Title

Reflection Principles and Ordinal Analysis

Author

Walsh, James

Publication Date

2020

Peer reviewed|Thesis/dissertation

Reflection Principles and Ordinal Analysis

by

James M Walsh

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Logic and the Methodology of Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Paolo Mancosu, Co-chair
Professor Antonio Montalbán, Co-chair
Associate Professor Wes Holliday
Assistant Professor Thomas Icard
Professor Thomas Scanlon

Spring 2020

Reflection Principles and Ordinal Analysis

Abstract

Reflection Principles and Ordinal Analysis

by

James M Walsh

Doctor of Philosophy in Logic and the Methodology of Science

University of California, Berkeley

Professor Paolo Mancosu, Co-chair

Professor Antonio Montalbán, Co-chair

It is a well-known empirical phenomenon that natural axiomatic theories are pre-well-ordered by proof-theoretic strength. However, without a precise definition of "natural," it is not clear how to explain this observation in a strictly mathematical way. One expression of the pre-well-ordering phenomenon comes from ordinal analysis, a research program wherein ordinals are assigned to natural theories; these ordinals are commonly understood as measuring the "strength" of the theories to which they are assigned. Our goal in this dissertation is to formulate and prove statements that provide insight into the pre-well-ordering phenomenon.

Part I contains an investigation of the pre-well-ordering phenomenon in the setting of first-order arithmetic. One method for determining the proof-theoretic ordinal of a theory T involves showing that T can be conservatively approximated by iterating consistency statements over a weak base theory. This technique yields informative results only for natural theories. Why can natural theories be conservatively approximated by iterating consistency statements? Our explanation is that the consistency operator and its iterates into the effective transfinite are canonical as functions on theories. To this end, we prove that any recursive monotone function on finitely axiomatized theories that is everywhere bounded by an iterate of the consistency operator must coincide somewhere with some iterate of the consistency operator. We also prove that any recursive monotone function on finitely axiomatized theories must be either as weak as the identity operator in the limit or as strong as the consistency operator in the limit.

Part II contains an investigation of the pre-well-ordering phenomenon in the setting of second-order arithmetic. In second-order arithmetic it is possible to formalize Π^1_1 reflection, which provides a more robust measurement of proof-theoretic strength than consistency strength. We prove that the behavior exhibited by natural theories with respect to consistency strength is exhibited in general with respect to Π^1_1 strength in the following sense: there is no sequence $(T_n)_{n<\omega}$ of Π^1_1 sound extensions of ACA_0 such that, for each n, T_n proves the Π^1_1 soundness of T_{n+1}. This means that it is possible to rank the Π^1_1 sound extensions of ACA_0 by proof-theoretic strength. We prove that for every Π^1_1 sound extension T of ACA_0^+

the rank of T is exactly the proof-theoretic ordinal of T; this vindicates the common wisdom that ordinal analysis measures the proof-theoretic strength of natural theories. Along the way we prove two far-reaching conservation theorems. First, α iterated Π^1_1 reflection of ACA_0 is $\Pi^1_1(\Pi^0_3)$ conservative over ε_α iterated $\Pi^1_1(\Pi^0_3)$ reflection of RCA_0. Second, for any Π^1_2 axiomatizable T, every set is contained in an ω model if and only if, for every well order α, α iterated Π^1_1 reflection for T is Π^1_1 sound. Finally, we use the techniques introduced to prove these results to provide a purely proof-theoretic demonstration of a classical theorem of recursion theory: there is no sequence $(A_n)_{n<\omega}$ such that, for each n, $\mathcal{O}^{A_{n+1}} \leqslant_H A_n$. This engenders a well-founded ranking of the real numbers; we prove for almost all reals A, the rank of A is ω_1^A. These results strengthen the analogy between the pre-well-ordering of natural theories by proof-theoretic strength in proof theory and the pre-well-ordering of reals under descriptive complexity.

Part III contains reflections on foundational issues. The main foundational applications of ordinal analysis to date are in Hilbert's program and modifications thereof. Instead, I emphasize the *explanatory* value of ordinal-theoretic consistency proofs over their *justificatory* value. Whether, for instance, Gentzen's consistency proof convinces us *that* arithmetic is consistent, it explains *why* arithmetic is consistent. On this basis, I argue against certain restrictive norms on mathematical practice. For instance, I argue against the norm that consistency proofs ought to be carried out using only constructive methods. Though this restriction is reasonable if one is attempting to convince a constructivist skeptic, it loses force when one instead seeks a combinatorial explanation of a theory's consistency.

Contents

Acknowledgments

Thanks to Antonio Montalbán for encouraging me to develop my research in directions that excited me; his suggestions and corrections have improved nearly every page of this document. Thanks to Paolo Mancosu for his professional support, mentorship, and teaching; I learned much in his seminars and his comments strengthened Chapter 7. Thanks to Wes Holliday and Thomas Icard for extensive intellectual and professional guidance and for all their work on behalf of the logic community in the Bay Area. Thanks to Tom Scanlon for joining my dissertation committee. Thanks also to John MacFarlane for preparing me for my qualifying exam.

Thanks to Vicky Lee, Jon Phillips, and Barb Waller for extensive administrative support.

I am grateful to my fellow students for their friendship and for creating an inspiring atmosphere. Thanks especially to Raha Ahmadianhosseini, Dan Appel, Doug Blue, Madeline Brandt, Reid Dale, Liz Ferme, Matthew Harrison-Trainor, Mikayla Kelley, Arc Kocurek, Richard Lawrence, Pen Long, Patrick Lutz, Matthew McCauley, Sven Neth, Adele Padgett, Nick Ramsey, Ed Scerbo, Pia Schneider, Noah Schweber, Anna Seigal, Benny Siskind, and Kentarô Yamamoto.

Thanks to Lev Beklemishev and Fedor Pakhomov for their hospitality in Russia.

Thanks to my co-authors Patrick Lutz, Antonio Montalbán, and Fedor Pakhomov for permitting me to include our joint work in this document. It was a pleasure to work with them all and I hope to do so again in the future.

Thanks to my trivia teammates for injecting levity into Sunday nights.

Most of all, thanks to my parents, Doreen, Tommy, Michael, and Donna.

Chapter 1

Introduction

1.1 The consistency strength hierarchy

This dissertation concerns the scope of the *axiomatic method*, a method whereby disciplines are organized around basic principles known as *axioms*. In mathematical logic the notions of axiom and proof are formalized so that they can be studied by mathematical means. Since the advent of modern mathematical logic, theoretical understanding of the axiomatic method has progressed significantly. Gödel's discovery of the incompleteness theorems marked an important moment in this development. Gödel's theorems establish fundamental restrictions on what can be proven on the basis of any reasonable axiomatic theories. By a *reasonable* axiomatic theory I mean a consistent, recursively axiomatized theory that interprets a modicum of arithmetic.

Theorem 1.1.1 (Gödel)**.** *No reasonable axiomatic theory is complete.*

Theorem 1.1.2 (Gödel)**.** *No reasonable axiomatic theory proves its own consistency.*

The first incompleteness theorem shows that, no matter what principles we endorse, there will be questions that are not resolvable on the basis of those principles. This means that there is no universal axiom system within which mathematics can be developed. Instead, we are left with a vast array of axiomatic theories. The second incompleteness theorem yields the primary method for mapping out this vast array of theories. For reasonable theories T and U, we say that $T \leqslant_{\mathsf{Con}} U$ if the consistency of U implies the consistency of T over a suitable base theory. As usual, $T \equiv_{\mathsf{Con}} U$ if $T \leqslant_{\mathsf{Con}} U$ and $U \leqslant_{\mathsf{Con}} T$. Also as usual, $T <_{\mathsf{Con}} U$ if $T \leqslant_{\mathsf{Con}} U$ and $U \nleqslant_{\mathsf{Con}} T$. The notion of strength that $<_{\mathsf{Con}}$ engenders is known as *consistency strength*.

The structure of *all* reasonable theories ordered by consistency strength is quite complicated. As far as I know, the following theorems are folklore:

Theorem 1.1.3 (Folklore)**.** *The ordering* $<_{\mathsf{Con}}$ *is not* pre-linear*, i.e., there are theories* T *and* U *such that* $T \not\equiv_{\mathsf{Con}} U$*,* $T \not<_{\mathsf{Con}} U$*, and* $U \not<_{\mathsf{Con}} T$*.*

Theorem 1.1.4 (Folklore)**.** *The ordering* $<_{\mathsf{Con}}$ *is* ill-founded*, i.e., there is a sequence* $T_0 >_{\mathsf{Con}} T_1 >_{\mathsf{Con}} T_2 >_{\mathsf{Con}} \ldots$ *where each* T_i *is consistent.*

All known instances of non-linearity and ill-foundedness have been discovered by defining theories in an ad-hoc manner using self-reference and other logical tricks. When one restricts one's attention to the *natural* axiomatic theories—those that arise in practice—ordered by $<_{\mathsf{Con}}$, the resulting structure is a *pre-well-ordering.*[1] Here is a tiny snapshot of such theories:

$$\mathsf{EA}, \mathsf{EA}^+, \mathsf{PRA}, I\Sigma_n, \mathsf{PA}, \mathsf{ATR}_0, \Pi^1_n\mathsf{CA}_0, \mathsf{PA}_n, \mathsf{ZF}, \mathsf{AD}^{L(\mathbb{R})}$$

The well-ordering phenomenon persists, taking a very liberal view of what constitutes a "natural" theory. Note that the theories just cited run the gamut from weak fragments of arithmetic to subsystems of analysis and all the way to strong extensions of set theory. These theories come from different areas of mathematics (e.g., arithmetic, analysis, set theory) and often codify different conceptions of mathematics. Indeed, many of the natural extensions of set theory that have been investigated are jointly inconsistent, yet comparable according to consistency strength.

Explaining the contrast between natural axiomatic theories and axiomatic theories in general is widely regarded as a major outstanding conceptual problem in mathematical logic. The following passage from a paper of S. Friedman, Rathjen, and Weiermann is representative:

> The fact that "natural" theories, i.e. theories which have something like an "idea" to them, are almost always linearly ordered with regard to logical strength has been called one of the great mysteries of the foundations of mathematics. (S. Friedman, Rathjen, Weiermann, [33], p. 382)

If it is true that natural axiomatic theories are pre-well-ordered by consistency strength, and not an illusion engendered by a paucity of examples, then one would like to prove that it is true. However, the claim that natural axiomatic theories are pre-well-ordered by consistency strength is not a strictly mathematical claim. The problem is that we lack a definition of "natural axiomatic theory." Without a precise definition of "natural," it is not clear how to prove this claim, or even how to state it mathematically.

1.2 The search for new axioms

I would like to consider how the well-ordering phenomenon plays out in practice, namely, in set theory. Indeed, the well-ordering phenomenon has been a focal point of interest in the search for new axioms extending the standard ZFC axioms for set theory.

[1] This is to say that the induced ordering on the equivalence classes modulo equiconsistency is a well-ordering.

