}_{\mathrm{T}_i + \neg\xi} \to \neg R(e, e, i, \xi, z)\big). \tag{5.17}$$

For the latter part, recall that by definition of ϕ,

$$\mathrm{I}\Delta_0 + \exp \vdash \exists z \mathrm{Pr}_{\mathrm{T}_i}(\ulcorner R(e,e,i,\xi,\dot{z}) \to \xi \urcorner) \to \exists z \phi(e,i,\xi,z), \quad (5.18)$$

so by the properties of the partial satisfaction predicate,

$$\mathrm{I}\Delta_0 + \exp \vdash \exists z \mathrm{Pr}_{\mathrm{T}_i}(\ulcorner R(e,e,i,\xi,\dot{z}) \to \xi \urcorner) \to \exists z \mathrm{Sat}_{\Sigma_1}(e,e,i,\xi,z), \quad (5.19)$$

and by assumption on R, working in $\mathrm{I}\Delta_0 + \exp$,

$$\mathrm{I}\Delta_0 + \exp \vdash \exists z \mathrm{Pr}_{\mathrm{T}_i}(\ulcorner R(e,e,i,\xi,\dot{z}) \to \xi \urcorner) \to \exists z R(e,e,i,\xi,z). \quad (5.20)$$

But then by (5.16) and (5.20),

$$\mathrm{I}\Delta_0 + \exp \vdash \exists z \mathrm{Pr}_{\mathrm{T}_i}(\ulcorner R(e,e,i,\xi,\dot{z}) \to \xi \urcorner) \to \mathrm{Pr}_{\mathrm{T}_i}(\ulcorner \xi \urcorner), \quad (5.21)$$

so it follows that

$$\mathrm{I}\Delta_0 + \exp \vdash \forall z \big(\mathrm{Con}_{\mathrm{T}_i + \neg\xi} \to \mathrm{Con}_{\mathrm{T}_i + R(e,e,i,\xi,z) + \neg\xi} \big). \qquad \square$$

Proof of Theorem 5.4. Let the T_i's and X_i's be as in the statement of the theorem, and let $\mathcal{M} \models \mathrm{I}\Sigma_1 + \forall i \mathrm{Con}_{\mathrm{T}_i}$. Then

$$\mathcal{M} \models \mathrm{I}\Sigma_1 + \forall i \forall \xi \in X_i \mathrm{Con}_{\mathrm{T}_i + \neg\xi}$$

by assumption. Let e be as in Lemma 5.5, and let

$$\gamma(x) := \exists u \exists y \exists z \big(R(e,e,u,y,z) \wedge \mathrm{Sat}_{\Sigma_n}(z,x) \big).$$

(1) By Lemma 5.5(1), $\mathcal{M} \models \forall i \forall \xi \in X_i \forall z \neg R(e,e,i,\xi,z)$. Hence $\mathcal{M} \models \forall x \neg\gamma(x)$.

(2) Pick $j \in \omega$, $\xi \in X_j$, and $\sigma(x) \in \Sigma_n$. By Lemma 5.5(2), it follows that $\mathcal{M} \models \mathrm{Con}_{\mathrm{T}_j + R(e,e,j,\ulcorner \xi \urcorner, \ulcorner \sigma \urcorner) + \neg\xi}$, and since $\mathcal{M} \models \mathrm{I}\Sigma_1$, the arithmetised completeness theorem (Fact 2.38) provides an end-extension $\mathcal{K}$ of $\mathcal{M}$ such that

$$\mathcal{K} \models \mathrm{T}_j + R(e,e,j,\ulcorner \xi \urcorner, \ulcorner \sigma \urcorner) + \neg\xi.$$

By a reasoning similar to the one in the proof of Theorem 5.1, the model is as desired. $\square$

It may seem too strong an assumption that the model $\mathcal{M}$ satisfies the consistency of all T_i's in the r.e. family of theories. Similar reservations may be made for the other internal quantification on i and ξ. This, however, is required to ensure that $\gamma(x)$ behaves in the intended way. A more useful formulation would perhaps be along the lines that 'if $\mathcal{M}$ satisfies $\mathrm{Con}_{\mathrm{T}_i}$ for a particular choice of i, then there is a suitable end-extension'. This can be accomplished by constructing a parametrised version $\gamma(x, i, \ulcorner \xi \urcorner)$.

Theorem 5.6. Suppose that $\{\mathrm{T}_i : i \in \omega\}$ is an r.e. family of consistent, r.e. theories extending Q. Suppose further that X_i is an r.e. family of r.e. sets such that for any choice of $i \in \omega$ and $\xi \in X_i$, $\mathrm{T}_i \nvdash \xi$.

Then there is a Σ_n formula $\gamma(x, y, z)$ such that for each $\sigma(x) \in \Sigma_n$, each $i \in \omega$ and each $\xi \in X_i$,

1. if $\mathcal{M} \models \mathrm{I}\Sigma_1 + \mathrm{Con}_{\mathrm{T}_i + \neg\xi}$, then $\mathcal{M} \models \forall x \neg\gamma(x, i, \ulcorner \xi \urcorner)$;
2. every model of $\mathrm{I}\Sigma_1 + \mathrm{Con}_{\mathrm{T}_i + \neg\xi}$ can be end-extended to a model of $\mathrm{T}_i + \forall x(\gamma(x, i, \ulcorner \xi \urcorner) \leftrightarrow \sigma(x)) + \neg\xi$.

The proof can be obtained by an easy modification of the proofs of Lemma 5.5 and Theorem 5.4.

5.3 Hierarchical generalisations

As in the case with Kripke's theorem, there are also hierarchical generalisations of the formalisation results above. Here, a method is sketched for converting the proofs of Theorem 5.1 and Lemma 5.2 to yield one such generalisation.

Recall that $\mathrm{Pr}_{\mathrm{T},\Sigma_{n+1}}(x)$ is the Σ_{n+1} formula

$$\exists z(\Sigma_{n+1}(z) \wedge \mathrm{Tr}_{\Sigma_{n+1}}(z) \wedge \mathrm{Pr}_{\mathrm{T}}(\ulcorner \dot{z} \rightarrow \dot{x} \urcorner)),$$

and that $\mathrm{Con}_{\mathrm{T},\Sigma_{n+1}}$ is the Π_{n+1} formula $\neg\mathrm{Pr}_{\mathrm{T},\Sigma_{n+1}}(\bot)$. Moreover, recall that the Σ_{n+1} formula $R^n(x, y, z)$ is defined as $\mathrm{Sel}\{\mathrm{Sat}_{\Sigma_{n+1}}\}(x, y, z)$. Hence z is functionally dependent on x and y in $\mathrm{I}\Sigma_{n+1}$.

Let $\phi(x, z)$ be the Σ_{n+1} formula $\mathrm{Pr}_{\mathrm{T},\Sigma_{n+1}}(\ulcorner \neg R^n(\dot{x}, \dot{x}, \dot{z}) \urcorner)$ and let $e = \ulcorner \phi \urcorner$. Let $\gamma(x)$ be the Σ_{n+1} formula $\exists z(R^n(e, e, z) \wedge \mathrm{Sat}_{\Sigma_{n+1}}(z, x))$.

Examine the proofs of Theorem 5.1 and Lemma 5.2 and replace all occurrences of Pr_{T}, $\mathrm{Con}_{\mathrm{T}}$, R and Sat_{Σ_1} with $\mathrm{Pr}_{\mathrm{T},\Sigma_{n+1}}$, $\mathrm{Con}_{\mathrm{T},\Sigma_{n+1}}$, R^n and $\mathrm{Sat}_{\Sigma_{n+1}}$, respectively. The resulting proofs yield:

Lemma 5.7. Let $n > 0$. If T is a Σ_{n+1}-definable, consistent extension of Q, then there is a Σ_{n+1} formula with Gödel number e, such that:

1. $\mathrm{I}\Sigma_n \vdash \forall z(\mathrm{Con}_{\mathrm{T},\Sigma_{n+1}} \rightarrow \neg R^n(e,e,z))$;
2. $\mathrm{I}\Sigma_n \vdash \forall z(\mathrm{Con}_{\mathrm{T},\Sigma_{n+1}} \leftrightarrow \mathrm{Con}_{\mathrm{T},\Sigma_{n+1}+R^n(e,e,z)})$.

Theorem 5.8. Let $n > 0$. Suppose that S is a consistent, r.e. extension of $\mathrm{I}\Sigma_{n+1}$, and that T is a Σ_{n+1}-definable, consistent extension of Q. There is a Σ_{n+1} formula $\gamma(x)$ such that:

1. $\mathrm{S} \vdash \mathrm{Con}_{\mathrm{T},\Sigma_{n+1}} \rightarrow \forall x \neg\gamma(x)$;
2. if $\sigma(x) \in \Sigma_{n+1}$, then every model $\mathcal{M}$ of $\mathrm{S} + \mathrm{Con}_{\mathrm{T},\Sigma_{n+1}}$ has a Σ_n-elementary extension to a model of $\mathrm{T} + \forall x(\gamma(x) \leftrightarrow \sigma(x))$.

Remark 5.9. That $\mathcal{K}$ is a Σ_n-elementary extension of $\mathcal{M}$ follows from the fact that if $\mathcal{K} \models \mathrm{Th}_{\Sigma_{n+1}}(\mathcal{M})$ (as is the case above), then it must also be the case that $\mathcal{K} \models \mathrm{Th}_{\Sigma_n}(\mathcal{M}) \cup \mathrm{Th}_{\Pi_n}(\mathcal{M})$. Hence there is no room for changing the truth-value of any Σ_n formula when passing to the extension.

$\mathrm{I}\Sigma_{n+1}$ is used to ensure the existence of the desired extension, while $\mathrm{I}\Sigma_n$ suffices to show that z in the Σ_{n+1} formula $R^n(x,y,z)$ is functionally dependent on x and y, hence unique.

6 Characterisations of partial conservativity

The previous chapter establishes the existence of certain kinds of end-extensions related to the concepts of flexibility and independence. As is known since the late 1970s, end-extensions play an important role in characterisations of interpretability and partial conservativity, both of which are useful tools in measuring and comparing the strength of theories. An illuminating passage is found in the introductory chapter of Hájek and Pudlák (1993).

> [W]hat more can we say about systems of arithmetic than that they are all incomplete? There are at least four directions in which the answer may be looked for:
>
> (1) For each formula ϕ unprovable and non-refutable in an arithmetic T we may ask, how *conservative* it is over T, i.e. for which formulas ψ the provability of ψ in $(T+\phi)$ implies the provability of ψ in T.
>
> (2) We may further ask if $(T+\phi)$ is *interpretable* in T, i.e. whether the notions of T may be redefined in T in such a way that for the new notions all axioms of $(T+\phi)$ are provable in T.
>
> (3) Given T we may look for various *natural* sentences true but unprovable in T (for example, various combinatorial principles).
>
> (4) Moreover, we may investigate *models* of T and look at how they visualize our syntactic notions and features.
>
> (Hájek and Pudlák, 1993, p. 3)

This chapter elaborates on the interesting relationship between these four directions, especially concerning (1), (2), and (4). In particular, characterisations of partial conservativity are given over *non-r.e.* theories, *weak* (especially non-reflexive) theories of arithmetic, and theories formulated in

an *extended language*. Each of these relaxations gives rise to their own difficulties. Some of the characterisations are necessary prerequisites for the results in Chapter 7.

6.1 The Orey-Hájek characterisation and its extensions

One of the cornerstones in the early study of the metamathematics of formal arithmetic is the Orey-Hájek characterisation of interpretability.

Theorem 6.1. Let T and S be consistent, r.e. extensions of PA in the same language. The following are equivalent:

1. S is interpretable in T;
2. $S|k$ is interpretable in T for all $k \in \omega$;
3. $T \vdash \mathrm{Con}_{S|k}$ for all $k \in \omega$.

The equivalence of (1) and (2) is due to Orey (1961): it is also known as Orey's compactness theorem. The equivalence of (1) and (3) is,

> implicit in Feferman (1960), all but explicit in Orey (1961), and fully explicit in Hájek (1971). (Lindström, 2003, p. 115)

By the independent work of Guaspari (1979) and Lindström (1979) it is possible to include a fourth equivalent condition:

4. S is Π_1-conservative over T.

It is well known that the success of the Orey-Hájek characterisation, as presented above, depends on certain features of the theories S and T. Of particular interest are reflection properties, sequentiality and whether or not the theories prove the totality of the exponentiation function. For finitely axiomatised theories, interpretability and Π_1-conservativity no longer coincide (Pudlák, 1985; Hájek, 1987; Visser, 1990; Shavrukov, 1997; Joosten, 2004).

The following theorem is a general characterisation for r.e. theories extending PA, and it is used repeatedly throughout Chapter 7, referred to

as the OHGL characterisation. The equivalence of (1) and (4) is due to Guaspari (1979), who writes that '[This method] for constructing end-extensions seems to be well known'. The condition (3) is rarely (if ever) included in these characterisations, but is almost definitionally equivalent to condition (4). It is also a convenient condition to work with.

Theorem 6.2 (The Orey-Hájek-Guaspari-Lindström characterisation)**.** Suppose that T and S are consistent, r.e. extensions of PA in the same language. Then the following are equivalent:

1. every model of T can be end-extended to a model of S;
2. every countable model of T can be end-extended a model of S;
3. for every model $\mathcal{M}$ of T, the theory $\mathrm{Th}_{\Sigma_1}(\mathcal{M}) + \mathrm{S}$ is consistent;
4. S is Π_1-conservative over T;
5. $\mathrm{T} \vdash \mathrm{Con}_{\mathrm{S}|k}$ for all $k \in \omega$;
6. S is interpretable in T.

By putting together a number of observations of Guaspari (1979), it is possible to state a hierarchical version of the above. To formulate this version, Guaspari's concept of provably Γ-faithful interpretations is convenient. Let t be an interpretation of S in T: then t is *provably* Γ*-faithful* if, for every $\phi \in \Gamma$, $\mathrm{T} \vdash t(\ulcorner \phi \urcorner) \to \phi$.

Theorem 6.3. Suppose that T and S are consistent, r.e. extensions of PA in the same language. Then the following are equivalent:

1. every model of T has a Σ_n-elementary extension to a model of S;
2. every countable model of T has a Σ_n-elementary extension to a model of S;
3. for every model $\mathcal{M}$ of T, the theory $\mathrm{Th}_{\Sigma_{n+1}}(\mathcal{M}) + \mathrm{S}$ is consistent;
4. S is Π_{n+1}-conservative over T;
5. $\mathrm{T} \vdash \mathrm{Con}_{(\mathrm{S},\Sigma_{n+1})|k}$ for all $k \in \omega$;
6. there is a provably Π_{n+1}-faithful interpretation of S in T.

6.2 A characterisation of partial conservativity over $I\Sigma_1$

As suggested in the previous section, the OHGL characterisation (and even the Orey-Hájek characterisation) breaks down when passing to weaker theories of arithmetic, e.g., extensions of $I\Sigma_1$. One reason is that whenever T is finitely axiomatisable, it is impossible to have $T \vdash \mathrm{Con}_{T|k}$ for all $k \in \omega$. Another reason is that, as noted above, interpretability and Π_1-conservativity no longer coincide. It is a major open problem whether every model of $I\Sigma_1$ has a proper end-extension to a model of $I\Sigma_1$. Even so, Fact 2.35 assures that every *countable* model of $I\Sigma_1$ is isomorphic to a proper initial segment of itself. This makes it possible to salvage parts of the characterisation of partial conservativity for extensions of $I\Sigma_1$.

Theorem 6.4 (Blanck and Enayat, 2017)**.** Suppose that T and S are consistent, r.e. extensions of $I\Sigma_1$ in the same language. Then the following are equivalent:

1. every countable model of T can be end-extended to a model of S;
2. for every model $\mathcal{M}$ of T, the theory $S + \mathrm{Th}_{\Sigma_1}(\mathcal{M})$ is consistent;
3. S is Π_1-conservative over T;
4. for each $n \in \omega$, $T \vdash \mathrm{Con}^n_{S,\Sigma_1}$.

Proof. (1) $\Rightarrow$ (2). Prove the contrapositive by supposing that (2) fails. There is then a countable model $\mathcal{M} \models T$ such that $\mathrm{Th}_{\Sigma_1}(\mathcal{M}) + S$ is inconsistent. Hence there is a $\sigma \in \mathrm{Th}_{\Sigma_1}(\mathcal{M})$ such that $S + \sigma \vdash \bot$. Since Σ_1 sentences are preserved when passing to an end-extension, this shows that every end-extension of $\mathcal{M}$ must fail to satisfy S, which makes it evident that (1) fails.

(2) $\Rightarrow$ (3). Prove the contrapositive by supposing that (3) fails. There is then a $\pi \in \Pi_1$ such that $S \vdash \pi$ and $T + \neg\pi$ is consistent. By the completeness theorem, there is some model $\mathcal{M} \models T + \neg\pi$. But since $\neg\pi \in \mathrm{Th}_{\Sigma_1}(\mathcal{M})$ and $S \vdash \pi$, (2) must fail.

(3) $\Rightarrow$ (4). Suppose (3), and let $n \in \omega$. By small reflection (Fact 2.25), $S \vdash \mathrm{Con}^n_{S,\Sigma_1}$. $\mathrm{Con}^n_{S,\Sigma_1}$ is a Π_1 sentence, and S is Π_1-conservative over T, so $T \vdash \mathrm{Con}^n_{S,\Sigma_1}$.

(4) $\Rightarrow$ (1). Suppose (4), and let $\mathcal{M}$ be a countable model of T. Assume, without loss of generality, that $\mathcal{M}$ is non-standard. By (4),

$$\forall n \in \omega \mathcal{M} \models \mathrm{Con}^n_{\mathrm{S},\Sigma_1},$$

and since $\mathcal{M} \models \mathrm{I}\Sigma_1$, $\mathcal{M}$ satisfies the overspill principle for Π_1 formulae, whence

$$\mathcal{M} \models \mathrm{Con}^m_{\mathrm{S},\Sigma_1}$$

for some non-standard $m \in M$. By McAloon's theorem (Fact 2.39), there is a submodel $\mathcal{M}_0 \models \mathrm{PA}$ that forms a non-standard initial segment of $\mathcal{M}$, all of whose elements are below m.

There is some $a \in M$ that codes the set $\{n \in \omega : \mathcal{M} \models \mathrm{Tr}_{\Sigma_1}(n)\} = \mathrm{Th}_{\Sigma_1}(\mathcal{M})$, i.e. the standard part of $\mathrm{True}_{\Sigma_1}(\mathcal{M})$, and this a can be chosen to be below m. In other words, $\mathcal{M} \models \forall x(x \varepsilon a \rightarrow \mathrm{Tr}_{\Sigma_1}(x))$, which together with $a < m$ ensures that

$$\mathcal{M}_0 \models \mathrm{Con}_{\mathrm{S}+\{n : n \varepsilon a\}}.$$

Since $\mathcal{M}_0 \models \mathrm{PA}$, the arithmetised completeness theorem (Fact 2.38) guarantees the existence of an end-extension $\mathcal{N}$ of $\mathcal{M}_0$ satisfying $\mathrm{S}+\{n : n \varepsilon a\}$, and therefore $\mathrm{S}+\mathrm{Th}_{\Sigma_1}(\mathcal{M})$. Since $\mathcal{M}_0 \subseteq_e \mathcal{M}$, and $\mathcal{M}_0 \subseteq_e \mathcal{N}$, it follows that $\mathrm{SSy}(\mathcal{M}) = \mathrm{SSy}(\mathcal{M}_0) = \mathrm{SSy}(\mathcal{N})$. This, together with the fact that $\mathcal{N} \models \mathrm{Th}_{\Sigma_1}(\mathcal{M})$ allows Fact 2.35 to embed $\mathcal{M}$ as an initial segment of $\mathcal{N}$. □

Remark 6.5. The equivalence of (3) and (4) seems to have been known to experts for some time: for example, it figures in an unpublished note from the early 1990s, due to Albert Visser. The same note contains a proof of what is essentially Fact 2.25. Similar results also appear in Joosten (2004, Chapter 2). The fact that (2) implies (1) follows, for extensions of PA, from Theorem 2 of Woodin (2011). That (2) and (3) are related is easy to see, and as the proof above suggests, the real difficulty lies in establishing that (4) implies (1).

The proof of the theorem above lends itself to prove a hierarchical generalisation. The equivalence of (1) and (3) for extensions of PA follows from Theorem 6.5(i) of Guaspari (1979).

Theorem 6.6 (Blanck and Enayat, 2017)**.** Let T be a consistent, r.e. extension of $\mathrm{I}\Sigma_{n+1}$ in the same language. Then the following are equivalent:

1. every countable model of T has a Σ_n-elementary extension to a model of S;
2. for every model $\mathcal{M}$ of T, the theory $\mathrm{S} + \mathrm{Th}_{\Sigma_{n+1}}(\mathcal{M})$ is consistent.
3. S is Π_{n+1}-conservative over T;
4. for each $k \in \omega$, $\mathrm{T} \vdash \mathrm{Con}^{k}_{\mathrm{S},\Sigma_{n+1}}$.

6.3 Language extensions

This section concerns rather specific improvements of Theorems 6.2 and 6.4 that are needed in the proofs of some of the results in Chapter 7. The first such improvement is essentially due to Guaspari: the equivalence of the new condition (3) with the others is immediate.

Theorem 6.7 (Guaspari, 1979, Theorem 6.5(i))**.** Let T be a consistent, r.e. extension of PA formulated in a finite language $\mathscr{L} \supseteq \mathscr{L}_A$. Then the following are equivalent for an $\mathscr{L}$-sentence ϕ:

1. every model of T can be end-extended to a model of $\mathrm{T} + \phi$;
2. every countable model of T can be end-extended to a model of $\mathrm{T}+\phi$;
3. for each model $\mathcal{M}$ of T, the theory $\mathrm{T} + \mathrm{Th}_{\Sigma_1(\mathscr{L})}(\mathcal{M}) + \phi$ is consistent;
4. $\mathrm{T} + \phi$ is $\Pi_1(\mathscr{L})$-conservative over T.

The next theorem is a new generalisation of Theorem 6.4 to extensions of $\mathrm{I}\Sigma_1$ formulated in an extended language $\mathscr{L}(c) = \mathscr{L}_A \cup \{c\}$, where c is a single new individual constant. Recall that if T is a theory formulated in $\mathscr{L}(c)$ and $\mathrm{T} \vdash \mathrm{I}\Sigma_n$, then $\mathrm{T} \vdash \mathrm{I}\Sigma_n(c)$. This generalisation, Theorem 6.8, is a refinement of Theorem 2 of Woodin (2011), which establishes the implication (2) $\Rightarrow$ (1) for theories extending PA. It is also exactly what is used to prove the main theorem of Section 7.1.

Theorem 6.8 (Blanck and Enayat, 2017)**.** Let T be a consistent, r.e. extension of $\mathrm{I}\Sigma_1$ formulated in a language $\mathscr{L}(c) = \mathscr{L}_A \cup \{c\}$. The following are equivalent for an $\mathscr{L}(c)$-sentence $\phi(c)$:

1. every countable model of T can be end-extended to a model of $\mathrm{T} + \phi(c)$;
2. for every model $(\mathcal{M}, s)$ of T, the theory $\mathrm{T} + \phi(c) + \mathrm{Th}_{\Sigma_1(c)}(\mathcal{M}, s)$ is consistent;
3. $\mathrm{T} + \phi(c)$ is $\Pi_1(c)$-conservative over T;
4. for each $n \in \omega$, $\mathrm{T} \vdash \mathrm{Con}^n_{\mathrm{T},\Sigma_1(c)+\phi(c)}$.

Proof. The proof is similar to the proof of Theorem 6.4, but some extra bookkeeping is required because of the expanded language.

(1) $\Rightarrow$ (2). Suppose that (2) fails. By the completeness theorem, and the Löwenheim-Skolem theorem, there is a countable model $(\mathcal{M}, s) \models \mathrm{T}$ such that $\mathrm{T} + \phi(c) + \mathrm{Th}_{\Sigma_1(c)}(\mathcal{M}, s)$ is inconsistent; it then follows that $\mathrm{T} + \mathrm{Th}_{\Sigma_1(c)} \vdash \neg\phi(c)$. Suppose that $(\mathcal{N}, t)$ is an $\mathscr{L}_A(c)$-structure end-extending $(\mathcal{M}, s)$, and that $(\mathcal{N}, t) \models \phi(c)$. It follows that $s = t$, and since $\Sigma_1(c)$ sentences are preserved when passing to $(\mathcal{N}, t)$, this shows that $(\mathcal{N}, t) \models \phi(c) \wedge \neg\phi(c)$, which in turn makes it evident that (1) fails.

(2) $\Rightarrow$ (3). Suppose that (3) fails. There is then a $\Pi_1(c)$ sentence π such that $\mathrm{T} + \phi(c) \vdash \pi$ and $\mathrm{T} + \neg\pi$ is consistent. Let $(\mathcal{M}, s)$ be a model of $\mathrm{T} + \neg\pi$. Then $\mathrm{T} + \phi(c) + \mathrm{Th}_{\Sigma_1(c)}(\mathcal{M}, s)$ is inconsistent, since $\mathrm{T} + \neg\pi \vdash \neg\phi(c)$ and $\neg\pi \in \Sigma_1(c)$.

(3) $\Rightarrow$ (4). Suppose that (3) holds, and pick any $n \in \omega$. Then by small reflection, $\mathrm{T} + \phi(c) \vdash \mathrm{Con}^n_{\mathrm{T},\Sigma_1(c)+\phi(c)}$, and since $\mathrm{Con}^n_{\mathrm{T},\Sigma_1(c)+\phi(c)}$ is a $\Pi_1(c)$ sentence, the conclusion follows from the $\Pi_1(c)$-conservativity of $\phi(c)$ over T.

(4) $\Rightarrow$ (1). Suppose that (4) holds, and let $(\mathcal{M}, s)$ be any countable model of T. Since $(\mathcal{M}, s) \models \mathrm{I}\Sigma_1$, it follows that $(\mathcal{M}, s) \models \mathrm{I}\Sigma_1(c)$. Therefore the model satisfies $\Pi_1(c)$-overspill, so it follows that

$$(\mathcal{M}, s) \models \mathrm{Con}^m_{\mathrm{T},\Sigma_1(c)+\phi(c)}$$

for some non-standard $m \in M$.

By reasoning as in the proof of the corresponding clause of Theorem 6.4, there is an initial submodel $\mathcal{M}_0 \models \mathrm{PA}$ of $(\mathcal{M}, s)$, all of whose elements are below m, and an $a < m$ that codes $\mathrm{Th}_{\Sigma_1(c)}(\mathcal{M}, s)$.