Set theory has been developed in an explicitly axiomatic fashion, at least since Zermelo isolated the axioms that are the precursors to the standard ZFC axioms. The early progenitors of set theory discovered that various mathematical structures could be realized in set-theoretic terms and that their properties could be established on the basis of set-theoretic reasoning. Thus, set theory constitutes a highly general mathematical framework. On the other hand, many of the central problems of set theory (such as the Continuum Hypothesis, the Projective Measure problem, and Suslin's Hypothesis) *cannot* be resolved on the basis of the ZFC axioms. This has motivated the search for new axioms for set theory.

Set theorists have investigated many extensions of ZFC, including large cardinal axioms, determinacy axioms, forcing axioms, and more. Is it possible to make rational judgments about these principles and their consequences? Maddy has promoted the maxim "maximize interpretability strength" as a guiding principle in the search for new axioms (see, e.g, [61, 62]). Let's call this principle MADDY'S MAXIM. For a large swathe of theories, including extensions of set theory, maximizing interpretability strength coincides with maximizing consistency strength. For a sentence φ that is independent of ZFC, one can imagine the following four possibilities:

1. φ increases interpretability strength but $\neg\varphi$ does not.
2. $\neg\varphi$ increases interpretability strength but φ does not.
3. Neither φ nor $\neg\varphi$ increases interpretability strength.
4. Both φ and $\neg\varphi$ increase interpretability strength.

It turns out that all four possibilities are realized; in the fourth case we *cannot* follow MADDY'S MAXIM or we would land ourselves in inconsistency. However, this is not typically taken as a refutation of MADDY'S MAXIM. The key point here is that when one restricts oneself to *natural* theories, only the first three possibilities are realized, since natural theories are linearly ordered by consistency strength. See [51] for a discussion of this point.

Consider once more the variety of extensions of ZFC that are investigated by set theorists: large cardinal axioms, axioms of definable determinacy, forcing axioms, and more. These axiom systems, which have different motivations and often codify different foundational conceptions of mathematics, are pre-well-ordered by consistency strength. Indeed, they are pre-well-ordered according to all the common notions of proof-theoretic strength (such as 1-consistency strength, Π^1_1 reflection strength, and so on). A consequence is that these axiom systems all converge on arithmetic statements and even analytic statements. As Steel writes:

> Thus the well-ordering of natural consistency strengths corresponds to a well-ordering by inclusion of theories of the natural numbers. There is no divergence at the arithmetic level, if one climbs the consistency strength hierarchy in any natural way we know of... Natural ways of climbing the consistency strength hierarchy do not diverge in their consequences for the reals.... Let T, U be natural

> theories of consistency strength at least that of "there are infinitely many Woodin cardinals"; then either $(\Pi^1_\omega)_T \subseteq (\Pi^1_\omega)_U$ or $(\Pi^1_\omega)_U \subseteq (\Pi^1_\omega)_T$. (Steel, [101], p. 159)

That is, at the level of statements about $\mathbb{R}$, all natural theories converge; MADDY'S MAXIM and its variants suggest that we ought to endorse the sentences on which they converge. Thus, the apparent pre-well-ordering of theories by consistency strength and other notions of proof-theoretic strength plays a central role in the search for new axioms in set theory.

1.3 Ordinal analysis

At this point it will be helpful to discuss two research programs that deal with questions of systematically reducing incompleteness. One is Gentzen's program of ordinal analysis and the other is Turing's program of completeness via iterated reflection principles. Both programs can be understood as reactions to the incompleteness theorems. The practitioners of ordinal analysis have attempted to reduce incompleteness by proving consistency statements in a systematic way, namely, from ever stronger transfinite induction principles. Turing's program, on the other hand, uses the second incompleteness theorem as the major engine for overcoming incompleteness; the practitioners of Turing's program try to systematically effect reductions in incompleteness by successively adding consistency statements to theories as new axioms. Both programs will play a major role in the approach to the "great mystery" advanced in this dissertation.

Ordinal analysis was developed in the context of *Hilbert's Program*, an early twentieth century research program pioneered by David Hilbert. To combat skepticism about the cogency of infinitary mathematics, Hilbert proposed to (i) axiomatize infinitary mathematics and (ii) prove the consistency of the axioms by finitary means. In 1931, Hilbert's program reached a major obstacle in the form of Gödel's [40] second incompleteness theorem. Indeed, it follows from Gödel's theorem that if the principles of finitistic mathematics are codifiable in a reasonable axiomatic theory, then they do not prove their own consistency, much less the consistency of stronger theories. Thus, it is generally agreed that Hilbert's program failed.

Gentzen was apparently undeterred, however. Not long after Gödel proved the incompleteness theorems, Gentzen [35] produced a consistency proof of arithmetic.

Theorem 1.3.1 (Gentzen)**.** *If ε_0 is well-founded, then arithmetic is consistent.*

The only principle invoked in Gentzen's proof that is not obviously finitistically acceptable is the well-foundedness of the ordinal number ε_0. Gentzen's consistency proof marked the beginning of a research program known as *ordinal analysis*, whereby similar consistency proofs have been discovered for a wide range of axiomatic theories. Developing such consistency proofs for a theory T involves, among other things, determining the *proof-theoretic ordinal* of T. Informally, the proof-theoretic ordinal of T is the least ordinal α such that induction along α suffices to prove the consistency of T. Making this informal definition precise is no easy task, as we shall see.

The methods of ordinal analysis have been used to analyze many theories of interest, including subsystems of first-order arithmetic, second-order arithmetic, and set theory. Ordinal analysis has not yet reached the level of full second-order arithmetic. Nevertheless, it is reasonable to expect that the existing results are part of a general connection between ordinals and consistency. Do Gentzen-style methods suffice for proving the consistency of any axiomatic theory? The following result [55] might *seem* to suggest a positive answer ($\mathsf{TI}^{\prec}_{\Pi_1}$ is a sentence expressing the validity of induction for Π_1 predicates along $\prec$):

Theorem 1.3.2 (Kreisel–Shoenfield–Wang)**.** *For any reasonable theory T, there is a presentation $\prec$ of a recursive ordinal such that* $\mathsf{PRA} + \mathsf{TI}^{\prec}_{\Pi_1} \vdash \mathsf{Con}(T)$.

This theorem is proved by showing that, for any true Π_1 sentence φ, one can encode the truth of φ into an ordinal notation system $\prec$ such that recognizing that $\prec$ has no Π_1 definable descending sequences is equivalent to recognizing the truth of φ. One could not recognize the validity of Π_1 transfinite induction along such a notation system without knowing the truth of φ, so one could not use such a transfinite induction principle to prove φ. Thus, the epistemic value of this theorem is limited.

As we will see, this is one version of a pervasive problem known as the *canonicity problem*. The ordinal notations that are devised to prove Theorem 1.3.2 are not *natural* notations. They form notation systems that one would introduce only in an ad-hoc manner to solve a problem in proof theory. Does the distinction between the pathological notation systems and the natural notation systems reflect some intrinsic mathematical properties of the notation systems? If one had a convincing definition of "natural," one might conjecture that, for any reasonable theory T, there is a *natural* presentation $\prec$ of a recursive ordinal such that $\mathsf{PRA} + \mathsf{TI}^{\prec}_{\Pi_1} \vdash \mathsf{Con}(T)$. However, at present there is no convincing evidence that it is possible to precisely define the "natural" notation systems.

The canonicity problem also makes it difficult to define "proof-theoretic ordinal," as suggested earlier. One might try to define the proof-theoretic ordinal of T as the least ordinal α such that induction along α (along with finitary methods) suffices to prove the consistency of T. The problem with this definition is that, in formalized theories, transfinite induction principles are stated in terms of *presentations* of ordinals, not the ordinals themselves. Kreisel has shown that it is always possible to prove the consistency of a reasonable theory by induction along a sufficiently pathological presentation of ω. Conversely, Beklemishev has shown that, for any reasonable theory T and recursive ordinal α, there is a sufficiently pathological presentation of α such that transfinite induction along that presentation (along with finitary methods) will not suffice to prove the consistency of T. When one restricts one's attention to "natural" presentations of ordinals, the definition seems to work, but the current state of affairs is vexing and unsatisfactory.

At present, many conflicting definitions of the "proof-theoretic ordinal" of a theory have been proposed; these definitions often coincide in crucial cases. Perhaps the most common definition is this: the proof-theoretic ordinal of a theory T is the supremum of the order-types of the primitive recursive well-orderings whose well-foundedness is provable in T. This

is sometimes called the Π^1_1 ordinal of a theory. Of course, this notion is only useful for measuring the proof-theoretic strength of theories in which well-foundedness is expressible, so it does not apply, e.g., to fragments of first-order arithmetic. Moreover, it is a somewhat coarse notion of strength, since Π^1_1 ordinals are invariant under the addition of true Σ^1_1 sentences to the object theory. Nevertheless, the Π^1_1 ordinal of a theory is an *ordinal*, not a presentation of an ordinal, so this is a somewhat robust notion.

1.4 Turing progressions

The second incompleteness theorem suggests a method for dealing with the first incompleteness theorem. According to Turing [109], if we endorse the axioms of a reasonable axiomatic theory T, then there is a principled way of extending T, namely, by adopting T's consistency statement as an axiom. Of course, if one adopts the statement $\mathsf{Con}(T)$ as an axiom, the statement $\mathsf{Con}(T + \mathsf{Con}(T))$ remains unprovable. However, there is a principled way of resolving this problem, namely, by adopting $\mathsf{Con}(T + \mathsf{Con}(T))$ as an axiom. Of course, this engenders a new problem, but it just as easily engenders a new solution, whence this process can iterated *ad infinitum*. Given presentations of recursive ordinals in the language of arithmetic, one can even extend this process into the effective transfinite.

Naively, we might try to define the iterations of consistency over a theory T as follows:

- $T_0 := T$
- $T_{\alpha+1} := T_\alpha + \mathsf{Con}(T_\alpha)$
- $T_\lambda := \bigcup_{\alpha<\lambda} T_\alpha$ for λ a limit.

However, such a definition does not even pin down the theory $T_{\omega+1}$. According to the definition, $T_{\omega+1}$ is just $T_\omega + \mathsf{Con}(T_\omega)$, but for $\mathsf{Con}(T_\omega)$ to be a statement of arithmetic we must have some effective presentation of T_ω, and there are many choices for the latter. Accordingly, when one defines iterated consistency statements, one must first fix an ordinal notation system $\prec$. One can then define the iterations of consistency along $\prec$ within arithmetic via Gödel's fixed point lemma.

Does iteratively endorsing consistency statements in this manner effect a significant reduction in incompleteness?

Theorem 1.4.1 (Turing)**.** *For any true* Π_1 *sentence* φ*, there is a presentation* $\prec$ *of* $\omega + 1$ *such that* $\mathsf{PA}_\prec \vdash \varphi$*.*

This result is known as *Turing's Completeness Theorem*. At first glance it may seem that this theorem is epistemically significant: we can come to know the truth of any Π_1 statement φ simply by iterating consistency statements. How are the consistency statements used in the proof of φ ? The disappointing response is that they are not used at all. Instead, the truth of φ is encoded into a non-standard description of the base theory PA, which becomes

available at iteration $\omega + 1$. Given this non-standard description of PA, discerning what the theory $\mathsf{PA}_\prec$ is committed to *requires* knowing the truth-value of φ. This is another appearance of the canonicity problem. As far as I know, this was the initial appearance of the problem, and Turing was the first to identify it.