Hence $\mathcal{M}_0 \models \mathrm{PA} + \mathrm{Con}_{\mathrm{S}+\{n:n\varepsilon a\}}$, and Fact 2.38 gives rise to an end-extension $\mathcal{N}$ of $\mathcal{M}_0$ such that $(\mathcal{N}, t)$ satisfies $\mathrm{S} + \{n : n\varepsilon a\}$. Since $\mathrm{SSy}(\mathcal{M}) = \mathrm{SSy}(\mathcal{M}_0) = \mathrm{SSy}(\mathcal{N})$, and $\mathrm{Th}_{\Sigma_1(c)}(\mathcal{M}, s)$ is contained in the set coded by a, Fact 2.35 assures the existence of an embedding f of $(\mathcal{M}, s)$ onto an initial segment of $(\mathcal{N}, t)$, with $f(s) = t$. □

This theorem can also be generalised in the spirit of Theorem 6.6, and again the equivalence of (1) and (3) for extensions of PA is due to Guaspari (1979). His proof yields the additional information that if T extends PA the assumption that $\mathcal{M}$ is countable can be removed, as in Theorem 6.7.

Theorem 6.9 (Blanck and Enayat, 2017)**.** Let T be a consistent, r.e. extension of $\mathrm{I}\Sigma_{n+1}$ formulated in the language $\mathscr{L}(c) = \mathscr{L}_A \cup \{c\}$. The following are equivalent for an $\mathscr{L}_A(c)$-sentence $\phi(c)$:

1. every countable model of T has a $\Sigma_n(c)$-elementary extension to a model of $\mathrm{T} + \phi(c)$;
2. for every model $(\mathcal{M}, s)$ of T, the theory $\mathrm{T}+\phi(c)+\mathrm{Th}_{\Sigma_{n+1}(c)}(\mathcal{M}, s)$ is consistent;
3. $\mathrm{T} + \phi(c)$ is $\Pi_{n+1}(c)$-conservative over T;
4. for each $k \in \omega$, $\mathrm{T} \vdash \mathrm{Con}^k_{\mathrm{T},\Sigma_{n+1}(c)+\phi(c)}$.

6.4 Theories that are not recursively enumerable

On the one hand, many theories featuring in the metatheory of first-order arithmetic are recursively enumerable (and therefore axiomatisable by a primitive recursive set of axioms, by Craig's trick). It is sometimes taken as a minimal requirement for something to be a theory: that it must be possible to check whether or not a statement is an axiom of the theory. On the other hand, in the study of models of arithmetic, non-r.e. sets of sentences feature regularly, e.g., $\mathrm{Th}_{\Pi_1}(\mathcal{M})$ for some model of arithmetic $\mathcal{M}$.

As noted by Guaspari (1979), the assumption that a theory is r.e. can be substituted by the assumption that the set of theorems of the theory is coded in the model. The theorem below provides a characterisation of partial conservativity for non-r.e. theories along the lines of the earlier results. Although this theorem is essentially well known (it can be extracted from Guaspari (1979), except for condition (2)), it is included here for sake of completeness, and because it is used at one point in Section 7.3.

Definition 6.10. T proves the local consistency of S if, for each finite subset F of S, $\mathrm{T} \vdash \mathrm{Con}_F$.[11]

Theorem 6.11. Suppose that T and S are consistent extensions of PA in the same language. Then the following are equivalent:

1. for each $\mathcal{M} \models \mathrm{T}$, if $\mathrm{S} \in \mathrm{SSy}(\mathcal{M})$, then $\mathcal{M}$ can be end-extended to a model of S;
2. for every model $\mathcal{M}$ of T, the theory $\mathrm{Th}_{\Sigma_1}(\mathcal{M}) + \mathrm{S}$ is consistent;
3. S is Π_1-conservative over T;
4. T proves the local consistency of S.

Proof. (1) $\Rightarrow$ (2). Suppose that (2) fails, i.e. that there is a model $\mathcal{M} \models \mathrm{T}$ such that $\mathrm{Th}_{\Sigma_1}(\mathcal{M}) + \mathrm{S}$ is inconsistent. Hence there is a $\sigma \in \mathrm{Th}_{\Sigma_1}(\mathcal{M})$ such that $\mathrm{S} + \sigma \vdash \bot$. Since Σ_1 sentences are preserved when passing to an end-extension, this shows that every end-extension of $\mathcal{M}$ must fail to satisfy S, making it evident that (1) fails.

(2) $\Rightarrow$ (3). Suppose (3) fails, i.e. that there is a $\pi \in \Pi_1$ such that $\mathrm{S} \vdash \pi$ and $\mathrm{T} + \neg\pi$ is consistent. By the completeness theorem, there is a model $\mathcal{M} \models \mathrm{T} + \neg\pi$. Since $\neg\pi \in \mathrm{Th}_{\Sigma_1}(\mathcal{M})$ and $\mathrm{S} \vdash \pi$, it follows that (2) fails.

(3) $\Rightarrow$ (4). Suppose that S is Π_1-conservative over T, and let F be any finite subtheory of S. Since S is essentially reflexive, $\mathrm{S} \vdash \mathrm{Con}_F$. But $\mathrm{Con}_F \in \Pi_1$ since F is finite, so $\mathrm{T} \vdash \mathrm{Con}_F$.

(4) $\Rightarrow$ (1). Suppose that T proves the local consistency of S, and let $\mathcal{M}$ be a model of T such that $\mathrm{S} \in \mathrm{SSy}(\mathcal{M})$.

[11] In Guaspari (1979) the terminology is 'S is strongly consistent with T'.

Assume that s is a code for S in $\mathcal{M}$, and let $\phi(x, y)$ be a Π_1 formula expressing 'the first x elements of y are consistent'. Then $\mathcal{M} \models \phi(n, s)$ for all $n \in \omega$. For suppose $\mathcal{M} \models \neg\phi(n, s)$ for some $n \in \omega$. Then there is a finite subtheory F of S such that $\mathcal{M} \models \neg\mathrm{Con}_F$. But T proves the local consistency of S, and $\mathcal{M} \models$ T, a contradiction.

By the overspill principle, there is an $m \in M$ such that $\mathcal{M} \models \phi(m, s)$. The existence of the desired end-extension now follows from Fact 2.38. $\square$

7 Uniformly flexible formulae and Solovay functions

Chapter 5 establishes the existence of a formula $\gamma(x) \in \Sigma_n$, such that for every $\sigma(x) \in \Sigma_n$, every model $\mathcal{M}$ of $\mathrm{T} + \mathrm{Con}_\mathrm{T}$ can be end-extended to a model of $\mathrm{T} + \forall x(\gamma(x) \leftrightarrow \sigma(x))$. The question now arises of whether the assumption that $\mathcal{M}$ satisfies Con_T can be removed from this, and other, results of Chapter 5. To see the relevance of this question, recall Gödel's second incompleteness theorem.

Theorem 7.1 (Gödel, 1931)**.** Let T be a consistent, r.e. extension of $\mathrm{I}\Delta_0 + \mathrm{exp}$. Then $\mathrm{T} + \neg\mathrm{Con}_\mathrm{T}$ is consistent.

As argued in the introduction to Chapter 4, the existence of independent formulae (and therefore also of flexible formulae) strengthens the first incompleteness theorem. Another way to improve the incompleteness theorem is to claim that $\mathrm{T} + \neg\mathrm{Con}_\mathrm{T}$ is not only consistent, but also interpretable in T. That this is indeed the case is the essence of Feferman's theorem on the interpretability of inconsistency.

Theorem 7.2 (Feferman, 1960)**.** Let T be a consistent, r.e. extension of $\mathrm{I}\Delta_0 + \mathrm{exp}$. Then $\mathrm{T} + \neg\mathrm{Con}_\mathrm{T}$ is interpretable in T.

In light of the OHGL characterisation (Theorem 6.2), this is equivalent to every model of T having an end-extension to a model of $\mathrm{T} + \neg\mathrm{Con}_\mathrm{T}$. Turning attention back to flexibility, it would be desirable to see a similar improvement of Kripke's theorem 4.3. Simply put, the question is:

Question 7.3. Is there, for any $n > 0$, a Σ_n formula $\gamma(x)$ such that for every $\sigma(x) \in \Sigma_n$, $\mathrm{T} + \forall x(\gamma(x) \leftrightarrow \sigma(x))$ is interpretable in T?

Indeed, if the assumption that $\mathcal{M} \models \mathrm{Con}_\mathrm{T}$ could be removed from the formalisation of Kripke's theorem, the question would have a positive

answer. The existence of such a *uniformly flexible formula* $\gamma(x)$ would then yield not only an improvement of the first incompleteness theorem, but also of the second, in the spirit of Feferman.

There are reasons to treat a number of special cases of this general problem separately. A somewhat recent result due to Woodin (2011) gives an important uniform flexibility result for formulae with bounded extensions. This result in conjunction with the formalisation of Kripke's theorem in Chapter 5 is what suggests the general question above. Woodin employs a *Solovay function* for his proof – a versatile technique introduced by Solovay (1976). Section 7.1 gives an introduction to the method by reproving Woodin's theorem, and also by giving some important generalisations.

In Section 7.3, an affirmative answer to the question is given for $n > 1$, while the trickier special case $n = 1$ is treated in Section 7.4. There, only partial results are given. Finally, hierarchical generalisations are discussed in Section 7.5.

7.1 Woodin's theorem and its extensions

In Woodin (2011), the following theorem is established.

Theorem 7.4. Suppose that T is a consistent, r.e. extension of PA. There is an r.e. set W_e such that:

1. $\mathrm{PA} \vdash \mathrm{Con_T} \rightarrow W_e = \emptyset$;
2. for each countable model $\mathcal{M} \models \mathrm{T}$, if s is an $\mathcal{M}$-finite set such that $\mathcal{M} \models W_e \subseteq s$, then there is an end-extension of $\mathcal{M}$ satisfying $\mathrm{T} + W_e = s$.[12]

In Woodin's paper, the purpose of the theorem is to drive a philosophical argument about the distinction between determinism and nondeterminism, an aspect not discussed in this thesis. Rather, the reason for including

[12]Strictly speaking, the expression $W_e = \emptyset$ is not a sentence in $\mathscr{L}_A$. Using the coding machinery outlined in Chapter 2, it is possible to take this notation as shorthand for the Π_1 sentence $\forall y \forall z \neg R(e, y, z)$, and the informal reading of this is 'the Turing machine with index e never produces any output'. Similarly, expressions like $W_e = s$ can be understood as shorthand for $\forall y \forall z (R(e, y, z) \leftrightarrow z \varepsilon c)$, where c is a new constant symbol. These considerations also apply to similar expressions in the sequel.

Woodin's result here is its relation to the question proposed in the introduction to this chapter, and its relation to flexible and independent formulae. The following paragraphs expound on these connections.

First, recall that there is a straightforward connection between r.e. sets W_n, and Σ_1 formulae, in that every r.e. set can be numerated (in an extension of Q) by a Σ_1 formula. Conversely, every set numerated in an r.e. extension of Q is r.e. Given certain assumptions on the enumeration of the r.e. sets and their representation within arithmetic, it is possible to arrange so that the formula $\phi(x)$ with Gödel number $\ulcorner\phi\urcorner$ numerates the r.e. set $W_{\ulcorner\phi\urcorner}$. Such a correspondence can be arranged to be verifiable in $\mathrm{I}\Delta_0 + \exp$, and W_e can therefore be viewed as a Σ_1 formula $\gamma(x)$.[13]

Secondly, an $\mathcal{M}$-finite set is one that is bounded within $\mathcal{M}$, i.e. a set X such that for all $x \in X$, $\mathcal{M} \models x \leq m$ for some $m \in M$. In particular, if $\mathcal{M} \models \mathrm{I}\Sigma_n$, then every set of the form $\{n \in \omega : \mathcal{M} \models \sigma(n)\}$ for $\sigma(x) \in \Sigma_n$ is $\mathcal{M}$-finite.

Moreover, the assumption that $\mathcal{M} \models W_e \subseteq s$ is necessary since Σ_1 sentences persist when passing to an end-extension: if $\mathcal{M} \models \sigma$ for $\sigma \in \Sigma_1$, and $\mathcal{M} \subseteq_e \mathcal{N}$, then $\mathcal{N} \models \sigma$. Since $n \in W_e$ is expressible by a Σ_1 formula, W_e can never shrink when passing to an end-extension, only grow. For instance, suppose that $\mathcal{M} \models W_e = \{k\}$ and $\mathcal{M} \models s = \{n\}$, with $k \neq n$. Then there can be no end-extension of $\mathcal{M}$ in which $W_e = s$, since then W_e would have to have 'lost' the element k in order to have the same extension as s, and this is not possible.

Taken together these observations suggests that Woodin's theorem can be regarded as a generalisation of Mostowski's theorem 4.5, rather than as a generalisation of Kripke's theorem 4.3. This is because Theorem 7.4 allows for the construction of an end-extension $\mathcal{N}$ in which W_e has a previously prescribed extension (i.e. one chosen from the finite sets of $\mathcal{M}$), rather than an extension which is identical to that of some Σ_1 formula $\sigma(x)$, whatever extension $\sigma(x)$ may happen to have in $\mathcal{N}$.

Following the publication of Woodin (2011), Ali Enayat and Volodya Shavrukov (in unpublished manuscripts) improved Woodin's theorem by removing the countability assumption on the base model $\mathcal{M}$.

[13] Cf. Fact 2.52 and the subsequent discussion.