Turing was careful not to overstate the epistemic significance of his theorem. Nevertheless, one might hope that iterating Con along *natural* ordinal notations would also suffice to prove any true Π_1 statement. Once again, if we had a precise definition of "natural" ordinal notation systems, we would be able to formulate a precise conjecture; however, no such characterization of "natural" ordinal notation systems is currently available.

One might naively try to use iterated consistency progressions to measure the strength of Π_1 statements: the measure of complexity we assign to a true Π_1 sentence φ is the least ordinal α such that $\mathsf{PA}_\alpha \vdash \varphi$. One could then extend this measure of complexity to theories: the complexity of T is the least ordinal such that $\mathsf{PA}_\alpha \vdash \varphi$ for each Π_1 theorem φ of T. Clearly, Turing's completeness theorem shows that iterations of consistency depend on presentations of ordinals and not just on ordinals themselves. The possibility remains, however, that one could prove informative results of this sort for large swathes of interesting theories by antecedently fixing some natural notation system; Beklemishev has recently pursued this possibility, as we will discuss in §1.5.

Turing's work was later pursued and greatly extended by Feferman [26]. Feferman shifted the focus from iterated consistency to iterated reflection principles of other sorts. Recall that $\mathsf{RFN}(T)$ is the uniform reflection schema for T, that is,

$$\mathsf{RFN}(T) := \{\forall \vec{x}\big(\mathsf{Pr}_T(\varphi(\vec{x})) \to \varphi(\vec{x})\big) : \varphi(\vec{x}) \in \mathcal{L}_A\}.$$

After fixing a presentation $\prec$ of a recursive ordinal, we can then define iterations of reflection over a theory T so as to satisfy the following conditions:

- $T_0^{\mathsf{RFN}} := T$
- $T_{\alpha+1}^{\mathsf{RFN}} := T_\alpha^{\mathsf{RFN}} + \mathsf{RFN}(T_\alpha^{\mathsf{RFN}})$
- $T_\lambda^{\mathsf{RFN}} := \bigcup_{\alpha<\lambda} T_\alpha^{\mathsf{RFN}}$ for λ a limit.

Feferman proved that, merely by iterating the uniform reflection schema along presentations of recursive ordinals, one can accrue resources sufficient, not only for proving any true Π_1 statement, but for proving *any* true arithmetical statement.

Theorem 1.4.2 (Feferman)**.** *For any true arithmetical sentence φ, there is a presentation $\prec$ of an ordinal $\alpha < \omega^{\omega^\omega+1}$ such that $\mathsf{PA}_\prec^{\mathsf{RFN}} \vdash \varphi$.*

This result is known as *Feferman's Completeness Theorem.* In the proof of this theorem, Feferman used Turing's technique of encoding the truth of statements into presentations of recursive ordinals, among other things. Feferman also proved that there are paths P through Kleene's $\mathcal{O}$ such that iterating uniform reflection along P is arithmetically complete.

However, any such path is Π^1_1 complete. That is, identifying such a path is *more difficult* than identifying the set of arithmetical truths. Thus, the epistemic significance of these results—as Feferman emphasized—is limited.

1.5 Ordinal analysis via iterated reflection

In recent years there has been interest in the interface between ordinal analysis and Turing progressions. This research has been motivated by a number of the drawbacks of the research discussed in the previous two sections.

As discussed in §1.3, the standard notion of the Π^1_1 ordinal of a theory is (i) only applicable to theories in which "well-foundedness" is expressible and (ii) insensitive to the true Σ^1_1 consequences of a theory. Building on earlier work of Schmerl [87], Beklemishev [7] introduced the notion of the Π_1 ordinal of a theory; this notion of "ordinal analysis" is both (i) suitable for subsystems of first-order arithmetic and (ii) sensitive to the Π_1 consequences of theories. Iterations of consistency in the style of Turing play a central role in the definition of Π_1 ordinals.

First we fix a base theory: Beklemishev uses EA, a weak subsystem of arithmetic that is just strong enough to carry out arithmetization of syntax in the standard way. We then fix some natural ordinal notation system and define the iterations of EA so as to satisfy the following conditions:

1. $\mathsf{EA}_0 := \mathsf{EA}$
2. $\mathsf{EA}_{\alpha+1} := \mathsf{EA}_\alpha + \mathsf{Con}(\mathsf{EA}_\alpha)$
3. $\mathsf{EA}_\lambda := \bigcup_{\alpha<\lambda} \mathsf{EA}_\alpha$ for λ a limit.

Given a target theory T, the Π_1 ordinal of T (relative to the base theory EA and notation system $\prec$) is defined as follows:

$$|T|_{\Pi_1} := \mathsf{sup}\{\alpha : \mathsf{EA}_\alpha \subseteq T\}$$

This definition yields interesting information only if T is conservatively approximated by iterations of consistency over EA; that is, only if $T \equiv_{\Pi_1} \mathsf{EA}_\alpha$ where $\alpha = |T|_{\Pi_1}$. Knowing that T is so approximated is useful, because there are elegant equations (first discovered by Schmerl) spelling out conservation relations between iterated consistency statements. One can use these equations to calculate the Π_1 proof-theoretic ordinals of theories that are conservatively approximated by iterated consistency statements in a uniform manner. This approach to ordinal analysis will play a major role in this dissertation.

Clearly, Π_1 ordinals are well-defined for theories of first-order arithmetic in which well-foundedness is not directly expressible. The definition is also sensitive to Π_1 sentences. For instance, whereas the Π_1 ordinal of PA is ε_0, the Π_1 ordinal of $\mathsf{PA} + \mathsf{Con}(\mathsf{PA})$ is $\varepsilon_0 \times 2$.

Note that the definition of Π_1 ordinals achieves Turing's goal of providing a metric of strength for theories in terms of the number of iterations of consistency required to capture their Π_1 theorems. It turns out that this metric of strength is the same one provided by ordinal analysis (indeed, the definition of Π_1 ordinals coincides with other standard definitions of proof-theoretic ordinal in standard cases). To provide such a metric of strength, we must fix (in advance) a natural ordinal notation system. Moreover, the definition only works for theories whose Π_1 fragments can be conservatively approximated by iterated consistency statements. This is apparently a feature of "natural theories" but not of all theories. We will explore this theme throughout this dissertation.

In recent years (see [8, 9]) Beklemishev has advocated for an approach to the canonicity problem for ordinal notation systems that makes use of reflection principles. I will briefly describe Beklemishev's notation system for PA and its fragments. The terms are generated by the constant symbol $\top$ and the function symbols $\mathsf{Con}_{\mathsf{EA}}(\cdot)$, $1\mathsf{Con}_{\mathsf{EA}}(\cdot)$, $2\mathsf{Con}_{\mathsf{EA}}(\cdot)$, etc. The ordering on the terms is given, roughly, by consistency strength over EA.[2] This notation system is well-suited to ordinal analysis via iterated reflection, since PA can be conservatively approximated over EA by the terms of the notation system, regarded as first-order theories.

The connection with the canonicity problem is this: If these reflection principles may be regarded as a *canonical means* of specifying the theory PA, then the ordinal notation system (not just the ordinal) has been extracted from a canonical presentation of PA. Various results in this dissertation will lend support to the notion that reflection principles *are* canonical, and to the notion that natural theories can be approximated by iterating reflection principles. However, the results in this dissertation seem too coarse grained to isolate Beklemishev's notation system as canonical in a way that certain notation systems that give "wrong" answers for ordinal analysis are not. Nevertheless, this is a potential theme for future research.

1.6 Turing degree theory

The pre-well-ordering of natural theories is paralleled by a phenomenon in Turing degree theory. The Turing degrees are not linearly ordered by $<_T$. That is, there are distinct degrees a and b such that $a \not\leqslant_T b$ and $b \not\leqslant_T a$. The Turing degrees are also ill-founded. That is, there are infinite sequences $(a_n)_{n<\omega}$ such that $a_k >_T a_{k+1}$ for all k. These two results mean that is neither possible to compare nor to rank Turing degrees in general.

The degrees that exhibit these pathological properties have been constructed using ad-hoc recursion-theoretic techniques, like the priority method. When one restricts one's attention to *natural* Turing degrees, the resulting structure is a well-order.

$$0, 0', 0'', ..., 0^{\omega}, ..., \mathcal{O}, ..., 0^{\sharp}, ...$$

[2] In fact, Beklemishev uses the ordering $\prec$ defined as follows: $\alpha \prec \beta$ if $\mathsf{GLP} \vdash \beta \rightarrow \Diamond_0 \alpha$, where GLP is a certain polymodal logic corresponding to provability over EA.

This state of affairs is remarkably similar to the state of affairs on the proof-theoretic side. Once again, it is not entirely clear how to address this problem. It is not a wholly mathematical problem since there is no precise mathematical definition of "natural" Turing degrees.

One oft-noted feature of natural Turing degrees is that their definitions *relativize*. Relativizing the definition of a Turing degree yields a degree invariant function on the reals, where a function f is degree invariant if $A \equiv_T B$ implies $f(A) \equiv_T f(B)$. For instance, relativizing the definition of $0'$ yields the Turing jump $\lambda X.X'$, relativizing the definition of Kleene's $\mathcal{O}$ yields the hyperjump $\lambda X.\mathcal{O}^X$, and so on.

Martin proposed a classification of the degree invariant functions in terms of their behavior *almost everywhere* in the sense of Martin Measure. Recall that a *cone* in the Turing degrees is any set of the form $\{a : a \geqslant_T b\}$. Assuming AD, Martin proved that every degree invariant set of reals either contains a cone or is disjoint from a cone. Moreover the intersection of countably many cones contains a cone. Thus, assuming AD, the function

$$\mu(A) = \begin{cases} 1 & \text{if } A \text{ contains a cone} \\ 0 & \text{if } A \text{ is disjoint from a cone} \end{cases}$$

is a countably additive measure on the σ algebra of degree invariant sets. This measure is called *Martin Measure*. In the statement of Martin's Conjecture, *almost everywhere* means almost everywhere with respect to Martin Measure.

Conjecture 1.6.1 (Martin)**.** *Assume* $\mathsf{ZF} + \mathsf{DC} + \mathsf{AD}$*. Then*

I. If $f : 2^\omega \to 2^\omega$ is degree invariant, and f is not increasing a.e. then f is constant a.e.

II. $\leqslant_m$ pre-well-orders the set of degree invariant functions that are increasing a.e. If f has $\leqslant_m$ rank α, then f' has $\leqslant_m$ rank $\alpha + 1$, where $f'(x) = f(x)'$ for all x.

One can view Martin's Conjecture as a conjecture about the functions in $L(\mathbb{R})$, since, assuming appropriate large cardinal axioms, $L(\mathbb{R})$ satisfies AD. Thus, Martin's Conjecture roughly states that the only definable degree invariant functions (up to almost everywhere equivalence) are constant functions, the identify function, and iterates of the Turing Jump.

Though Martin's Conjecture is presently open, many informative partial results and special cases are known (see [100, 94]). We remind the reader that a function f is *uniformly degree invariant* if there is a recursive ϕ such that if $A \leqslant_T B$ via e then $f(A) \leqslant_T f(B)$ via $\phi(e)$. A function f is *order-preserving* if $A \leqslant_T B$ implies $f(A) \leqslant_T f(B)$. Finally, a function f is increasing if, for all A, $A \leqslant_T f(A)$.

Theorem 1.6.2 (Slaman–Steel)**.** *Part I of Martin's Conjecture holds for all uniformly degree invariant functions.*

Theorem 1.6.3 (Steel)**.** *Part II of Martin's Conjecture holds for all uniformly degree invariant functions.*

Theorem 1.6.4 (Slaman–Steel)**.** *If f is a Borel order preserving function that is increasing a.e., then there exists an $\alpha < \omega_1$ such that $f(x) \equiv_T x^{(\alpha)}$ a.e.*

There are other results that speak to this phenomenon. For instance, Steel [98] has shown that there cannot be any "simple" descending sequences in the Turing jump hierarchy, in the following sense:

Theorem 1.6.5 (Steel)**.** *Let $P \subset \mathbb{R}^2$ be arithmetic. Then there is no sequence $(x_n)_{n<\omega}$ such that for every n,*

(i) $x_n \geqslant_T x'_{n+1}$ *and*

(ii) x_{n+1} *is the unique y such that $P(x_n, y)$.*

We will explore the analogy between the recursion-theoretic and proof-theoretic well-ordering phenomena throughout this dissertation. On the one hand, the statements and proofs of many purely proof-theoretic results are inspired by this recursion-theoretic research. On the other hand, the proof-theoretic theorems we produce inspire some purely recursion-theoretic results in Chapter 6.

1.7 Summary of upcoming chapters

My goal is to bring precision to the question of the pre-well-orderedness of natural theories and to offer (at least partial) solutions. Before summarizing the chapters to come, I would like to mention three themes that will be interwoven.

The first theme is that many natural theories can be axiomatized by reflection principles over natural base theories. Indeed, the fragments of natural theories corresponding to different syntactic complexity classes can often be conservatively approximated by iterated reflection principles of the appropriate complexity class.

The second theme is that ordinal analysis is an expression of the well-ordering phenomenon. For different notions of "proof-theoretic strength," there are corresponding notions of "proof-theoretic ordinals." Insofar as the proof-theoretic ordinal afforded by some definition of "proof-theoretic ordinal" measures the proof-theoretic strength of theories, the attendant method of ordinal analysis well-orders the theories within its ken according to that notion of strength. The heuristic that we will try to vindicate in this dissertation is the following: to calculate the proof-theoretic ordinal of a theory T is to determine T's rank in the hierarchy of natural theories ordered by proof-theoretic strength.

The third theme is that reflection principles play the same role in proof theory that jumps play in recursion theory. Just as natural Turing degrees are apparently equivalent to ordinal iterates of the Turing jump, natural theories are apparently equivalent to ordinal iterates (along natural presentations of well-orderings) of reflection principles. Just as natural Turing degrees can be obliquely studied in terms of jumps, natural axiomatic theories can be obliquely studied in terms of reflection principles.

On the inevitability of the consistency operator

In Chapter 2, co-authored with Antonio Montalbán, we present an approach to the well-ordering of natural theories that is inspired by Martin's Conjecture. In particular, we shift our focus from theories to recursive monotone functions on theories. A function $\mathfrak{g}$ on finitely axiomatized extensions of T is *monotone* if whenever $T \vdash \varphi \to \psi$, then also $T \vdash \mathfrak{g}(\varphi) \to \mathfrak{g}(\psi)$. Whereas the notion of a "natural" theory is informal and non-mathematical, the notion of a recursive monotone function is perfectly mathematical.

One recursive monotone function is the consistency operator $\varphi \mapsto \mathsf{Con}_T(\varphi)$. Our goal is to show that the consistency operator and its iterates into the effective transfinite are canonical, by analogy with Martin's Conjecture for order-preserving functions. Our first step in this direction is the following theorem. Note that we write $[\varphi]_T$ to denote the equivalence class of φ modulo T provable equivalence, i.e., $[\varphi]_T = \{\psi : T \vdash \varphi \leftrightarrow \psi\}$.

Theorem 1.7.1. *Let T be a sound, recursively axiomatized extension of elementary arithmetic. Let $\mathfrak{g}$ be recursive and monotone. Suppose that for all consistent φ,*

1. *$T + \varphi + \mathsf{Con}_T(\varphi) \vdash \mathfrak{g}(\varphi)$*
2. *$T + \mathfrak{g}(\varphi) \vdash \varphi$ and $T + \varphi \nvdash \mathfrak{g}(\varphi)$.*

Then for every true φ, there is a true ψ such that $T + \psi \vdash \varphi$ and $[\mathfrak{g}(\psi)]_T = [\psi \wedge \mathsf{Con}_T(\psi)]_T$.

We also generalize this theorem into the effective transfinite. In particular, fixing a nice ordinal notation system, we show that for any recursive monotone $\mathfrak{g}$, for *arbitrarily strong* true inputs, $\mathfrak{g}$ must be either as strong as Con_T^α or as weak as Con_T^β for some $\beta \prec \alpha$.

The main theorem of Chapter 2 states that if the strength of any sufficiently nice function $\mathfrak{g}$ is "bounded" by some iterate of the consistency operator, then $\mathfrak{g}$ must somewhere coincide with an iterate of the consistency operator.

Theorem 1.7.2. *Let T be a sound, recursively axiomatized extension of elementary arithmetic. Let $\mathfrak{g}$ be recursive and monotone such that, for all φ, $\mathfrak{g}(\varphi)$ is Π_1. Then one of the following holds:*

1. *For some φ, $T + \varphi + \mathsf{Con}_T^\alpha(\varphi)$ does not prove $\mathfrak{g}(\varphi)$.*
2. *For some $\beta \preceq \alpha$ and some φ, $[\varphi + \mathfrak{g}(\varphi)]_T = [\varphi + \mathsf{Con}_T^\beta(\varphi)]_T \neq [\bot]_T$.*

A note on the consistency operator

In Chapter 3 we extend the results of the previous chapter. We begin by formulating a conjecture, according to which recursive monotone functions are linearly ordered according to the amount of strength that they systematically add to theories. This conjecture is stated in terms of recursive functions that produce only Π_1 sentences. This results in no loss of generality, since the consistency strength of a theory is determined by its Π_1 consequences. We

conjecture that any monotone algorithm that adds only Π_1 sentences to theories is equivalent, in the limit, to some (potentially non-standard) iterate of the consistency operator.

We then prove the first case of this hoped for classification. In particular, we show that any monotone algorithm that adds only Π_1 sentences to theories is either as weak as the identity operator or as strong as the consistency operator in the limit.

Theorem 1.7.3. *Let T be a sound, recursively axiomatized extension of elementary arithmetic. Let $\mathfrak{g}$ be recursive and monotone such that, for all φ, $\mathfrak{g}(\varphi)$ is Π_1. Then one of the following holds:*

1. *There is a true ψ such that, for all φ for which $T + \varphi$ proves ψ,*

$$T + \varphi \vdash \mathfrak{g}(\varphi).$$

2. *There is a true ψ such that, for all φ for which $T + \varphi$ proves ψ,*

$$T + \varphi + \mathfrak{g}(\varphi) \vdash \mathsf{Con}_T(\varphi).$$

We also established some limitative results on the scope of this approach. In particular, we show that the assumption that $\mathfrak{g}$ is recursive is necessary in the statement of Theorem 1.7.3. We do this by exhibiting a function $\mathfrak{g}$ that is limit-recursive but not recursive which meets the other hypotheses of Theorem 1.7.3 but does not satisfy the conclusion. In particular, $\mathfrak{g}$ vacillates between behaving like the identity operator and the consistency operator, without converging on either.

Theorem 1.7.4. *Let T be a sound, recursively axiomatized extension of elementary arithmetic. There is a limit-recursive monotone function $\mathfrak{g}$ such that, for every φ, $\mathfrak{g}(\varphi)$ is Π_1, yet for arbitrarily strong true sentences*

$$[\varphi + \mathfrak{g}(\varphi)]_T = [\varphi + \mathsf{Con}_T(\varphi)]_T$$

and for arbitrarily strong true sentences

$$[\varphi + \mathfrak{g}(\varphi)]_T = [\varphi]_T.$$

Reflection ranks and ordinal analysis

To understand the apparent well-foundedness of theories by proof-theoretic strength, we focus our attention on descending sequences of theories in these hierarchies. We want to show that no such descending sequence can be simple, in some formal sense. For instance, we would like to show that all descending sequences are descriptively complex, or contain elements that are unsound.

In Chapter 4, co-authored with Fedor Pakhomov, we prove a number of theorems along these lines. The first such theorem concerns, not the ordering $\prec_{\mathsf{Con}}$ on axiomatic theories,

but a closely related structure. We say that a theory is Σ_2 sound if all its Σ_2 consequences are true. We prove that there are no "simple" descending sequences in the Σ_2 soundness hierarchy.

Theorem 1.7.5. *There is no recursively enumerable sequence $(T_n)_{n\in\mathbb{N}}$ of Σ_2 sound extensions of $\mathsf{B}\Sigma_1$ such that, for each n, T_n proves the Σ_2 soundness of T_{n+1}.*[3]

Theorem 1.7.5 is, in a sense to be made precise, the best possible. To state this precisely, we recall that a theory is Σ_1 sound if all its Σ_1 consequences are true. The following theorem demonstrates that Theorem 1.7.5 is the best possible.

Theorem 1.7.6. *There exists a recursive sequence $(T_n)_{n\in\mathbb{N}}$ of Σ_1 sound extensions of elementary arithmetic such that, for each n, T_n proves the Σ_1 soundness of T_{n+1}.*

The method introduced to prove Theorem 1.7.5 easily adapts to rule out *all* descending sequences in another hierarchy of proof-theoretic strength, namely, the one given by Π^1_1 soundness.

Theorem 1.7.7. *There is no sequence $(T_n)_{n\in\mathbb{N}}$ of Π^1_1 sound extensions of ACA_0 such that, for each n, T_n proves the Π^1_1 soundness of T_{n+1}.*

Theorem 1.7.7 rules out descending sequences of theories according to a certain metric of logical strength. This makes it possible to rank theories according to this metric of strength. Let's call the rank of a theory in this hierarchy its *reflection rank*. Recall that in ordinal analysis, values called *proof-theoretic ordinals* are systematically assigned to theories and that, according to conventional wisdom, proof-theoretic ordinals measure the proof-theoretic strength of theories. We vindicate this conventional wisdom by proving that for most theories T, the reflection rank of T equals the proof-theoretic ordinal of T. We make the notion of "most theories" precise in the following way (ACA^+_0 is an axiomatic theory extending ACA_0 with the axiom "every set is contained in an ω model of ACA_0").