Theorem 7.5 (Woodin/Enayat and Shavrukov)**.** Suppose that T is a consistent, r.e. extension of PA. There is an r.e. set W_e, such that:

1. $\mathrm{PA} \vdash \mathrm{Con}_{\mathrm{T}} \rightarrow W_e = \emptyset$;
2. for each model $\mathcal{M} \models \mathrm{T}$, if s is an $\mathcal{M}$-finite set such that $\mathcal{M} \models W_e \subseteq s$, then there is an end-extension of $\mathcal{M}$ satisfying $\mathrm{T}+W_e = s$.

The Enayat-Shavrukov proof applies only to reflexive theories, e.g. theories extending PA. It is also possible to derive the theorem directly from Theorem 7.4, by using the strength of Theorem 6.7. However, for finitely axiomatised theories such as $\mathrm{I}\Sigma_1$, more is needed to establish a similar result. This is indeed possible by using methods inspired by the characterisation of the modal logic of Π_1-conservativity over extensions of $\mathrm{I}\Sigma_1$ (Hájek and Montagna, 1990; Japaridze, 1994), thus establishing:

Theorem 7.6 (Blanck and Enayat, 2017)**.** Suppose that T is a consistent, r.e. extension of $\mathrm{I}\Sigma_1$. There is an r.e. set W_e, such that:

1. $\mathrm{I}\Sigma_1 \vdash \mathrm{Con}_{\mathrm{T}} \rightarrow W_e = \emptyset$;
2. for each countable model $\mathcal{M} \models \mathrm{T}$, if s is an $\mathcal{M}$-finite set such that $\mathcal{M} \models W_e \subseteq s$, then there is an end-extension of $\mathcal{M}$ satisfying $\mathrm{T} + W_e = s$.

The three theorems above are derived, using the characterisations of partial conservativity in Chapter 6, from the following lemma, which is an adaptation of the construction used by Shavrukov in the original proof of Theorem 7.5. The proof method for establishing these results differs from the method used in Chapter 5, in that the set W_e is not obtained by means of a partial satisfaction predicate, but rather constructed as a so called Solovay function. The intellectual heritage of the construction below goes from Solovay (1976), via Japaridze (1994), Woodin (2011), and the Enayat-Shavrukov manuscripts, to Blanck and Enayat (2017). The present concoction is inspired by all of these works.

Lemma 7.7. Suppose that T is a consistent, r.e. extension of $\mathrm{I}\Sigma_1$ in the same language. There is an r.e. set W_e, such that:

1. $\mathrm{I}\Sigma_1 \vdash \mathrm{Con}_\mathrm{T} \to W_e = \emptyset$;

2. for each $k \in \omega$, $\mathrm{T} \vdash \forall$ finite set $s(W_e \subseteq s \to \mathrm{Con}^k_{\mathrm{T},\Sigma_1+W_e=s})$.

Proof. The set W_e is defined in $\mathrm{I}\Sigma_1$, using the recursion theorem, by the stages $W_{e,x}$ in which it acquires its elements. An auxiliary function $r(x)$ is simultaneously defined.

Stage 0: Set $W_{e,0} = \emptyset$, and $r(0) = \infty$.[14]

Stage $x + 1$: Suppose $r(x) = m$. There are two cases:

> **Case A**: $s \supseteq W_{e,x}$, $n < m$, and x witnesses a Σ_1 sentence $\sigma(s)$ such that n is a proof in T of $\forall t(\sigma(t) \to W_e \neq t)$. Then set $W_{e,x+1} = s$ and $r(x + 1) = n$;
>
> **Case B**: otherwise, set $W_{e,x+1} = W_{e,x}$ and $r(x + 1) = m$.

Let $W_e = \bigcup_x W_{e,x}$.

Provably in $\mathrm{I}\Sigma_1$, $W_{e,x+1} \supseteq W_{e,x}$, and $r(x + 1) \leq r(x)$. That a limit $R = \lim_x r(x)$ exists is shown by repeating an argument due to Beklemishev and Visser (2005). Reason in $\mathrm{I}\Sigma_1$:

> By the Σ_1-least number principle, let R be such that
>
> $$\exists x(r(x) = R)) \wedge \forall y{<}R \neg\exists x(r(x) = y)$$
>
> i.e., R is the least value attained by r; then $\forall x(r(x) \geq R)$. If m is such that $r(m) = R$, then $\forall x{\geq}m(R \geq r(x))$ since $\forall x(r(x) \geq r(x + 1))$. It follows that $\forall x{\geq}m(r(x) = R)$, so the limit of r exists.

For each x with $W_{e,x+1} \neq W_{e,x}$, $\mathrm{I}\Sigma_1$ proves $r(x + 1) < r(x)$, whence there are only finitely many such x. So $\mathrm{I}\Sigma_1 \vdash$ "W_e is finite".

Note that for each $k \in \omega$, $\mathrm{T} \vdash R > k$. Fix $k \in \omega$ and argue in T:

> Suppose $R \leq k$. Let y be minimal such that $r(y + 1) = R$. Then $W_e = W_{e,y+1} = s$ for some s such that R is a proof of $\forall t(\sigma(t) \to W_e \neq t)$, where $\sigma(s)$ is a true Σ_1 sentence.

[14] Here, and in what follows, ∞ is a fictitious number greater than all the natural numbers. It is possible to replace ∞ with an ordinary number, at the cost of making the definition of W_e and $r(x)$ less transparent.

But, by small reflection (Fact 2.25),

> since $W_e \neq s$ is proved from a true Σ_1 sentence with a T-proof not exceeding k, it must be true. The contradiction proves $R > k$.

To prove (1), argue for the contrapositive statement in $\mathrm{I}\Sigma_1$:

> If $W_e = s \neq \emptyset$, then $\mathrm{Pr}_{\mathrm{T}}^n(\ulcorner \forall t(\sigma(t) \to W_e \neq t)\urcorner)$ for some n.
>
> Since W_e is finite, $s \subseteq W_e$ is Σ_1. Then $\mathrm{Pr}_{\mathrm{T}}(\ulcorner s \subseteq W_e \urcorner)$ follows by formalised Σ_1-completeness. Now reason inside Pr_{T}:
>
> > There is $u = W_e$ with $u \supseteq s$, so by construction, $\sigma(u)$ is true, and $\mathrm{Pr}_{\mathrm{T}}^m(\ulcorner \forall t(\sigma(t) \to W_e \neq t)\urcorner)$ for some $m \leq n$.
>
> Using formalised small reflection, continue reasoning inside Pr_{T}:
>
> > Then $\forall t(\sigma(t) \to W_e \neq t)$ and $\sigma(u)$, so $W_e \neq u$.
>
> But then $\mathrm{Pr}_{\mathrm{T}}(\ulcorner W_e = u \wedge W_e \neq u\urcorner)$, so $\neg\mathrm{Con}_{\mathrm{T}}$ as desired.

To prove (2), first fix $k \in \omega$. By small reflection, there is a proof n in T of

$$\forall t(\mathrm{Pr}_{\mathrm{T},\Sigma_1}^k(\ulcorner W_e \neq \dot{t}\urcorner) \to W_e \neq t).$$

Now reason in T:

> Consider any finite $s \supseteq W_e$. Suppose x is a k-proof of $W_e \neq s$ in $\mathrm{T} + \mathrm{Th}_{\Sigma_1}(\mathbb{N})$. Then $s \supseteq W_{e,x+1}$, and therefore $r(x+1) \leq n$ by construction of $r(x+1)$: here $\mathrm{Pr}_{\mathrm{T},\Sigma_1}^k(\ulcorner W_e \neq s\urcorner)$ is a true sentence playing the role of $\sigma(s)$. But $n \leq R < r(x+1)$, and the contradiction proves $\mathrm{Con}_{\mathrm{T},\Sigma_1+W_e=s}^k$. □

With this lemma in place, it is easy to derive Theorems 7.4 through 7.6.

Proof of Theorems 7.4 and 7.6. Let T be a consistent, r.e. extension of $\mathrm{I}\Sigma_1$, and let W_e be as in Lemma 7.7. Let $\mathcal{M}$ be a countable model of T, and let s be an $\mathcal{M}$-finite set such that $\mathcal{M} \models W_e \subseteq s$.

By Lemma 7.7, for each $k \in \omega$, $\mathrm{T} \vdash W_e \subseteq s \to \mathrm{Con}_{\mathrm{T},\Sigma_1+W_e=s}^k$. But by Theorem 6.8 ((4) ⇒ (1)), with $\mathcal{M}$ being countable, this implies that $\mathcal{M}$ can be end-extended to a model satisfying $\mathrm{T} + W_e = s$. □

Proof of Theorem 7.5. This follows directly from Theorem 7.4, by coupling its conclusion with Theorem 6.7. □

7.2 Digression: On coding schemes

In fact, the theorem proved in Woodin (2011) is not phrased in terms of r.e. sets, but rather in terms of Turing machines and finite binary sequences. The correspondence between Turing machines, r.e. sets and Σ_1-formulae is easy to establish, but the translation between finite binary sequences and finite sets in this setting require a few words on coding.

Theorem 7.8 (Woodin, 2011, Theorem 5)**.** There exists $e_0 \in \omega$ such that for all countable models $\mathcal{M} \models \mathrm{PA}$, if s is the output of the Turing machine with program e_0 within $\mathcal{M}$, and if t is an internal binary sequence of $\mathcal{M}$ such that s is a proper initial segment of t, then there exists a countable model $\mathcal{N} \models \mathrm{PA}$ such that $\mathcal{M}$ is a proper initial segment of $\mathcal{N}$ and such that t is the output of the Turing machine with program e_0 within $\mathcal{N}$.

This theorem can be derived from Theorem 7.4 in the following manner. Fix a recursive function $h : \omega \to {}^{<\omega}2$ such that for every $t \in {}^{<\omega}2$ there are infinitely many i such that $h(i) = t$. For the base case, suppose that n is the first stage at which W_e is non-empty, and that $W_{e,n} = s_0$ for some finite set s_0. Then the desired program e_0 enumerates s_0 in increasing order as $\{a_i : i < k\}$ for k the size of s_0 and outputs the sequence

$$f(s_0) = h(a_0)^\frown \ldots {}^\frown h(a_{k-1}).$$

In general, if $W_{e,x+1} \neq W_{e,x}$, then e_0 enumerates $W_{e,x+1} \setminus W_{e,x}$ in increasing order as $\{b_j : j \leq m\}$ and replaces its previous output $f(s_x)$ by $f(s_{x+1}) = f(s_x)^\frown u$, where $u = h(b_0)^\frown \ldots {}^\frown h(b_{m-1})$.

Suppose that the output of e_0 within $\mathcal{M}$ is s and let t be a proper, $\mathcal{M}$-finite prolongation of s. Then $s = h(a_0)^\frown \ldots {}^\frown h(a_i)$ for some $a_0, \ldots, a_i$ such that $A = \{a_0, \ldots, a_i\} = W_e$ within $\mathcal{M}$. Let u be a sequence such that $s^\frown u = t$. By definition of h, there is an $n \in \omega$ such that $n > a_i$ for all $a_i \in A$ and such that $h(n) = u$. By Theorem 7.4, there is an end-extension $\mathcal{N}$ of $\mathcal{M}$ in which $W_e = A \cup \{n\}$. But then the output of e_0 within $\mathcal{N}$ is precisely $s^\frown u = t$, as desired. This coding scheme can also be applied to obtain sequence versions of Theorems 7.5 and 7.6.

To derive Theorem 7.4 from Theorem 7.8, first fix a recursive function $g : {}^{<\omega}2 \to \omega^{<\omega}$. This g is required to have the additional property that

for every $t \in {}^{<\omega}2$, and every $s \in \omega^{<\omega}$, there is an u, properly extending t, such that $g(u) = s$. For example, fix an enumeration $\langle s_i, i \in \omega \rangle$ of $\omega^{<\omega}$ in which every element occurs infinitely often, and for any $t \in {}^{<\omega}2$, let $g(t) = s_n$ where n is the length of t. Now, define W_e in stages: if t_0 is the first output of e_0, then set $W_{e,1} = g(t_0)$, and generally, as soon as e_0 generates a new output t_{x+1}, then $W_{e,x+1} = g(t_x) \cup g(t_{x+1})$.

The paper Woodin (2011) also contains a notable philosophical interpretation of (a corollary to the proof of) Theorem 7.8. Although a thorough treatment of Woodin's philosophical argument is beyond the scope of this thesis, there is reason to elaborate on the construction of the index e_0^* that is used in his argument. Woodin writes:

> Our construction of e_0 actually gives rise to a Turing program e_0^* that witnesses a more dramatic version of the property just discussed. (Woodin, 2011, p. 120)

Even though Woodin formally, and informally, describes the behaviour of this new program e_0^*, there are no details on how to actually construct the desired program, and that lacuna is bridged here.

Suppose that s is the output of Woodin's program e_0 within a model $\mathcal{M}$ of PA. A central point of Woodin's argumentation is that the piece of information by which the sequence s is prolonged can be added as the one and only new entry in an end-extension $\mathcal{N}$, that is, *in a single step*. Let t be a sequence to be added to the output of e_0, i.e. such that $s^\frown t$ is the total output of e_0 in $\mathcal{N}$.

If t can be added in a single step, then there can be no intermediate end-extension $\mathcal{K}$ such that $\mathcal{M} \subset_e \mathcal{K} \subset_e \mathcal{N}$, unless $W_{e_0}^{\mathcal{K}}$ (that is, W_{e_0} as calculated within $\mathcal{K}$) is equal to s or $s^\frown t$. Otherwise, the program e_0 would have to have added t in two separate parts t_1 and t_2, such that $W_{e_0}^{\mathcal{K}} = s^\frown t_1$ and $W_{e_0}^{\mathcal{N}} = s^\frown t_1{}^\frown t_2 = s^\frown t$, and this represent two steps, as the terminology is used.