Theorem 1.7.8. *For any Π^1_1-sound extension T of ACA^+_0, the reflection rank of T equals the Π^1_1 proof-theoretic ordinal of T.*

We derive Theorem 1.7.8 from a variant of Schmerl's formula in the context of second-order arithmetic. To precisely state this variant of Schmerl's formula, we introduce some terminology. Let $\Pi^1_1(\Pi^0_3)$ be the syntactic complexity class consisting of formulas of the form $\forall X\varphi(X)$ where $\varphi(X)$ is Π^0_3. Though all Π^1_1 formulas are provably equivalent to such formulas in ACA_0, this is not the case in RCA_0. We define $\mathbf{R}^\alpha_\Gamma(T)$ as the result of iterating Γ reflection

[3]In Chapter 4 this is stated as a result about EA rather than $\mathsf{B}\Sigma_1$. This is because we formalize our results in terms of "smooth provability" instead of using the ordinary provability predicate. Our proof works for smooth provability, but not ordinary provability, over EA, whereas it works for both over $\mathsf{B}\Sigma_1$. Rather than explicate the notion of smooth provability here we state our result about extensions of $\mathsf{B}\Sigma_1$.

α many times over T. Roughly, the iterates of reflection are defined as the solutions of the following equation:

$$\mathbf{R}^\alpha_\Gamma(T) := \mathsf{RFN}_\Gamma\big(T + \{\mathbf{R}^\beta_\Gamma(T) : \beta \prec \alpha\}\big)$$

We are now ready to state our variant of Schmerl's formula for second-order arithmetic:

Theorem 1.7.9. $\mathbf{R}^\alpha_{\Pi^1_1}(\mathsf{ACA}_0)$ *is* $\Pi^1_1(\Pi^0_3)$ *conservative over* $\mathbf{R}^{\varepsilon_\alpha}_{\Pi^1_1(\Pi^0_3)}(\mathsf{RCA}_0)$.

A reduction principle for ω-model reflection

In Chapter 5, co-authored with Fedor Pakhomov, we prove a conservation theorem relating different reflection principles in second-order arithmetic. In particular, our result reduces ω model reflection to iterated syntactic reflection principles. There is a thorough proof-theoretic understanding of the latter in terms, e.g., of ordinal analysis. Accordingly, these reductions yield proof-theoretic analyses of ω model reflection principles.

We formalize all our results in ACA_0 about the language $\mathbf{L}_2$ that extends the standard language of second-order arithmetic with set-constants C_X for all sets X. Boldface notation is used to define the standard syntactic complexity classes for $\mathbf{L}_2$.

The main syntactic reflection principle we consider, $\mathbf{\Pi}^1_1$-$\mathsf{RFN}(T)$, informally says "all $\mathbf{\Pi}^1_1$ theorems of T are true." We will also be interested in the theories $\mathbf{\Pi}^1_1$-$\mathbf{R}^\alpha(T)$ that result from iterating this principle along well-orderings. We give precise definitions of these theories via Gödel's fixed point lemma in §5.2. Informally, one can think of them as defined inductively, according to the following equation:

$$\mathbf{\Pi}^1_1\text{-}\mathbf{R}^\alpha(T) := T + \{\mathbf{\Pi}^1_1\text{-}\mathsf{RFN}\big(T + \mathbf{\Pi}^1_1\text{-}\mathbf{R}^\beta(T)\big) : \beta < \alpha\}$$

We will also consider ω model reflection principles. An ω model is an L_2 structure whose first-order part is $\mathbb{N}$ and whose second-order part is some subset of $\mathcal{P}(\mathbb{N})$. The semantic reflection principle we work with is "every set is contained in an ω model of T."

The main theorem of Chapter 5 is the following:

Theorem 1.7.10 (ACA_0)**.** *For any* $\mathbf{\Pi}^1_2$*-axiomatizable* T*, the following are equivalent:*

1. *Every set is contained in an* ω *model of* T*.*
2. $\forall\alpha\big(\mathsf{WO}(\alpha) \to \mathbf{\Pi}^1_1\text{-}\mathsf{RFN}(\mathbf{\Pi}^1_1\text{-}\mathbf{R}^\alpha(T))\big)$.

Incompleteness and jump hierarchies

Theorem 1.7.7 demonstrates that the proof-theoretic hierarchy based on Π^1_1 reflection is actually well-founded, not just well-founded for natural theories. In fact, a similar phenomenon happens on the recursion-theoretic side when we shift our attention from the Turing degrees to the *hyperdegrees*. For reals A and B, we say that $A \leqslant_H B$ if A is hyperarithmetical in B. For any real A, there is a canonical real $\mathcal{O}^A$, known as the *hyperjump* of A, such that

$A <_H \mathcal{O}^A$. In Chapter 6, co-authored with Patrick Lutz, we refine the technique introduced to prove Theorem 1.7.7 to provide a novel proof of a classical theorem (originally due to Spector):

Theorem 1.7.11. *There is no sequence $(A_n)_{n\in\mathbb{N}}$ of reals such that, for each n, $\mathcal{O}^{A_{n+1}} \leqslant_H A_n$.*

Whereas Spector's proof relies on the theory of admissible ordinals, our proof uses Gödel's second incompleteness theorem. We then use Theorem 1.7.11 to give a novel proof of a semantic incompleteness theorem originally due to Mummert and Simpson:

Theorem 1.7.12. *Let T be an L_2 theory. For each $n \geqslant 1$, if there is a β_n-model of T then there is a β_n-model of T which contains no countable coded β_n-models of T.*

The analogy with Theorem 1.7.7 raises the question: what is the analogue of ordinal analysis in the hyperdegress? Indeed, Theorem 1.7.11 states that the relation $A \prec B$ defined by $\mathcal{O}^A \leqslant_H B$ is a well founded partial order. We call the $\prec$ rank of a real its *Spector rank*. There is a recursion-theoretically natural characterization of the Spector ranks of reals:

Theorem 1.7.13. *For any real A, the Spector rank of A is α just in case ω_1^A is the $(1+\alpha)^{th}$ admissible ordinal.*

Corollary 1.7.14. *Assuming suitable large cardinal hypotheses, for all X on a cone, the Spector rank of X is ω_1^X.*

These new proofs and results strengthen the analogy between the proof-theoretic and recursion-theoretic well-ordering phenomena.

Epistemic aspects of consistency proofs

Ordinal analysis plays a major role in this dissertation, as a means of addressing the problem of the well-ordering of natural theories. In Chapter 7 I discuss the role of ordinal analysis in developing consistency proofs. With respect to the epistemic significance of Gentzen-style consistency proofs, I emphasize *explanation* over *justification*. Philosophers have argued about the prospects of using Gentzen's proof to justify the claim that arithmetic is consistent, or to enhance our credence that arithmetic is consistent. I will argue that, independent of its force as a credence-enhancer, the proof has utility as an understanding-enhancer. Whether Gentzen's proof convinces us *that* arithmetic is consistent, it explains *why* arithmetic is consistent. Gentzen's proof bears many of the marks of explanation: for instance, it has facilitated new discoveries, including surprising discoveries about unfamiliar contexts, and it is articulated at a level of grain that is conducive to generalization and even to reversal. I argue that all the major theories of mathematical explanation predict that explanatory proofs should bear these marks; this provides strong evidence, which lies in the overlapping consensus of the most prominent theories of mathematical explanation, that Gentzen's proof is explanatory. On this basis, I argue against certain restrictive norms on mathematical practice. For instance, I argue against the norm that consistency proofs ought to be carried out

using only constructive methods. Though this restriction is reasonable if one is attempting to convince a constructivist skeptic, it loses force when one instead is attempting to present a combinatorial explanation of a theory's consistency.

Part I

First-Order Arithmetic

Chapter 2

On the inevitability of the consistency operator

This chapter contains joint work with Antonio Montalbán.

2.1 Introduction

It is a well-known empirical phenomenon that natural axiomatic theories are well-ordered by their consistency strength. However, without a precise mathematical definition of "natural," it is difficult to explain this observation in a strictly mathematical way. One expression of this phenomenon comes from *ordinal analysis*, a research program whereby recursive ordinals are assigned to theories as a measurement of their consistency strength. One method for calculating the proof-theoretic ordinal of a theory T involves demonstrating that T can be approximated over a weak base theory by a class of formulas that are well understood. In particular, the Π^0_1 fragments of natural theories are often proof-theoretically equivalent to iterated consistency statements over a weak base theory, making these theories amenable to ordinal analysis. For discussion, see, e.g., Beklemishev [7, 8] and Joosten [46].

Why are the Π^0_1 fragments of natural theories proof-theoretically equivalent to iterated consistency statements? Our approach to this question is inspired by Martin's approach to another famous question from mathematical logic: why are natural Turing degrees well-ordered by Turing reducibility? Martin conjectured that (i) the non-constant degree invariant functions meeting a certain simplicity condition ($f \in L(\mathbb{R})$)[1] are pre-well-ordered by the relation "$f(a) \leqslant_T g(a)$ on a cone in the Turing degrees" and (ii) the successor for this well-ordering is induced by the Turing jump. Martin's conjecture is meant to capture the idea that the Turing jump and its iterates into the transfinite are the only natural non-trivial degree invariant functions.

[1] Martin's Conjecture is stated under the hypothesis $\mathsf{ZF} + \mathsf{AD} + \mathsf{DC}$, which is satisfied by $L(\mathbb{R})$ assuming that there are ω many Woodin cardinals with a measurable above them all.

In this paper we investigate analogous hypotheses concerning jumps on consistent axiomatic theories, namely, consistency statements. We fix elementary arithmetic EA as our base theory. EA is a subsystem of PA that is often used as a base theory in ordinal analysis and in which standard approaches to arithmetization of syntax can be carried out without substantial changes; see [9] for details. We write $[\varphi]$ to denote the equivalence class of φ modulo EA-provable equivalence. We write $\varphi \vdash \psi$ if $\mathsf{EA} \vdash \varphi \to \psi$ and say that φ *implies* ψ. If $\varphi \vdash \psi$ but $\psi \nvdash \varphi$ we say that φ *strictly implies* ψ. The *Lindenbaum algebra* of EA is the set of equivalence classes of sentences ordered by $\vdash$. We focus on recursive functions f that are *monotonic*, i.e.,

$$\text{if } \varphi \vdash \psi, \text{ then } f(\varphi) \vdash f(\psi).$$

We note that (i) a function f is monotonic just in case f preserves implication over EA and (ii) all monotonic functions induce functions on the Lindenbaum algebra of EA. We adopt the convention that all functions named "f" in this paper are recursive.

Our goal is to demonstrate that $\varphi \mapsto (\varphi \wedge \mathsf{Con}(\varphi))$ and its iterates into the transfinite are canonical among monotonic functions. Our first theorem to this end is the following.

Theorem 2.1.1. *Let f be monotonic. Suppose that for all consistent φ,*
(i) $\varphi \wedge \mathsf{Con}(\varphi)$ implies $f(\varphi)$ and
(ii) $f(\varphi)$ strictly implies φ.
Then for every true φ, there is a true ψ such that $\psi \vdash \varphi$ and $[f(\psi)] = [\psi \wedge \mathsf{Con}(\psi)]$.

Corollary 2.1.2. *There is no monotonic function f such that for all consistent φ,*
(i) $\varphi \wedge \mathsf{Con}(\varphi)$ strictly implies $f(\varphi)$ and
(ii) $f(\varphi)$ strictly implies φ.

We note that this result depends essentially on the condition of monotonicity. Shavrukov and Visser [90] studied recursive functions f that are *extensional* over the Lindenbaum algebra of PA, i.e.,

$$\text{if } \mathsf{PA} \vdash (\varphi \leftrightarrow \psi), \text{ then } \mathsf{PA} \vdash (f(\varphi) \leftrightarrow f(\psi)),$$

and proved the following theorem.

Theorem 2.1.3. *(Shavrukov–Visser) There is a recursive extensional function f such that for all consistent φ,*
(i) $\varphi \wedge \mathsf{Con}(\varphi)$ strictly implies $f(\varphi)$ and
(ii) $f(\varphi)$ strictly implies φ.

In particular, Shavrukov and Visser proved that for any consistent φ, the sentence

$$\varphi^\star := \varphi \wedge \forall x \big(\mathsf{Con}(I\Sigma_x + \varphi) \to \mathsf{Con}(I\Sigma_x + \varphi + \mathsf{Con}(I\Sigma_x + \varphi)) \big)$$

has deductive strength strictly between φ and $\varphi \wedge \mathsf{Con}(\varphi)$, and that the map $\varphi \mapsto \varphi^\star$ is extensional. By a theorem of Kripke and Pour-El [77], the Lindenbaum algebras of PA and EA are effectively isomorphic, whence Theorem 2.1.3 also applies to EA. Thus, Corollary

2.1.2 cannot be strengthened by weakening the hypothesis of monotonicity to the hypothesis of extensionality.

We also note that Friedman, Rathjen, and Weiermann [33] introduced a notion of *slow consistency* with which they produced a Π^0_1 sentence $\mathsf{SlowCon}(\mathsf{PA})$ with deductive strength strictly between PA and $\mathsf{PA} + \mathsf{Con}(\mathsf{PA})$. In general, the statement $\mathsf{SlowCon}(\varphi)$ has the form

$$\forall x(F_{\epsilon_0}(x) \downarrow \rightarrow \mathsf{Con}(I\Sigma_x + \varphi))$$

where F_{ϵ_0} is a standard representation of a recursive function that is not provably total in PA. This is not in conflict with Corollary 2.1.2, however, since $\varphi \wedge \mathsf{Con}(\varphi)$ and $\varphi \wedge \mathsf{SlowCon}(\varphi)$ are provably equivalent for all φ such that $\varphi \vdash \forall x F_{\epsilon_0}(x) \downarrow$. On the other hand, changing the definition of the $\mathsf{SlowCon}(\varphi)$ so that the function in the antecedent varies with the input φ results in a map that is not monotonic.

Theorem 2.1.1 generalizes to the iterates of Con into the effective transfinite. For an elementary presentation α of a recursive well-ordering (see Definition 2.3.1) and a sentence φ, we define sentences $\mathsf{Con}^\beta(\varphi)$ for every $\beta < \alpha$.

$$\begin{aligned}\mathsf{Con}^0(\varphi) :=& \top \\ \mathsf{Con}^{\beta+1}(\varphi) :=& \mathsf{Con}(\varphi \wedge \mathsf{Con}^\beta(\varphi)) \\ \mathsf{Con}^\lambda(\varphi) :=& \forall \beta < \lambda(\mathsf{Con}^\beta(\varphi))\end{aligned}$$

For a precise definition using Gödel's fixed point lemma, see Definition 2.3.2. Note that for every φ, $[\mathsf{Con}^1(\varphi)] = [\mathsf{Con}(\varphi)]$.

Remark 2.1.4. We warn the reader that there is some discrepancy between our notation and the notation used by other authors. Our iteration scheme $\mathsf{Con}^{\alpha+1}(\varphi) \equiv \mathsf{Con}(\varphi \wedge \mathsf{Con}^\alpha(\varphi))$ is sometimes denoted $\mathsf{Con}((\mathsf{EA}+\varphi)_\alpha)$, e.g., [6]. Moreoever, the notation $\mathsf{Con}^{\alpha+1}(\varphi)$ is sometimes used to denote $\mathsf{Con}(\mathsf{Con}^\alpha(\varphi))$, e.g., [15].

With each predicate Con^α we associate a function

$$\varphi \mapsto (\varphi \wedge \mathsf{Con}^\alpha(\varphi)).$$

Theorem 2.1.1 then generalizes into the effective transfinite as follows.

Theorem 2.1.5. *Let f be monotonic. Suppose that for all φ,*
(i) $\varphi \wedge \mathsf{Con}^\alpha(\varphi)$ implies $f(\varphi)$,
(ii) if $[f(\varphi)] \neq [\bot]$, then $f(\varphi)$ strictly implies $\varphi \wedge \mathsf{Con}^\beta(\varphi)$ for all $\beta < \alpha$.
Then for every true φ, there is a true ψ such that $\psi \vdash \varphi$ and $[f(\psi)] = [\psi \wedge \mathsf{Con}^\alpha(\psi)]$.

Corollary 2.1.6. *There is no monotonic f such that for all φ, if $[\varphi \wedge \mathsf{Con}^\alpha(\varphi)] \neq [\bot]$, then both*
(i) $\varphi \wedge \mathsf{Con}^\alpha(\varphi)$ strictly implies $f(\varphi)$ and
(ii) $f(\varphi)$ strictly implies $\varphi \wedge \mathsf{Con}^\beta(\varphi)$ for all $\beta < \alpha$.

Thus, if the range of a monotonic function f is sufficiently constrained, then for some φ and some α,

$$[f(\varphi)] = [\varphi \wedge \mathsf{Con}^{\alpha}(\varphi)] \neq [\bot].$$

This property still holds even when these constraints on the range of f are relaxed considerably. More precisely, if a monotonic function is everywhere bounded by a finite iterate of Con, then it must be somewhere equivalent to an iterate of Con.

Theorem 2.1.7. *Let $n \in \mathbb{N}$. Let f be a monotonic function such that for every φ,*
(i) $\varphi \wedge \mathsf{Con}^{n}(\varphi)$ implies $f(\varphi)$ and
(ii) $f(\varphi)$ implies φ.
Then for some φ and some $k \leqslant n$, $[f(\varphi)] = [\varphi \wedge \mathsf{Con}^{k}(\varphi)] \neq [\bot]$.

To generalize this result into the effective transfinite, we focus on a particular class of monotonic functions that we call Π^0_1.

Definition 2.1.8. A function f is Π^0_1 if $f(\varphi) \in \Pi^0_1$ for all φ.

Our main theorem is the following: if a monotonic function is everywhere bounded by a transfinite iterate of Con, then it must be somewhere equivalent to an iterate of Con. This to say that the iterates of the consistency operator are *inevitable*; no monotonic function that is everywhere bounded by some iterate of Con can avoid all of the iterates of Con.

Theorem 2.1.9. *Let $\varphi \mapsto f(\varphi)$ be a monotonic Π^0_1 function Then either*
(i) for some $\beta \leqslant \alpha$ and some φ, $[\varphi \wedge f(\varphi)] = [\varphi \wedge \mathsf{Con}^{\beta}(\varphi)] \neq [\bot]$ or
(ii) for some φ, $(\varphi \wedge \mathsf{Con}^{\alpha}(\varphi)) \nvdash f(\varphi)$.

The main theorem bears a striking similarity to the following theorem of Slaman and Steel [94].

Theorem 2.1.10. *(Slaman–Steel) Suppose $f : 2^{\omega} \to 2^{\omega}$ is Borel, order-preserving with respect to $\leqslant_T$, and increasing on a cone. Then for any $\alpha < \omega_1$ either*
(i) for some $\beta \leqslant \alpha$, $f(x) \equiv_T x^{(\beta)}$ cofinally or
(ii) $(x^{(\alpha)} <_T f(x))$ cofinally.

There are two notable disanalogies between Theorem 2.1.9 and Theorem 2.1.10. First, Theorem 2.1.9 guarantees only that sufficiently constrained functions are *somewhere* equivalent to an iterate of Con, whereas Theorem 2.1.10 guarantees *cofinal* equivalence with an iterate of the Turing jump. Second, by assuming AD, Slaman and Steel inferred that this behavior happens not only cofinally but also *on a cone* in the Turing degrees. There is no obvious analogue of AD from which one can infer that if cofinally many Lindenbaum degrees have a property then every element in some non-trivial ideal of Lindenbaum degrees has that property.

We then turn our attention to a generalization of consistency, namely, 1-consistency. Recall that a theory T is *1-consistent* if T is consistent with the true Π^0_1 theory of arithmetic.

Just as the Π^0_1 fragments of natural theories are often proof-theoretically equivalent to iterated consistency statements over a weak base theory, the Π^0_2 fragments of natural theories are often proof-theoretically equivalent to iterated 1-consistency statements over a weak base theory

Conservativity theorems relating 1-consistency and iterated consistency play an important role in the proof-theoretic analysis of arithmetic theories. For instance, it is a consequence of Beklemshev's *reduction principle* [9] that for any Π^0_1 φ,

$$\mathsf{EA} + 1\mathsf{Con}(\mathsf{EA}) \vdash \varphi \text{ if and only if } \mathsf{EA} + \{\mathsf{Con}^k(\mathsf{EA}) : k < \omega\} \vdash \varphi.$$

This fact plays an integral role in Beklemishev's [8] consistency proof of PA. We show that this conservativity result is drastically violated *in the limit*. For functions f and g, we say that f *majorizes* g if there is a consistent φ such that for all ψ, if $\psi \vdash \varphi$ then $f(\psi) \vdash g(\psi)$; if in addition φ is true then we say that f *majorizes* g *on a true ideal.*

Proposition 2.1.11. *For any elementary presentation α of a recursive well-ordering,* $1\mathsf{Con}$ *majorizes* Con^α *on a true ideal.*

It is tempting to conjecture on the basis of this result that $1\mathsf{Con}$ is the weakest monotonic function majorizing each Con^α for α a recursive well-ordering. We prove that this is not the case.

Theorem 2.1.12. *There are infinitely many monotonic functions f such that for every recursive ordinal α, there is an elementary presentation a of α such that f majorizes* Con^a *on a true ideal but also* $1\mathsf{Con}$ *majorizes f on a true ideal.*

Theorem 2.1.1 demonstrates that for any monotonic f with a sufficiently constrained range, f must agree cofinally with Con. We would like to strengthen *cofinally* to *on a true ideal*. One strategy for establishing this claim would be to show that every set that is closed under EA provable equivalence and that contains cofinally many true sentences also contains every sentence in some true ideal. We show that this strategy fails.

Proposition 2.1.13. *There is a recursively enumerable set $\mathcal{A}$ that contains arbitrarily strong true sentences and that is closed under* EA *provable equivalence but does not contain any true ideals.*

It is not clear whether Theorem 2.1.1 can be strengthened in the desired manner.

2.2 No monotonic function is strictly between the identity and Con

In this section we prove that no monotonic function sends every consistent φ to a sentence with deductive strength strictly between φ and $(\varphi \wedge \mathsf{Con}(\varphi))$. Most of the work is contained in the proof of the following lemma.