The point of the following coding scheme is to describe how a program e_0^* can be constructed from the program e used in the proof of Theorem 7.5. Hence, this coding scheme treats the set-versions of Woodin's theorem. It is easy to see how a single-element extension of a set can be obtained in a single step, but not as much so in the case of e.g. a two-element extension.

Let $C(n)$ be the non-empty finite set canonically coded by $n+1$, e.g. by writing $n+1$ in base 2 and looking at the positions in which there is a 1. Let e_0^* be the Turing machine program which, at any given stage of computation, outputs $C(a_1) \cup \cdots \cup C(a_n)$ if the output of e at the same stage is $\{a_1, \ldots, a_n\}$.

Suppose that, in $\mathcal{M}$, $W_{e_0^*} = s$ for some $\mathcal{M}$-finite set s, and that t is an $\mathcal{M}$-finite set properly extending s. By definition of e_0^*, this means that $W_e = \{a_1, \ldots, a_n\}$ with $\mathcal{M} \models s = C(a_1) \cup \cdots \cup C(a_n)$. Let $u = t \setminus s$; then u is an $\mathcal{M}$-finite set. Choose $m \in M$ such that $\mathcal{M} \models u = C(m)$. Note that $m \notin s$. There is then an end-extension $\mathcal{N}$ of $\mathcal{M}$ such that $W_e = \{a_1, \ldots, a_n, m\}$ as calculated in $\mathcal{N}$, which in turn shows that

$$W_{e_0^*}^{\mathcal{N}} = C(a_1) \cup \cdots \cup C(a_n) \cup C(m) = t.$$

Hence the additional output $u = C(m)$ is given in a single step.

The recently established fact that the Woodin-like theorems can be modified in such a way that any additional output can be added in a single step has some repercussions on the structure of possible end-extensions obtained by those theorems. For example, fix $\mathcal{M}$ and suppose $W_e^{\mathcal{M}} = s$. Let $t = \{m_1, m_2\}$, with $m_1, m_2 \in M$. By Theorem 7.5, there is an end-extension $\mathcal{N}_1$ of $\mathcal{M}$ such that $W_e^{\mathcal{N}_1} = s \cup \{m_1\}$, and an end-extension $\mathcal{N}_2$ of $\mathcal{M}$ such that $W_e^{\mathcal{N}_2} = s \cup \{m_2\}$. Moreover, there is an end-extension $\mathcal{N}_{1,2}$ of $\mathcal{M}$ such that $W_e^{\mathcal{N}_{1,2}} = s \cup \{m_1, m_2\}$. If the additional outputs all are given in a single step, then $\mathcal{N}_1 \not\subseteq_e \mathcal{N}_{1,2}$ and $\mathcal{N}_2 \not\subseteq_e \mathcal{N}_{1,2}$ even though each of $\mathcal{N}_1$, $\mathcal{N}_2$ and $\mathcal{N}_{1,2}$ end-extends $\mathcal{M}$.

Put in another way: if an element m is added to W_e at a stage k, then it is only possible to add new elements to W_e at stages $> k$. Supposing that $\mathcal{M}$ is such that the latest addition to W_e occurred at stage k, there are end-extensions $\mathcal{N}_1$ and $\mathcal{N}_{1,2}$ as above, both in which the latest addition to W_e occurred at stage $k+1$, but such that 'the set which was added to W_e at stage $k+1$' is $\{m_1\}$ in the first model and $\{m_1, m_2\}$ in the latter. Then, even if $\mathcal{N}_1$ is end-extended to a model $\mathcal{N}_1'$ in which m_2 is added to W_e, so that $W_e^{\mathcal{N}_1'} = W_e^{\mathcal{N}_{1,2}}$, the Σ_1-theory of any of these two models is inconsistent with the theory of the other model.

7.3 Uniformly flexible formulae

The purpose of this section is to establish the following:

Theorem 7.9. Suppose that T is a consistent, r.e. extension of PA. For all $n > 1$, there is a Σ_n formula $\gamma(x)$ such that for any $\sigma(x) \in \Sigma_n$, $\mathrm{T} + \forall x(\gamma(x) \leftrightarrow \sigma(x))$ is interpretable in T.

The proof combines Kripke's trick from Chapter 4 with the Solovay functions of Section 7.1, and can very roughly be outlined as follows. Modify the set W_e of Theorem 7.6 so that it adds a single element every time it grows. Then the desired formula $\gamma(x)$ can be chosen as 'x satisfies the formula whose Gödel number is the latest addition to W_e'.

In more detail, the suggested modification amounts to turning W_e into a Solovay function $f(x)$, acquiring a new input-output pair when passing to an end-extension. This function is defined in T using the recursion theorem, in such a way that, as x increases, f reaches a limit after a finite number of steps. Formally, the limit of f is defined as the unique z satisfying the Σ_2 formula $\lambda(z) := \exists x \forall y \geq x(f(y) = z)$.

The function f is constructed to satisfy the additional property that, for all $n, k \in \omega$, $\mathrm{T} \vdash \mathrm{Con}_{\mathrm{T}|n+\lambda(k)}$. The desired formula $\gamma(x)$ is chosen to express 'x satisfies the formula whose Gödel number is the limit of f'.

For every $\mathcal{M} \models \mathrm{T}$ and every $\sigma(x) \in \Sigma_n$, the existence of a non-standard $m \in M$ such that $\mathcal{M} \models \mathrm{Con}_{\mathrm{T}|m+\lambda(\ulcorner\sigma\urcorner)}$ follows by the overspill principle. Fact 2.38 allows for the construction of an end-extension satisfying $\mathrm{T} + \lambda(\ulcorner\sigma\urcorner)$, and therefore, by Kripke's trick, $\forall x(\gamma(x) \leftrightarrow \sigma(x))$. The existence of the interpretation then follows from the OHGL characterisation. It remains to construct a Solovay function f with the desired properties, and give a formal definition of $\gamma(x)$.

Opposed to the focus of the earlier sections, the point here is not to qualify for what theories T this may hold, but rather to investigate if there can be a general method allowing for the construction of such a formula $\gamma(x)$. In order to avoid the proofs becoming needlessly complicated, suppose, for the remainder of this chapter, that every theory mentioned is essentially reflexive, e.g. an extension of PA. With some additional work, it is possible to replace PA with $\mathrm{I}\Sigma_1$ in the lemma below, but since the latter theory is

finitely axiomatisable and therefore not reflexive, the interpretability result would not follow.

Lemma 7.10. Let T be a consistent, r.e. extension of PA. There is a recursive function f such that with

$$\lambda(z) := \exists x \forall y \geq x (f(y) = z)$$

the following holds:

1. $\mathrm{PA} \vdash \exists z \lambda(z)$;
2. $\forall n, k \in \omega$, $\mathrm{T} \vdash \mathrm{Con}_{\mathrm{T}|n+\lambda(k)}$.

Proof. The recursive function $f(x)$ is defined in PA, using the recursion theorem. An auxiliary rank function $r(x)$ is simultaneously defined.

Stage 0: $f(0) = 0$, $r(0) = \infty$.

Stage $x + 1$: Suppose $r(x) = m$. There are two cases:

> **Case A**: $n < m$ and x is a proof in $\mathrm{T}|n$ of $\neg\lambda(k)$. Then set $f(x+1) = k$, and $r(x+1) = n$;
>
> **Case B**: otherwise, set $f(x+1) = f(x)$, and $r(x+1) = m$.

Provably in PA, $r(x+1) \leq r(x)$, so there is an $R = \lim_x r(x)$. Every time the value of f changes, $r(x+1) < r(x)$, hence there can be at most finitely many such x. It follows that (1) $\mathrm{PA} \vdash \exists z \lambda(z)$.

As in the proof of Lemma 7.7, note that for all $n \in \omega$, $\mathrm{T} \vdash R > n$. Fix $n \in \omega$, and reason in T:

> Suppose $R \leq n$. Let x be minimal such that $r(x+1) = R$. Then $f(x+1) = k$ for some k, but since R is the limit of $r(x)$, k must be the limit of $f(x)$. Hence $\lambda(k)$ holds. By construction of f and r, $\mathrm{T}|R$ proves $\neg\lambda(k)$.

In view of essential reflexivity of T,

> since $\neg\lambda(k)$ is proved by $\mathrm{T}|R$ and therefore by $\mathrm{T}|n$, it must be true. The contradiction proves $R > n$.

For (2), fix $n, k \in \omega$ and reason in T:

> Suppose x is a proof of $\neg\lambda(k)$ in $\mathrm{T}|n$. By construction of r, $r(x+1) \leq n$, but as shown above, $n < R \leq r(x+1)$ holds. The contradiction proves $\mathrm{Con}_{\mathrm{T}|n+\lambda(k)}$. □

Proof of Theorem 7.9. Pick $n > 1$. Let f be as in Lemma 7.10, and let

$$\gamma(x) := \exists z(\lambda(z) \wedge \mathrm{Sat}_{\Sigma_n}(z, x)).$$

By definition of $\lambda(z)$, $\gamma(x)$ is $\Sigma_{\max(2,n)}$. Let $\sigma(x)$ be any Σ_n formula, and let $\mathcal{M} \models \mathrm{T}$. It follows from Lemma 7.10 that $\mathcal{M} \models \mathrm{Con}_{\mathrm{T}|n+\lambda(\ulcorner\sigma\urcorner)}$ for all $n \in \omega$, so by overspill, $\mathcal{M} \models \mathrm{Con}_{\mathrm{T}|m+\lambda(\ulcorner\sigma\urcorner)}$ for some non-standard $m \in M$. The arithmetised completeness theorem (Fact 2.38) can now be used to construct the desired end-extension $\mathcal{K} \models \mathrm{T} + \lambda(\ulcorner\sigma\urcorner)$. But since there can only be one object satisfying $\lambda(z)$ in $\mathcal{K}$, it follows that $\mathcal{K} \models \mathrm{T} + \forall x(\gamma(x) \leftrightarrow \sigma(x))$. Since $\mathcal{M}$ was arbitrary, $\mathrm{T} + \forall x(\gamma(x) \leftrightarrow \sigma(x))$ is interpretable in T by the OHGL characterisation. □

This result can be used to prove a variation of Hamkins's theorem 5.3. The result below is stronger in that it provides end-extensions of any model of PA, but weaker in that the formula $\xi(x)$ can no longer be assured to be Σ_1.

Corollary 7.11. There is a Σ_2 formula $\xi(x)$ such that if $\mathcal{M} \models \mathrm{T}$, and f is an $\mathcal{M}$-definable function in ${}^M2$, then $\mathcal{M}$ can be end-extended to a model satisfying $\mathrm{T} + \{\xi(m)^{f(m)} : m \in M\}$.

Proof. Let $\mathcal{M}$ be any model of T, let $f \in {}^M2$ be $\mathcal{M}$-definable, let $\xi(x)$ be a Σ_2 formula as in Theorem 7.9, and let $X = \{\xi(m)^{f(m)} : m \in M\}$.

The proof is similar to the proof of Theorem 5.3, but using the flexible formula $\gamma(x)$ of Theorem 7.9 in place of the one from Theorem 5.1. □

This theorem is in one sense the best possible, in that the restriction to $\mathcal{M}$-definable functions is essential.

Remark 7.12. Let $\xi(x)$ be any formula. There is a function $f \in {}^\omega 2$ and a model $\mathcal{M} \models \mathrm{PA}$ such that $\mathcal{M}$ has no end-extension satisfying $\mathrm{PA} + \{\xi(n)^{f(n)} : n \in \omega\}$.

Proof. Let $\xi(x)$ be any formula and let $\mathcal{M}$ be a countable non-standard model of PA. Suppose that, for each $f \in {}^{\omega}2$, there is an end-extension $\mathcal{N}_f$ of $\mathcal{M}$, satisfying $\mathrm{T}_f = \mathrm{PA} + \{\xi(n)^{f(n)} : n \in \omega\}$.

If $\xi(x)$ is not independent over PA, this immediately leads to a contradiction. Hence, assume that $\xi(x)$ is independent over PA so that for any choice of f, the theory T_f is consistent and has a model $\mathcal{N}_f$.

Suppose that $\mathcal{N}_f$ end-extends $\mathcal{M}$: then $\xi(x)$ represents a set X_f in $\mathcal{N}_f$, and by Fact 2.34, $X_f \in \mathrm{SSy}(\mathcal{M})$. For any two different f, g, it must be the case that $X_f \neq X_g$. But since there are $2^{\aleph_0}$ different f's, and $\mathcal{M}$ was chosen to be countable, it is impossible for $\mathrm{SSy}(\mathcal{M})$ to contain all the possible X_f's. □

This result is related to a question posed by Taishi Kurahashi in private communication. He asked whether there can be a Σ_1 formula such that $\mathrm{T} + \{\neg\xi(n) : n \in \omega\}$ is consistent, and for every $f \in {}^{\omega}2$, the theory $\mathrm{T} + \{\xi(n)^{f(n)} : n \in \omega\}$ is Π_1-conservative over $\mathrm{T}+\{\neg\xi(n) : n \in \omega\}$. In light of Theorem 6.11, the examples produced above are not counterexamples to the possible Π_1-conservativity, since that would require an f such that $\{n \in \omega : \mathcal{M} \models \xi(n)^{f(n)}\}$ is in $\mathrm{SSy}(\mathcal{M})$, but such that $\mathcal{M}$ has no end-extension to a model satisfying $\mathrm{T} + \{\xi(n)^{f(n)} : n \in \omega\}$.