Lemma 2.2.1. *Let f be a monotonic function such that for all consistent φ, $f(\varphi)$ strictly implies φ. Then for every true sentence φ there is a true sentence θ such that $\theta \vdash \varphi$ and $f(\theta) \vdash (\theta \wedge \mathsf{Con}(\theta))$.*

Proof. Let f be as in the statement of the theorem. By assumption the following statement is true.

$$\chi := \forall\zeta(\mathsf{Con}(\zeta) \rightarrow \mathsf{Con}(\zeta \wedge \neg f(\zeta)))$$

Let φ be a true sentence. Then the sentence $\psi := \varphi \wedge \chi$ is true. Let

$$\theta := (\psi \wedge (f(\psi) \rightarrow \mathsf{Con}(\psi))).$$

Note that $\theta \vdash \varphi$.

Claim. $f(\theta) \vdash (\theta \wedge f(\psi))$.

Clearly $\theta \vdash \psi$. So $f(\theta) \vdash f(\psi)$ since f is monotonic. Also $f(\theta) \vdash \theta$ by assumption.

Claim. $(\theta \wedge f(\psi)) \vdash (\psi \wedge \mathsf{Con}(\psi))$.

Immediate from the definition of θ.

Claim. $(\psi \wedge \mathsf{Con}(\psi)) \vdash (\theta \wedge \mathsf{Con}(\theta))$.

Clearly $(\psi \wedge \mathsf{Con}(\psi)) \vdash \theta$. It suffices to show that

$$(\psi \wedge \mathsf{Con}(\psi)) \vdash \mathsf{Con}(\theta).$$

We reason as follows.

$$\begin{aligned}(\psi \wedge \mathsf{Con}(\psi)) &\vdash \forall\zeta(\mathsf{Con}(\zeta) \rightarrow \mathsf{Con}(\zeta \wedge \neg f(\zeta))) \text{ by choice of } \psi.\\ &\vdash \mathsf{Con}(\psi) \rightarrow \mathsf{Con}(\psi \wedge \neg f(\psi)) \text{ by instantiation.}\\ &\vdash \mathsf{Con}(\psi \wedge \neg f(\psi)) \text{ by logic.}\\ &\vdash \mathsf{Con}(\theta) \text{ by the definition of } \theta.\end{aligned}$$

It is immediate from the preceding claims that $f(\theta) \vdash (\theta \wedge \mathsf{Con}(\theta))$. ❑

A number of results follow immediately from the lemma.

Theorem 2.2.2 (Restatement of Theorem 2.1.1)**.** *Let f be monotonic. Suppose that for all consistent φ,*
(i) $\varphi \wedge \mathsf{Con}(\varphi)$ implies $f(\varphi)$ and
(ii) $f(\varphi)$ strictly implies φ.
Then for every true φ, there is a true ψ such that $\psi \vdash \varphi$ and $[f(\psi)] = [\psi \wedge \mathsf{Con}(\psi)]$.

Proof. By the lemma, for every true φ there is a true ψ such that $\psi \vdash \varphi$ and $f(\psi) \vdash (\psi \wedge \mathsf{Con}(\psi))$. Since we are assuming that $(\psi \wedge \mathsf{Con}(\psi)) \vdash f(\psi)$, it follows that $[f(\psi)] = [\psi \wedge \mathsf{Con}(\psi)]$. ❑

We note that this theorem applies to a number of previously studied operators. For instance, the theorem applies to the notion of *cut-free consistency*, i.e., consistency with respect to cut-free proofs. EA does not prove the cut-elimination theorem, which is equivalent to the totality of super-exponentiation (over EA), and does not prove the equivalence of cut-free consistency and consistency. Another such operator is the Friedman-Rathjen-Weiermann *slow consistency operator* discussed in §1. Theorem 4.1.2 implies that these operators exhibit the same behavior as the consistency operator "in the limit." Indeed, for any φ such that φ proves the cut-elimination theorem, $\varphi \wedge \mathsf{Con}(\varphi)$ and $\varphi \wedge \mathsf{Con}_{\mathsf{CF}}(\varphi)$ are EA-provably equivalent. Likewise, for any φ that proves the totality of F_{ϵ_0}, $\varphi \wedge \mathsf{Con}(\varphi)$ and $\varphi \wedge \mathsf{SlowCon}(\varphi)$ are EA-provably equivalent.

As a corollary of Theorem 4.1.2 we note that no monotonic function reliably produces sentences strictly between those produced by the identity and by Con.

Corollary 2.2.3 (Restatement of Corollary 2.1.2)**.** *There is no monotonic function f such that for all consistent φ,*
(i) $\varphi \wedge \mathsf{Con}(\varphi)$ strictly implies $f(\varphi)$ and
(ii) $f(\varphi)$ strictly implies φ.

Shavrukov and Visser [90] studied functions over Lindenbaum algberas and discovered a recursive *extensional uniform density function* g for the Lindenbaum algebra of EA, i.e., (i) for any φ and ψ such that ψ strictly implies φ, $g(\langle \varphi, \psi \rangle)$ is a sentence with deductive strength strictly between φ and ψ and (ii) if $\mathsf{EA} \vdash (\varphi \leftrightarrow \psi)$ then, for any θ, $[g(\langle \varphi, \theta \rangle)] = [g(\langle \psi, \theta \rangle)]$ and $[g(\langle \theta, \varphi \rangle)] = [g(\langle \theta, \psi \rangle)]$. They asked whether this result could be strengthened by exhibiting a recursive uniform density function that is monotonic in both its coordinates. As a corollary of our theorem we answer their question negatively.

Corollary 2.2.4. *There is no monotonic uniform density function for the Lindenbaum algebra of* EA*.*

Proof. Suppose there were such a function g over the Lindenbaum algebra of EA. Then given any input of the form $\langle \varphi, (\varphi \wedge \mathsf{Con}(\varphi)) \rangle$, g would produce a sentence with deductive strength strictly between φ and $(\varphi \wedge \mathsf{Con}(\varphi))$. We then note that $f : \varphi \mapsto g(\langle \varphi, (\varphi \wedge \mathsf{Con}(\varphi)) \rangle)$ is monotonic, but that for every consistent φ, $\varphi \wedge \mathsf{Con}(\varphi)$ strictly implies $f(\varphi)$ and $f(\varphi)$ strictly implies φ, contradicting the previous theorem. ❑

Our negative answer to the question raised by Shavrukov and Visser makes use of a Π^0_2 sentence $\forall \zeta(\mathsf{Con}(\zeta) \rightarrow \mathsf{Con}(\zeta \wedge \neg f(\zeta)))$. Shavrukov and Visser raised the following question in private communication.

Question 2.2.5. *Is there a recursive uniform density function for the lattice of Π^0_1 sentences over EA that is monotonic in both its coordinates?*

Remark 2.2.6. It is clear from the proof of the lemma that any monotonic f meeting the hypotheses of Theorem 4.1.2 is not only cofinally equivalent to Con; for every true ψ that

implies

$$\chi := \forall\zeta(\mathsf{Con}(\zeta) \to \mathsf{Con}(\zeta \wedge \neg f(\zeta))),$$

there is a true θ such that $\theta \vdash \psi$ and $[\psi \wedge \mathsf{Con}(\psi)] = [\theta \wedge \mathsf{Con}(\theta)] = [f(\theta)]$.

This observation points the way toward a corollary of our theorem; namely that any monotonic function strictly meeting the hypotheses of the theorem must have the same range as $\varphi \mapsto (\varphi \wedge \mathsf{Con}(\varphi))$ in the limit. To prove this, we first prove a version of jump inversion—$\varphi \mapsto (\varphi \wedge \mathsf{Con}(\varphi))$ inversion—for Lindenbaum algebras. This is to say that the range of Con contains a true ideal in the Lindenbaum algebra. A similar result is established for true Π^0_2 sentences in [1].

Proposition 2.2.7. *Suppose $\varphi \vdash \mathsf{Con}(\top)$. Then for some ψ, $[\varphi] = [(\psi \wedge \mathsf{Con}(\psi))]$.*

Proof. Let $\psi := \mathsf{Con}(\top) \to \varphi$.

Claim. $\varphi \vdash (\psi \wedge \mathsf{Con}(\psi))$.

Trivially, $\varphi \vdash \psi$. Since $\varphi \vdash \mathsf{Con}(\top)$, it follows that from the formalized second incompleteness theorem, i.e., $\mathsf{Con}(\top) \vdash \mathsf{Con}(\neg\mathsf{Con}(\top))$, that $\varphi \vdash \mathsf{Con}(\neg\mathsf{Con}(\top))$. But $\neg\mathsf{Con}(\top)$ is the first disjunct of ψ, so $\varphi \vdash \mathsf{Con}(\psi)$.

Claim. $(\psi \wedge \mathsf{Con}(\psi)) \vdash \varphi$.

Note that $\mathsf{Con}(\psi) \vdash \mathsf{Con}(\top)$. The claim then follows since clearly $(\psi \wedge \mathsf{Con}(\top)) \vdash \varphi$. ❑

Corollary 2.2.8. *Let f be monotonic. Suppose that for all consistent φ,*
(i) $\varphi \wedge \mathsf{Con}(\varphi)$ implies $f(\varphi)$ and
(ii) $f(\varphi)$ strictly implies φ.
Then the intersection of the ranges of f and Con in the Lindenbaum algebra contains a true ideal.

Proof. Let φ be a sentence such that $\varphi \vdash \mathsf{Con}(\top)$ and

$$\varphi \vdash \forall\zeta(\mathsf{Con}(\zeta) \to \mathsf{Con}(\zeta \wedge \neg f(\zeta))).$$

Note that both of these sentences are true, and hence φ is in an element of a true ideal. By the previous proposition, there is a ψ such that $[\psi \wedge \mathsf{Con}(\psi)] = [\varphi]$. By Remark 2.2.6 there is a θ such that $[f(\theta)] = [\psi \wedge \mathsf{Con}(\psi)]$, that is, φ is in the range of f. ❑

2.3 Iterating Con into the transfinite

By analogy with Martin's Conjecture, we would like to show that there is a natural well-ordered hierarchy of monotonic functions and that the successor for this well-ordering is induced by Con. Thus, we define the iterates of Con along elementary presentations of well-orderings.

Definition 2.3.1. By an *elementary presentation* of a recursive well-ordering we mean a pair $(\mathcal{D}, <)$ of elementary formulas, such that (i) the relation $<$ well-orders $\mathcal{D}$ in the standard model of arithmetic and (ii) EA proves that $<$ linearly orders the elements satisfying $\mathcal{D}$, (iii) it is elementarily calculable whether an element represents zero or a successor or a limit and (iv) the elementary formulas defining the set of limit ordinals and the successor relation provably in EA satisfy their corresponding first order definitions in terms of $<$.

Definition 2.3.2. Given an elementary presentation $\langle \alpha, < \rangle$ of a recursive well-ordering and a sentence φ, we use Gödel's fixed point lemma to define sentences $\mathbf{Con}^\star(\varphi, \beta)$ for $\beta < \alpha$ as follows.

$$\mathsf{EA} \vdash \mathbf{Con}^\star(\varphi, \beta) \leftrightarrow \forall \gamma < \beta, \mathsf{Con}(\varphi \wedge \mathbf{Con}^\star(\varphi, \gamma)).$$

We use the notation $\mathsf{Con}^\beta(\varphi)$ for $\mathbf{Con}^\star(\varphi, \beta)$.