7.4 Partial results on uniformly flexible Σ_1 formulae

With the strong result of the previous section safely in hand, the question remains whether there can exist a Σ_1 formula with similar properties. It is easy to see that the answer to the unqualified version of Question 7.3 is negative for $n = 1$:

Suppose, that there is a Σ_1 formula $\gamma(x)$ that is uniformly flexible for Σ_1 over PA. Then by the OHGL characterisation, for every $\sigma(x) \in \Sigma_1$, every $\mathcal{M} \models \mathrm{PA}$ has an end-extension to a model of $\mathrm{PA} + \forall x(\gamma(x) \leftrightarrow \sigma(x))$. If $\mathcal{M}$ is chosen such that $\mathcal{M} \models \exists x\gamma(x)$ and $\sigma(x)$ is chosen as the Σ_1 formula $x \neq x$, then this immediately leads to a contradiction. Hence the question has to be qualified to rule out pathological counterexamples, suggesting the following version which is phrased in terms of end-extensions:

Question 7.13. Is there a Σ_1 formula $\gamma(x)$ such that for every $\sigma(x) \in \Sigma_1$, each model of $\mathrm{PA} + \forall x(\gamma(x) \to \sigma(x))$ has an end-extension to a model of $\mathrm{PA} + \forall x(\gamma(x) \leftrightarrow \sigma(x))$?

This question has a trivial positive answer, by letting $\gamma(x) := x = x$. Then every model satisfying $\mathrm{PA} + \forall x(\gamma(x) \to \sigma(x))$ must also satisfy $\forall x(\gamma(x) \leftrightarrow \sigma(x))$. A further qualification is:

Question 7.14. Is there a Σ_1 formula $\gamma(x)$ such that for every $\sigma(x) \in \Sigma_1$, and every $\mathcal{M} \models \mathrm{PA}$,

1. if $\mathcal{M} \models \mathrm{Con}_{\mathrm{PA}}$, then $\mathcal{M} \models \forall x \neg\gamma(x)$;
2. if $\mathcal{M} \models \mathrm{PA} + \forall x(\gamma(x) \to \sigma(x))$ then there is an end-extension of $\mathcal{M}$ satisfying $\mathrm{PA} + \forall x(\gamma(x) \leftrightarrow \sigma(x))$?

By reformulating Theorem 7.5 in terms of Σ_1 formulae rather than indices for r.e. sets it is easy to see that the answer to the question is positive when the choice of $\sigma(x)$ is restricted to formulae with finite extensions in the base model $\mathcal{M}$. Theorem 5.1 shows that if the answer to the question is negative in general, then a counterexample must come in the shape of a $\sigma(x) \in \Sigma_1$ and a model of $\mathrm{PA} + \neg\mathrm{Con}_{\mathrm{PA}} + \forall x(\gamma(x) \to \sigma(x))$ with no end-extension to a model of $\mathrm{PA} + \forall x(\gamma(x) \leftrightarrow \sigma(x))$. However, if attention is restricted to precisely models of $\mathrm{PA} + \neg\mathrm{Con}_{\mathrm{PA}}$, it is easy to construct a silly formula having *almost* the properties asked for:

Remark 7.15. There is a Σ_1 formula $\gamma(x)$, such that for every $\sigma(x) \in \Sigma_1$ and every $\mathcal{M} \models \mathrm{PA}$,

1. if $\mathcal{M} \models \mathrm{Con}_{\mathrm{PA}}$, then $\mathcal{M} \models \forall x \neg\gamma(x)$;
2. if $\mathcal{M} \models \neg\mathrm{Con}_{\mathrm{PA}} + \forall x(\gamma(x) \to \sigma(x))$, then there is an end-extension of $\mathcal{M}$ satisfying $\mathrm{PA} + \forall x(\gamma(x) \leftrightarrow \sigma(x))$.

Proof. Let $\gamma(x) := \neg\mathrm{Con}_{\mathrm{PA}} \wedge (x = x)$. Then (1) is immediate. For (2), pick any $\mathcal{M}$ that satisfies $\mathrm{PA} + \neg\mathrm{Con}_{\mathrm{PA}} + \forall x(\gamma(x) \to \sigma(x))$. Then $\mathrm{PA} + \forall x(\gamma(x) \leftrightarrow \sigma(x)) + \mathrm{Th}_{\Sigma_1}(\mathcal{M})$ is consistent. The existence of the desired end-extension is now a consequence of Theorem 6.2. □

This result does not answer Question 7.14, since it is not guaranteed that any model of PA + $\mathrm{Con}_{\mathrm{PA}}$ can be end-extended to a model of PA + $\forall x(\gamma(x) \leftrightarrow \sigma(x))$. Neither does the next result or its corollaries give exactly what is asked for, but here the failure comes in the possibly non-standard shape of the formula being satisfied in the end-extension.

Theorem 7.16. There is a Σ_1 formula $\gamma(x)$ such that for every $\mathcal{M} \models \mathrm{PA}$:

1. if $\mathcal{M} \models \mathrm{Con}_{\mathrm{PA}}$, then $\mathcal{M} \models \forall x \neg\gamma(x)$, and for each $\sigma(x) \in \Sigma_1$, there is an end-extension $\mathcal{N}_0 \models \mathrm{PA}$ of $\mathcal{M}$ such that
$$\mathcal{N}_0 \models \forall x(\gamma(x) \leftrightarrow \sigma(x));$$
2. otherwise, there is an $\mathcal{M}$-finite set of possibly non-standard Σ_1 formulae $\sigma_0(x), \ldots, \sigma_s(x)$ such that
$$\mathcal{M} \models \forall x(\gamma(x) \leftrightarrow \sigma_0(x) \vee \cdots \vee \sigma_s(x))$$
and for every $\sigma(x) \in \Sigma_1$, there is an end-extension $\mathcal{N}_1 \models \mathrm{PA}$ of $\mathcal{M}$ such that
$$\mathcal{N}_1 \models \forall x(\gamma(x) \leftrightarrow \sigma_0(x) \vee \cdots \vee \sigma_s(x) \vee \sigma(x)).$$

Proof. Suppose, without loss of generality, that $\mathrm{Sat}_{\Sigma_1}(z, x)$ is such that every value of z encodes a Σ_1 formula. Let W_e be as in Theorem 7.5, and let, abusing language, $\gamma(x)$ be the Σ_1 formula $\exists z(z \in W_e \wedge \mathrm{Sat}_{\Sigma_1}(z, x))$.

For (1), suppose that $\mathcal{M} \models \mathrm{Con}_{\mathrm{PA}}$. Then by Theorem 7.5, $W_e = \emptyset$ as calculated in $\mathcal{M}$. Hence $\gamma(x)$ is false for every x. Let $\sigma(x)$ be any Σ_1 formula. Again by the theorem, there is an end-extension of $\mathcal{M}$ in which $W_e = \{\ulcorner\sigma\urcorner\}$, and the conclusion follows.

For (2), let $\{a_0, a_1, \ldots, a_s\}$ be W_e as calculated within $\mathcal{M}$. For each $i \leq s$ let $\sigma_i(x)$ be the Σ_1 formula $\mathrm{Sat}_{\Sigma_1}(a_i, x)$. By reasoning within $\mathcal{M}$ it is easy to ascertain that
$$\mathcal{M} \models \forall x(\gamma(x) \leftrightarrow \sigma_0(x) \vee \cdots \vee \sigma_s(x)).$$
Let $\sigma(x)$ be any Σ_1 formula. By Theorem 7.5, there is an end-extension $\mathcal{N}_1 \models \mathrm{PA}$ of $\mathcal{M}$ in which $W_e = \{a_0, \ldots, a_s, \ulcorner\sigma\urcorner\}$. It follows that
$$\mathcal{N}_1 \models \forall x(\gamma(x) \leftrightarrow \sigma_0(x) \vee \cdots \vee \sigma_s(x) \vee \sigma(x)). \qquad \square$$

Remark 7.17. This is as good as it gets when it comes to applying Kripke's method to the problem at hand. The general form of the formulae $\gamma(x)$ used in Kripke-style proofs is $\exists z(\phi(z) \wedge \mathrm{Sat}_{\Sigma_n}(z, x))$, where the unique object satisfying $\phi(z)$ can change when passing to an end-extension. The uniqueness is necessary for $\gamma(x)$ to have the same extension as the chosen formula $\sigma(x)$. If the condition $\phi(z)$ is Σ_1, new witnesses to $\phi(z)$ can be added, but the old ones can not be lost due to Σ_1-persistence. This means that there can never be a new unique object satisfying $\phi(z)$ in the end-extension.

Corollary 7.18. Suppose that $\gamma(x)$ is as in Theorem 7.16, and that W_e as calculated within $\mathcal{M}$ is a finite set of natural numbers when viewed externally. Then there is a standard formula $\sigma_0(x) := \bigvee_{a_i \in W_e} \mathrm{Sat}_{\Sigma_1}(a_i, x)$ such that $\mathcal{M} \models \forall x(\gamma(x) \leftrightarrow \sigma_0(x))$, and for every $\sigma(x) \in \Sigma_1$, there is an end-extension $\mathcal{N}_2 \models \mathrm{PA}$ of $\mathcal{M}$ such that

$$\mathcal{N}_2 \models \forall x(\gamma(x) \leftrightarrow \sigma_0(x) \vee \sigma(x)).$$

Corollary 7.19. If $\mathcal{M}$, $\gamma(x)$ and $\sigma_0(x)$ are as in Corollary 7.18, then if $\sigma(x) \in \Sigma_1$ is chosen such that $\mathrm{PA} \vdash \forall x(\sigma_0(x) \rightarrow \sigma(x))$, there is an end-extension $\mathcal{N}_3 \models \mathrm{PA}$ of $\mathcal{M}$ such that

$$\mathcal{N}_3 \models \forall x(\gamma(x) \leftrightarrow \sigma(x)).$$

The best partial result available at this point is the one below. It is due to Volodya Shavrukov, and it is included here with his graceful permission. For technical reasons, the result is phrased in terms of (indices for) r.e. sets, rather than Σ_1-formulae, but the translation between these is, as always, straightforward.

Theorem 7.20 (Shavrukov, unpublished)**.** Suppose that T is a consistent, r.e. extension of PA. There is an r.e. set W_e such that:

1. $\mathbb{N} \models W_e = \emptyset$;
2. for each $j \in \omega$, every model $\mathcal{M} \models \mathrm{T}$ can be end-extended to a model $\mathcal{K} \models \mathrm{T} + W_e =^* W_j$, where $=^*$ is equality modulo finite differences.

Proof. Using the recursion theorem, the set W_e is defined (provably in PA) by the stages $W_{e,x}$ in which it acquires its elements, together with an auxiliary function $r(x)$. Assume that if nothing is added to a set W_n at stage x, then $W_{n,x} = \emptyset$.

Stage 0: $W_{e,0} = \emptyset$. $r(0) = \infty$. Assume that the ordering of the r.e. sets is such that W_∞ is the empty set.[15]

Stage $x + 1$: Suppose $r(x) = m$.

> **Case A**: there is a proof in $\mathrm{T}|n$ of $W_e \neq^* W_m$ and $n < m$. Then let $W_{e,x+1}$ be $W_{n,x+1}$, and set $r(x + 1) = n$;
>
> **Case B**: otherwise, $W_{e,x+1} = W_{m,x+1}$, and $r(x + 1) = m$.

Let $W_e = \bigcup_{x \in \omega} W_{e,x}$.

The idea is that W_e remains empty until the procedure encounters an n such that $\mathrm{T}|n$ proves that $W_e \neq^* \emptyset$. At this stage, the procedure starts enumerating W_n instead, until it encounters a smaller n' such that $\mathrm{T}|n'$ proves that $W_e \neq^* W_{n'}$, et cetera. Hence, at each stage x, x limits the number of objects that have been put in W_e, as well as the number of objects having been left out from the r.e. set $W_{r(x)}$ currently enumerated, suggesting that if this process comes to an end, for some y, W_e is equal to $W_{r(y)}$ *modulo finite differences*.

Provably in PA, $r(x + 1) \leq r(x)$, so there exists $R = \lim_x r(x)$. Since at each stage x at which W_e starts enumerating a new r.e. set, it must be the case that $r(x + 1) < r(x)$, so there can be at most finitely many such x. Hence:

> PA $\vdash$ "After a finite number of steps, W_e settles on which r.e. set it enumerates."

It also follows that $\mathrm{T} \vdash R > n$ for all $n \in \omega$. Fix n and reason in T:

> Suppose $R \leq n$. Let x be minimal such that $r(x + 1) = R$. Then for all $z > x$, $z \in W_e$ iff $z \in W_R$, which implies $W_e =^* W_R$. But by construction of $r(x + 1)$ it also follows that $\mathrm{T}|R \vdash W_e \neq^* W_R$.

[15] Should this assumption on the ordering of r.e. sets feel uncomfortable, it is easy to divide the construction of W_e in two separate procedures, bypassing the use of W_∞.

Since T is reflexive:

> $W_e \neq^* W_R$ is proved by $\mathrm{T}|R$ and therefore by $\mathrm{T}|n$, so it must be true. The contradiction proves $R > n$.

Since $R \leq n$ is T-equivalent to a Σ_1 sentence, it would imply its own provability in T. Since T is consistent, this implies $R = \infty$ in the real world. Hence $\mathbb{N} \models W_e = \emptyset$.

For (2), fix n and argue in T:

> Pick j and suppose x is a proof of $W_e \neq^* W_j$ in $\mathrm{T}|n$. By construction of r, $r(x+1) \leq n$. But $n < R \leq r(x+1)$, and the contradiction proves $\mathrm{Con}(\mathrm{T}|n + W_e =^* W_j)$.

Let $\mathcal{M} \models \mathrm{T}$, and pick a $j \in \omega$. By overspill, and the argument above, $\mathcal{M} \models \mathrm{Con}_{\mathrm{T}|m+W_e=^*W_j}$ for some non-standard $m \in M$. Fact 2.38 can now be used to construct the end-extension $\mathcal{K} \models \mathrm{T} + W_e =^* W_j$. □

7.5 Hierarchical generalisations: Asking the right question

As in the preceding chapters, it is time to discuss hierarchical generalisations. By modifying the proofs of Lemma 7.10 and Theorem 7.9 in a way that is detailed below, it is possible to prove the following:

Theorem 7.21. Let T be a consistent, r.e. extension of PA. There is a Σ_{n+2} formula $\gamma(x)$ such that for every $\sigma(x) \in \Sigma_{n+2}$, every model of T has a Σ_n-elementary extension satisfying $\mathrm{T} + \forall x(\gamma(x) \leftrightarrow \sigma(x))$.

The necessary modifications can be outlined as follows.

Proof sketch. Fix n and let a Δ_{n+1} function $f(x)$ be defined in PA by:

Stage 0: $f(0) = 0$, $r(0) = \infty$.

Stage $x+1$: Suppose $r(x) = m$. There are two cases:

> **Case A**: $i < m$ and x is a proof in $(\mathrm{T} + \mathrm{Th}_{\Sigma_{n+1}}(\mathbb{N}))|i$ of $\neg\lambda(k)$. Then set $f(x+1) = k$, and $r(x+1) = i$;
>
> **Case B**: otherwise, set $f(x+1) = f(x)$, and $r(x+1) = m$.

This construction is admissible by the recursion theorem (Fact 2.53), and f is then Δ_{n+1}, since by Craig's trick, $\mathrm{T} + \mathrm{Th}_{\Sigma_{n+1}}(\mathbb{N})$ has a deductively equivalent Π_n definition. Since f is Δ_{n+1}, the formula $\lambda(x)$ expressing that x is the limit of f is Σ_{n+2}. Most of the proof then goes through as it stands. For the final part, Let $\mathcal{M} \models \mathrm{T}$, fix $n, m, k \in \omega$ and reason in T:

> Suppose that x is a proof in $(\mathrm{T} + \mathrm{Th}_{\Sigma_{n+1}}(\mathbb{N}))|m$ of $\neg\lambda(k)$. Then $r(x+1) < m$ and $m < R \leq r(x+1)$. The contradiction proves $\mathrm{Con}_{(\mathrm{T},\Sigma_{n+1})|m+\lambda(k)}$.

Let $\gamma(x)$ be $\exists z(\lambda(z) \wedge \mathrm{Sat}_{\Sigma_{n+2}}(z, x))$ and let $\sigma(x)$ be any Σ_{n+2} formula. Since $\mathcal{M}$ satisfies overspill, there is a non-standard $m \in M$ such that $\mathcal{M} \models \mathrm{Con}_{(\mathrm{T},\Sigma_{n+1})|m+\lambda(\ulcorner\sigma\urcorner)}$. Now use Fact 2.38 to construct an end-extension $\mathcal{K}$ satisfying $\mathrm{T} + \mathrm{Th}_{\Sigma_{n+1}}(\mathcal{M}) + \lambda(\ulcorner\sigma\urcorner)$. Using Kripke's trick, conclude that $\mathcal{K} \models \mathrm{T} + \mathrm{Th}_{\Sigma_{n+1}}(\mathcal{M}) + \forall x(\gamma(x) \leftrightarrow \sigma(x))$, whence $\mathcal{K}$ is a Σ_n-elementary extension of $\mathcal{M}$. □

With this theorem in hand, it turns out that the success of Theorem 7.9 is a special case of a more general phenomenon. Hence, a more general question, informed by the questions asked at the outset of this chapter and the theorem above, can be phrased as:

Question 7.22. Is there a Σ_{n+1} formula $\gamma(x)$, such that $\mathbb{N} \models \forall x \neg\gamma(x)$, and for each Σ_{n+1} formula $\sigma(x)$, every model of $\mathrm{T} + \forall x(\gamma(x) \rightarrow \sigma(x))$ has a Σ_n-elementary extension satisfying $\mathrm{T} + \forall x(\gamma(x) \leftrightarrow \sigma(x))$?

Remark 7.23. The additional assumption that the base model satisfies $\forall x(\gamma(x) \rightarrow \sigma(x))$ reappears here to take into account the elementarity for Σ_n formulae. Similar to the case of Σ_1-persistence over end-extensions, Σ_{n+1} sentences persist when passing to a Σ_n-elementary extension.

An affirmative answer to this question would improve Theorems 7.5 and 7.21, as well as cast some light on the question asked in Section 7.4. Given the general formulations of the hierarchical generalisations throughout this thesis, it seems reasonable to claim that there is nothing special about the case $n = 0$, i.e., the case concerning Σ_1 formulae. Following this line of thought – if an answer is given for any n, it is likely that the same proof or disproof could be easily transformed to encompass the other cases as well.

8 Concluding remarks

A number of more or less interesting problems are left open in this thesis. The first three of these originate in Chapter 3:

Question 8.1. Can the collection of sets of fixed points over T be characterised among the creative sets?

Question 8.2. Is there a Γ formula $\theta(x)$ such that $\mathrm{Fix}_\Gamma(\theta)$ is neither recursive nor creative?

Question 8.3. Is $\mathscr{F}_\Gamma$ an ultrafilter on $\mathscr{R}_\Gamma$?

An observation is that it seems difficult to say something in general about the number of fixed points (up to provable equivalence) of a non-extensional formula. This difficulty is related to the long-standing problem of whether all Rosser sentences are provably equivalent or not.

A lesson to be learned from this thesis is the success of Kripke's method in establishing numerous generalisations of the first incompleteness theorem. The method is prevalent throughout Chapters 4, 5 and 7, and is used in some form in almost every proof. On the other hand, even when combined with the powerful method of Solovay functions, each instance of a very general problem remain unsolved:

Question 8.4. Suppose that T is a consistent, r.e. extension of PA. Is there a Σ_{n+1} formula $\gamma(x)$, such that $\mathbb{N} \models \forall x \neg\gamma(x)$, and for each Σ_{n+1} formula $\sigma(x)$, every model of $\mathrm{T} + \forall x(\gamma(x) \to \sigma(x))$ has a Σ_n-elementary extension satisfying $\mathrm{T} + \forall x(\gamma(x) \leftrightarrow \sigma(x))$?

In particular, in the case $n = 0$, an affirmative answer would establish the interpretability of $\mathrm{T} + \forall x(\gamma(x) \leftrightarrow \sigma(x))$ in $\mathrm{T} + \forall x(\gamma(x) \to \sigma(x))$, which would be a strengthening of Feferman's theorem on the interpretability of inconsistency. By complexity calculations, the possibility that Kripke's method can be used to establish such a result can be ruled out.

As a contrast, Theorem 7.21 can be interpreted as saying that very much can change at the Σ_{n+2} level of a model of PA, while leaving the Σ_n level completely untouched. Ensuring that the Σ_{n+1} level can be preserved seems to be much more difficult.

The final question listed here is due to Taishi Kurahashi.

Question 8.5. Is there a Σ_1 formula $\xi(x)$ such that $\mathrm{T} + \{\neg\xi(n) : n \in \omega\}$ is consistent, and for each $f \in {}^{\omega}2$, the theory $\mathrm{T} + \{\xi(n)^{f(n)} : n \in \omega\}$ is Π_1-conservative over $\mathrm{T} + \{\neg\xi(n) : n \in \omega\}$?

References

Beklemishev, L. (1998). Review: Per Lindström, Aspects of Incompleteness. *The Journal of Symbolic Logic*, 63(4):1606–1608.

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

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

Bennet, C. (1986). Lindenbaum algebras and partial conservativity. *Proceedings of the American Mathematical Society*, 97(2):323–327.

Bernardi, C. (1981). On the relation provable equivalence and on partitions in effectively inseparable sets. *Studia Logica*, 40(1):29–37.

Blanck, R. (2011). *Metamathematical fixed points*. Licentiate thesis, University of Gothenburg.

Blanck, R. (2016). Flexibility in fragments of Peano arithmetic. In Cégielski, P., Enayat, A., and Kossak, R., editors, *New Studies in Weak Arithmetics, Vol. 3*, number 217 in CSLI Lecture Notes, pages 1–20. CSLI Publications.

Blanck, R. and Enayat, A. (2017). Marginalia on a theorem of Woodin. *The Journal of Symbolic Logic*, 82(1):359–374.

Carnap, R. (1937). *The Logical Syntax of Language*. Routledge & Kegan Paul.

Chaitin, G. J. (1974). Information-theoretic limitations of formal systems. *Journal of the ACM*, 21:403–424.

Cornaros, Ch. and Dimitracopoulos, C. (2000). A note on end extensions. *Archive for Mathematical Logic*, 39:459–463.

Craig, W. (1953). On axiomatizability within a system. *The Journal of Symbolic Logic*, 18(1):30–32.

D'Aquino, P. (1993). A sharpened version of McAloon's theorem on initial segments of models of $I\Delta_0$. *Annals of Pure and Applied Logic*, 61:49–62.

Dimitracopoulos, C. and Paris, J. (1988). A note on a theorem of H. Friedman. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 34:13–17.

Ehrenfeucht, A. and Feferman, S. (1960). Representability of recursively enumerable sets in formal theories. *Archiv für mathematische Logik und Grundlagenforschung*, 5(1–2):37–41.

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

Feferman, S. (1962). Transfinite recursive progressions of axiomatic theories. *The Journal of Symbolic Logic*, 27(3):259–316.

Feferman, S., Kreisel, G., and Orey, S. (1962). 1-consistency and faithful interpretations. *Archiv für mathematische Logik und Grundlagenforschung*, 6(1–2):52–63.

Franzén, T. (2005). *Gödel's Theorem: An Incomplete Guide to its Use and Abuse*. A. K. Peters.

Gödel, K. (1930). Die Vollständigkeit der Axiome des logischen Funktionenkalküls. *Monatshefte für Mathematik und Physik*, 37:349–360.

Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme, I. *Monatshefte für Mathematik und Physik*, 38:173–198.

Grzegorczyk, A., Mostowski, A., and Ryll-Nardzewski, C. (1958). The classical and the ω-complete arithmetic. *The Journal of Symbolic Logic*, 23(2):188–206.

Guaspari, D. (1979). Partially conservative extensions of arithmetic. *Transactions of the American Mathematical Society*, 254:47–68.

Hájek, P. (1971). On interpretability in set theories. *Commentationes Mathematicae Universitatis Carolinae*, 12(1):73–79.

Hájek, P. (1987). Partial conservativity revisited. *Commentationes Mathematicae Universitatis Carolinae*, 28(4):679–690.

Hájek, P. (1993). Interpretability and fragments of arithmetic. In Clote, P. and Krajíček, J., editors, *Arithmetic, Proof Theory, and Computational Complexity*, volume 23 of *Oxford Logic Guides*, pages 185–196. Clarendon Press.

Hájek, P. and Montagna, F. (1990). The logic of Π_1-conservativity. *Archive for Mathematical Logic*, 30:113–123.

Hájek, P. and Pudlák, P. (1993). *Metamathematics of First Order Arithmetic*. Perspectives in Mathematical Logic. Springer-Verlag.

Halbach, V. and Visser, A. (2014). Self-reference in arithmetic I. *Review of Symbolic Logic*, 7(4):671–691.

Hamkins, J. D. (2016). Every function can be computable. `http://jdh.hamkins.org/every-function-can-be-computable/`. [Online; accessed 2017-04-30].

Harary, F. (1961). A very independent axiom system. *The American Mathematical Monthly*, 68(2):159–162.

Hilbert, D. and Bernays, P. (1934–1939). *Grundlagen der Mathematik*, volume 1–2. Springer-Verlag.

Japaridze, G. (1994). A simple proof of arithmetical completeness for Π_1-conservativity logic. *Notre Dame Journal of Formal Logic*, 35(3):346–354.

Jensen, D. and Ehrenfeucht, A. (1976). Some problem in elementary arithmetics. *Fundamenta Mathematicae*, 92(3):223–245.

Jeroslow, R. G. (1971). Consistency statements in formal theories. *Fundamenta Mathematicae*, 72(1):17–40.

Jeroslow, R. G. (1975). Experimental logics and Δ_2^0-theories. *Journal of Philosophical Logic*, 4(3):253–267.

Joosten, J. J. (2004). *Interpretability formalized*. PhD thesis, Utrecht University.

Kaså, M. (2012). Experimental logics, mechanism and knowable consistency. *Theoria*, 78(3):213–224.

Kaye, R. (1991). *Models of Peano Arithmetic*, volume 15 of *Oxford Logic Guides*. Clarendon Press.

Kikuchi, M. and Kurahashi, T. (2016). Illusory models of Peano arithmetic. *The Journal of Symbolic Logic*, 81(3):1163–1175.

Kikuchi, M. and Kurahashi, T. (20XX). Generalizations of Gödel's incompleteness theorems for Σ_n-definable theories of arithmetic. Manuscript.

Kleene, S. C. (1952). *Introduction to Metamathematics*. North-Holland.

Kossak, R. and Schmerl, J. (2006). *The Structure of Models of Peano Arithmetic*, volume 50 of *Oxford Logic Guides*. Clarendon Press.

Kreisel, G. (1962). On weak completeness of intuitionistic predicate logic. *The Journal of Symbolic Logic*, 27(2):139–158.

Kripke, S. A. (1962). "Flexible" predicates of formal number theory. *Proceedings of the American Mathematical Society*, 13(4):647–650.

van Lambalgen, M. (1989). Algorithmic information theory. *The Journal of Symbolic Logic*, 54(4):1389–1400.

Li, M. and Vitányi, P. (1993). *An Introduction to Kolmogorov Complexity and Its Applications*. Springer-Verlag.

Lindström, P. (1979). Some results on interpretability. In *Proceedings of the 5th Scandinavian Logic Symposium 1979*, pages 329–361.

Lindström, P. (1984). A note on independent formulas. In *Notes on formulas with prescribed properties in arithmetical theories*, number 25 in Philosophical Communications, Red Series, pages 1–5. Göteborgs Universitet.

Lindström, P. (2003). *Aspects of Incompleteness*. Number 10 in Lecture Notes in Logic. A. K. Peters, 2nd edition.

Lindström, P. and Shavrukov, V. Yu. (2008). The $\forall\exists$ theory of Peano Σ_1 sentences. *Journal of Mathematical Logic*, 8(2):251–280.

Löb, M. H. (1955). Solution of a problem of Leon Henkin. *The Journal of Symbolic Logic*, 20(2):115–118.

Löwenheim, L. (1915). Über Möglichkeiten im Relativkalkül. *Mathematische Annalen*, 76(4):447–470.

Maltsev, A. (1936). Untersuchungen aus dem Gebiete der mathematischen Logik. *Matématičéskij sbornik, n.s.*, 1:323–336.

McAloon, K. (1982). On the complexity of models of arithmetic. *The Journal of Symbolic Logic*, 47(2):403–415.

Montagna, F. (1982). Relatively precomplete numerations and arithmetic. *Journal of Philosophical Logic*, 11(4):419–430.

Montague, R. (1962). Theories incomparable with respect to relative interpretability. *The Journal of Symbolic Logic*, 27(2):195–211.

Mostowski, A. (1952). On models of axiomatic systems. *Fundamenta Mathematicae*, 39(1):133–158.

Mostowski, A. (1961). A generalization of the incompleteness theorem. *Fundamenta Mathematicae*, 49(2):205–232.

Myhill, J. (1955). Creative sets. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 1:97–108.

Myhill, J. (1972). An absolutely independent set of Σ^0_1-sentences. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 18:107–109.

Orey, S. (1961). Relative interpretations. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 7:146–153.

Paris, J. B. (1981). Some conservations results for fragments of arithmetic. In Berline, C., McAloon, K., and Ressayre, J., editors, *Model Theory and Arithmetic*, volume 890 of *Lecture notes in Mathematics*, pages 251–262. Springer-Verlag.

Post, E. L. (1948). Degrees of recursive unsolvability. *Bulletin of the American Mathematical Society*, 54(7):641–642.

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

Putnam, H. and Smullyan, R. M. (1960). Exact separation of recursively enumerable sets within theories. *Proceedings of the American Mathematical Society*, 11:574–577.

Raatikainen, P. (1998). On interpreting Chaitin's incompleteness theorem. *Journal of Philosophical Logic*, 27(6):569–586.

Ressayre, J.-P. (1987). Nonstandard universes with strong embeddings, and their finite approximations. In *Logic and Combinatorics*, volume 65 of *Contemporary Mathematics*, pages 333–358. American Mathematical Society.

Robinson, A. (1963). On languages which are based on non-standard arithmetic. *Nagoya Mathematical Journal*, 22:83–117.

Rogers, Jr., H. (1967). *Theory of Recursive Functions and Effective Computability*. McGraw-Hill.

Rosser, J. B. (1936). Extensions of some theorems of Gödel and Church. *The Journal of Symbolic Logic*, 1(3):87–91.

Salehi, S. and Seraji, P. (2016). Gödel-Rosser's incompleteness theorem, generalized and optimized for definable theories. *Journal of Logic and Computation*. Advance online article, doi:10.1093/logcom/exw025.

Scott, D. (1962). Algebras of sets binumerable in complete extensions of arithmetic. In Dekker, J. C. E., editor, *Recursive Function Theory*, volume 5 of *Proceedings of Symposia in Pure Mathematics*, pages 117–121. American Mathematical Society.

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

Simpson, S. G. (1999). *Subsystems of Second Order Arithmetic*. Springer-Verlag.

Sjögren, J. (2008). On explicating the concept *the power of an arithmetical theory*. *Journal of Philosophical Logic*, 37:183–202.

Skolem, T. (1920). Logisch-kombinatorische Untersuchungen über die Erfüllbarkeit oder Beweisbarkeit mathematischer Sätze nebst einem Theoreme über dichte Mengen. *Videnskapsselskapets Skrifter, I. Matematisk-naturvidenskabelig Klasse*, 4:1–36.

Smoryński, C. (1984). Lectures on nonstandard models of arithmetic. In Lolli, G., Longo, G., and Marcja, A., editors, *Logic Colloquium '82*, volume 112 of *Studies in Logic and the Foundation of Mathematics*, pages 1–70. North-Holland.

Smoryński, C. (1985). *Self-Reference and Modal Logic*. Springer-Verlag.

Smoryński, C. (1987). Quantified modal logic and self-reference. *Notre Dame Journal of Formal Logic*, 28(3):356–370.

Smullyan, R. M. (1961). *Theory of Formal Systems*, volume 47 of *Annals of Mathematics Studies*. Princeton University Press.

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

Sommaruga-Rosolemos, G. (1991). *Fixed Point Constructions in Various Theories of Mathematical Logic*. Bibliopolis.

Tarski, A. (1933). The concept of truth in formalized languages. In Corcoran, J., editor, *Logic, Semantics, Metamathematics*, pages 152–278. Hackett Publishing Company.

Tarski, A., Mostowski, A., and Robinson, R. M. (1953). *Undecidable Theories*. North-Holland.

Verbrugge, R. and Visser, A. (1994). A small reflection principle for bounded arithmetic. *The Journal of Symbolic Logic*, 59(3):785–812.

Visser, A. (1980). Numerations, λ-calculus & arithmetic. In Seldin, J. P. and Hindley, J. R., editors, *To H. B. Curry: Essays on Combinatory Logic, Lambda Calculus and Formalism*, pages 259–284. Academic Press.

Visser, A. (1990). Interpretability logic. In Petkov, P., editor, *Mathematical Logic*, pages 175–209. Plenum Press.

Wilkie, A. J. (1977). On the theories of end-extensions of models of arithmetic. In *Set Theory and Hierarchy Theory*, volume 619 of *Lecture Notes in Mathematics*, pages 305–310. Springer-Verlag.

Wong, T. L. (2016). Interpreting weak König's lemma using the arithmetized completeness theorem. *Proceedings of the American Mathematical Society*, 144(9):4021–4024.

Woodin, W. H. (2011). A potential subtlety concerning the distinction between determininsm and nondeterminism. In Heller and Woodin, editors, *Infinity, New Research Frontiers*, pages 119–129. Cambridge University Press.

Sammanfattning

1 Inledning

En viktig insikt som den matematiska logiken gör är att sanning och bevisbarhet är komplicerade begrepp. Den här avhandlingen bidrar till studiet av den invecklade relationen mellan sanning och bevisbarhet i sådana formella teorier som lämpar sig för att beskriva de naturliga talen 0, 1, 2, ….

Det mest inflytelserika tekniska resultatet som beskriver denna relation är Gödels första ofullständighetssats: Välj ett motsägelsefritt formellt system som är sådant att det finns en mekanisk metod för att avgöra om en given sats är ett axiom i systemet eller ej. Om detta system är tillräckligt starkt för att kunna uttrycka en viss del av den elementära aritmetiken, så är det också tillräckligt starkt för att kunna konstruera en sats som handlar om tal, som är sann men omöjlig att bevisa inom systemet (Gödel, 1931).

Avhandlingen behandlar flera typer av generaliseringar av ofullständighetssatserna, främst genom studiet av så kallade oberoende och flexibla formler. Det första temat är att försöka besvara frågan om vilka teorier som kan bevisa vilka ofullständighetsresultat; detta är ett bidrag till studiet av *svaga aritmetiska teorier* där Hájek och Pudlák (1993) är ett nyckelverk. Ett annat tema rör möjligheten att generalisera ofullständighetsresultat till teorier som inte är rekursivt enumerabla. Liknande frågor har behandlats av till exempel Jeroslow (1975); Kaså (2012); Salehi och Seraji (2016). Det sista temat rör Fefermans (1960) generalisering av Gödels andra ofullständighetssats. Fefermans resultat säger att påståendet ”teorin T är motsägelsefri” är interpreterbart (i en teknisk mening) i T, och här görs försök att bevisa liknande resultat för oberoende och flexibla formler.

2 Bakgrund

I detta kapitel presenteras nödvändigt bakgrundsmaterial. Läsaren förutsätts vara bekant med första ordningens logik, teorierna Q (Robinsons aritmetik) och PA (Peanos aritmetik), den grundläggande teorin för rekursiva funktioner, och naiv mängdteori.

3 Fixpunktsmängder

Fixpunktssatsen och dess variationer används ofta för att konstruera ”självrefererande” satser i form av fixpunkter: ϕ är en fixpunkt till $\theta(x)$ i T omm $\mathrm{T} \vdash \phi \leftrightarrow \theta(\ulcorner \phi \urcorner)$. Flera klassiska resultat kan bevisas med hjälp av detta verktyg, däribland Gödels första ofullständighetssats, Tarskis resultat att aritmetisk sanning inte är aritmetiskt definierbar, samt Löbs sats. Lindström (2003) ger en bild av hur mångsidig tekniken är. I detta kapitel studeras fixpunkter från ett annat perspektiv: givet en $\mathscr{L}_A$-formel $\theta(x)$, vad kan vi säga om mängden av fixpunkter till $\theta(x)$?

Det viktigaste resultatet är att varje fixpunktsmängd är *kreativ* i Rogers (1967) tekniska bemärkelse, och att detta kan generaliseras till de begränsade fixpunktsmängder som är disjunkta från någon annan sådan fixpunktsmängd. Detta resultat ger en marginell förstärkning av ett reslutat av Halbach och Visser (2014). Bidrag görs också till studiet av den algebraiska struktur som erhålls när rekursiva, begränsade fixpunktsmängder ordnas under mängdinklusion.

4 Flexibilitet i fragment

Detta kapitel fyller flera syften. Först introduceras de centrala begreppen oberoende och flexibla formler, och därefter ges en litteraturöversikt från fältets begynnelse under tidigt 1960-tal fram till 2016. Ett annat syfte är att framhålla Kripkes metod (1962) som överlägsen i att konstruera oberoende och flexibla formler. Därutöver tas tillfället att relatera de klassiska resultaten till det modernare studiet av svaga aritmetiker, genom att uppskatta hur mycket matematisk induktion som behövs för att bevisa dessa olika resultat.

5 Formalisering och ändextensioner

Resultaten i kapitel 4 garanterar, tillsammans med fullständighetssatsen för första ordningens logik, existensen av en rekursiv funktion f och en siffra e som är sådana att det för varje $k \in \omega$ finns en modell som satisfierar $\mathrm{T} + f(e) = k$. Varje sådan modell är trivialt en ändextension av standardmodellen. Frågan som behandlas här är om detta är en specialegenskap hos standardmodellen, eller om det finns andra strukturer som har liknande ändextensioner. Det viktigaste resultatet i detta kapitel är att varje modell till $\mathrm{T} + \mathrm{Con}_{\mathrm{T}}$ har sådana ändextensioner, och bevismetoderna kan finslipas för att bevisa starkare resultat av liknande slag.

6 Karaktäriseringar av partiell konservativitet

Det är känt sedan det sena 1970-talet att ändextensioner av modeller till aritmetik spelar en viktig roll i karaktäriseringar av interpreterbarhet och partiell konservativitet, som båda är användbara begrepp för att jämföra formella teoriers relativa styrka. Här generaliseras den så kallade OHGL-karaktäriseringen (efter Orey, Hájek, Guaspari och Lindström) på tre sätt. Karaktäriseringar ges av partiell konservativitet över svaga teorier, teorier formulerade i utökade språk, samt teorier som inte är rekursivt enumerabla.

7 Uniformt flexibla formler och Solovayfunktioner

I kapitel 5 fastställs att det för varje $n > 0$ finns en formel $\gamma(x) \in \Sigma_n$, som är sådan att för varje $\sigma(x) \in \Sigma_n$ och varje modell $\mathcal{M} \models \mathrm{T} + \mathrm{Con}_{\mathrm{T}}$ så finns en ändextension av $\mathcal{M}$ som satisfierar $\mathrm{T} + \forall x(\gamma(x) \leftrightarrow \sigma(x))$. Frågan som behandlas i det här kapitlet är om antagandet att T satisfierar $\mathrm{Con}_{\mathrm{T}}$ kan strykas. Om detta är möjligt så ger OHGL-karaktäriseringen upphov till en interpretation av $\mathrm{T} + \forall x(\gamma(x) \leftrightarrow \sigma(x))$ i T, vilket skulle ge en förstärkning av den generalisering av Gödels andra ofullständighetssats som härrör från Feferman (1960). Här bevisas, med hjälp av så kallade Solovayfunktioner, att antagandet kan strykas om $n > 1$. Det mest intressanta fallet är dock just $n = 1$ och i detta fall ges partiella resultat bland annat genom att generalisera ett resultat från Woodin (2011).