Remark 2.3.3. Note that, since the following clauses are provable in EA.

- $\mathsf{Con}^0(\varphi) \leftrightarrow \top$
- $\mathsf{Con}^{\gamma+1}(\varphi) \leftrightarrow \mathsf{Con}(\varphi \wedge \mathsf{Con}^\gamma(\varphi))$
- $\mathsf{Con}^\lambda(\varphi) \leftrightarrow \forall \gamma < \lambda, \mathsf{Con}^\gamma(\varphi)$ for λ a limit.

Note that this hierarchy is proper for true φ by Gödel's second incompleteness theorem. We need to prove that for transfinite α, Con^α is monotonic over the Lindenbaum algebra of EA. Before proving this claim we recall Schmerl's [87] technique of *reflexive transfinite induction*. Note that "$\mathsf{Pr}(\varphi)$" means that φ is provable in EA.

Proposition 2.3.4. *(Schmerl) Suppose that $<$ is an elementary linear order and that $\mathsf{EA} \vdash \forall \alpha(\mathsf{Pr}(\forall \beta < \alpha, A(\beta)) \rightarrow A(\alpha))$. Then $\mathsf{EA} \vdash \forall \alpha A(\alpha)$.*

Proof. From $\mathsf{EA} \vdash \forall \alpha(\mathsf{Pr}(\forall \beta < \alpha, A(\beta)) \rightarrow A(\alpha))$ we infer

$$\begin{aligned} \mathsf{EA} \vdash \mathsf{Pr}(\forall \alpha A(\alpha)) &\rightarrow \forall \alpha \mathsf{Pr}(\forall \beta < \alpha, A(\beta)) \\ &\rightarrow \forall \alpha A(\alpha). \end{aligned}$$

Löb's theorem, i.e.,

$$\text{if } \mathsf{EA} \vdash \mathsf{Pr}(\zeta) \rightarrow \zeta, \text{ then } \mathsf{EA} \vdash \zeta,$$

then yields $\mathsf{EA} \vdash \forall \alpha A(\alpha)$. ❑

Proposition 2.3.5. *If $\varphi \vdash \psi$, then $\mathsf{Con}^\alpha(\varphi) \vdash \mathsf{Con}^\alpha(\psi)$.*

Proof. Let $\mathcal{A}(\beta)$ denote the claim that $\mathsf{Con}^\beta(\varphi) \vdash \mathsf{Con}^\beta(\psi)$.

We want to prove that $\mathcal{A}(\alpha)$, without placing any restrictions on α. We prove the equivalent claim that $\mathsf{EA} \vdash \mathcal{A}(\alpha)$. By Proposition 2.3.4, it suffices to show that

$$\mathsf{EA} \vdash \forall \alpha(\mathsf{Pr}(\forall \beta < \alpha, \mathcal{A}(\beta)) \rightarrow \mathcal{A}(\alpha)).$$

Reason within EA**.** Suppose that $\mathsf{Pr}(\forall\beta < \alpha, \mathcal{A}(\beta))$, which is to say that

$$\mathsf{Pr}(\forall\beta < \alpha, \mathsf{Pr}(\mathsf{Con}^\beta(\varphi) \to \mathsf{Con}^\beta(\psi))).$$

Since $\mathsf{Con}^\alpha(\varphi)$ contains EA, we infer that

$$\mathsf{Con}^\alpha(\varphi) \vdash \forall\beta < \alpha \mathsf{Pr}(\mathsf{Con}^\beta(\varphi) \to \mathsf{Con}^\beta(\psi)).$$

Since $\mathsf{Con}^\alpha(\varphi)$ proves that for all $\beta < \alpha$, $\mathsf{EA} \nvdash \neg\mathsf{Con}^\beta(\varphi)$ we infer that

$$\mathsf{Con}^\alpha(\varphi) \vdash \forall\beta < \alpha \mathsf{Con}(\mathsf{Con}^\beta(\psi)).$$

Thus,

$$\mathsf{Con}^\alpha(\varphi) \vdash \forall\beta < \alpha(\mathsf{Con}^\beta(\psi)).$$

This concludes the proof of the proposition. ❑

Thus, for each predicate Con^α the function

$$\varphi \mapsto (\varphi \wedge \mathsf{Con}^\alpha(\varphi))$$

is monotonic over the Lindenbaum algebra of EA.

In this section we show that the functions given by iterated consistency are minimal with respect to each other. We fix an elementary presentation α of a recursive well-ordering. We assume that f is a monotonic function such that for every consistent φ, $f(\varphi)$ strictly implies $\varphi \wedge \mathsf{Con}^\beta(\varphi)$ for all $\beta < \alpha$. We would like to relativize the proof of Lemma 2.2.1 to Con^β. However, the proof of Lemma 2.2.1 relied on the truth of the principle

$$\forall\zeta(\mathsf{Con}(\zeta) \to \mathsf{Con}(\zeta \wedge \neg f(\zeta))).$$

It is not in general clear that $\mathsf{Con}^\alpha(\varphi)$ implies $\mathsf{Con}^\alpha(\varphi \wedge \neg f(\varphi))$. To solve this problem, we define a sequence of true sentences $(\theta_\beta)_{\beta\leqslant\alpha}$ such that for every sentence φ, if $\varphi \vdash \theta_\beta$ then $\mathsf{Con}^\beta(\varphi)$ implies $\mathsf{Con}^\beta(\varphi \wedge \neg f(\varphi))$. Thus, we are able to relativize the proof of Lemma 2.2.1 for Con^β to sentences that imply θ_β.

Definition 2.3.6. Given an elementary presentation α of a recursive well-ordering, we use Gödel's fixed point lemma to define sentences $\theta^\star(\beta)$ for $\beta < \alpha$ as follows.

$\mathsf{EA} \vdash \theta^\star(\beta) \leftrightarrow$

$$\forall\gamma < \beta(\mathsf{True}_{\Pi_3}(\theta^\star(\gamma))) \wedge \forall\zeta\Big(\big(\forall\gamma < \beta \mathsf{Pr}(\zeta \to \theta^\star(\gamma))\big) \to \big(\mathsf{Con}^\beta(\zeta) \to \mathsf{Con}^\beta(\zeta \wedge \neg f(\zeta))\big)\Big).$$

We use the notation θ_β for $\theta^\star(\beta)$.

Remark 2.3.7. Note that every sentence in the sequence $(\theta_\beta)_{\beta\leqslant\alpha}$ has complexity Π^0_3. Note moreover that for a successor $\beta + 1$, $\theta_{\beta+1}$ is equivalent to

$$\theta_\beta \wedge \forall\zeta\big(\mathsf{Pr}(\zeta \to \theta_\beta) \to \big(\mathsf{Con}^{\beta+1}(\zeta) \to \mathsf{Con}^{\beta+1}(\zeta \wedge \neg f(\zeta))\big)\big).$$

Lemma 2.3.8. *Let f be monotonic such that, for all φ,*
(i) $\varphi \wedge \mathsf{Con}^{\alpha}(\varphi)$ implies $f(\varphi)$,
(ii) if $[f(\varphi)] \neq [\bot]$, then $f(\varphi)$ strictly implies $\varphi \wedge \mathsf{Con}^{\beta}(\varphi)$ for all $\beta < \alpha$.
Then for each $\beta \leqslant \alpha$, the sentence θ_β is true.

Proof. Let f be as in the statement of the lemma. We prove the claim by induction on $\beta \leqslant \alpha$. The **base case** $\beta = 0$ is trivial.

For the **successor case** we assume that $\beta < \alpha$ and that θ_β is true; we want to show that $\theta_{\beta+1}$ is true. So let ζ be a sentence such that $\zeta \vdash \theta_\beta$. We want to show that $\mathsf{Con}^{\beta+1}(\zeta)$ implies $\mathsf{Con}^{\beta+1}(\zeta \wedge \neg f(\zeta))$. We prove the contrapositive, that $\neg\mathsf{Con}^{\beta+1}(\zeta \wedge \neg f(\zeta))$ implies $\neg\mathsf{Con}^{\beta+1}(\zeta)$. So suppose $\neg\mathsf{Con}^{\beta+1}(\zeta \wedge \neg f(\zeta))$, i.e.,

$$\zeta \wedge \neg f(\zeta) \vdash \neg\mathsf{Con}^{\beta}(\zeta \wedge \neg f(\zeta)). \tag{$\dagger$}$$

We reason as follows.

Since $\zeta \vdash \theta_\beta$, $\zeta \vdash \forall\gamma < \beta, \mathsf{True}_{\Pi_3}(\theta_\gamma)$. From this we infer

$$\zeta \vdash \mathsf{Pr}(\zeta \to \forall\gamma < \beta, \mathsf{True}_{\Pi_3}(\theta_\gamma)) \tag{$\star$}$$

by Σ_1^0 completeness. Moreover, since $\zeta \vdash \theta_\beta$,

$$\begin{aligned}
\zeta &\vdash \forall\varphi\big(\big(\forall\gamma < \beta \mathsf{Pr}(\varphi \to \theta_\gamma)\big) \to \big(\mathsf{Con}^{\beta}(\varphi) \to \mathsf{Con}^{\beta}(\varphi \wedge \neg f(\varphi))\big)\big) \text{ by the definition of } \theta_\beta.\\
&\vdash \forall\gamma < \beta \mathsf{Pr}\big(\zeta \to \theta_\gamma\big) \to \big(\mathsf{Con}^{\beta}(\zeta) \to \mathsf{Con}^{\beta}(\zeta \wedge \neg f(\zeta))\big) \text{ by instantiation.}\\
&\vdash \mathsf{Con}^{\beta}(\zeta) \to \mathsf{Con}^{\beta}(\zeta \wedge \neg f(\zeta)) \text{ by } (\star).\\
\zeta \wedge \neg f(\zeta) &\vdash \neg\mathsf{Con}^{\beta}(\zeta \wedge \neg f(\zeta)) \text{ by } (\dagger).\\
&\vdash \neg\mathsf{Con}^{\beta}(\zeta) \text{ by logic.}\\
\zeta &\vdash \mathsf{Con}^{\beta}(\zeta) \to f(\zeta) \text{ by logic.}
\end{aligned}$$

Thus, $(\zeta \wedge \mathsf{Con}^{\beta}(\zeta)) \vdash f(\zeta)$. Since $f(\varphi)$ always strictly implies $\varphi \wedge \mathsf{Con}^{\beta}(\varphi)$, we infer that

$$[\zeta \wedge \mathsf{Con}^{\beta}(\zeta)] = [\bot].$$

This is to say that $\neg\mathsf{Con}^{\beta+1}(\zeta)$.

For the **limit case** we let β be a limit ordinal and assume that for every $\gamma < \beta$, θ_γ is true. We want to show that θ_β is true. Let ζ be a sentence such that for every $\gamma < \beta$, $\zeta \vdash \theta_\gamma$. We want to show that $\mathsf{Con}^{\beta}(\zeta)$ implies $\mathsf{Con}^{\beta}(\zeta \wedge \neg f(\zeta))$. So assume that $\mathsf{Con}^{\beta}(\zeta)$, i.e., for every $\gamma < \beta, \mathsf{Con}^{\gamma}(\zeta)$. Let $\gamma < \beta$. Since β is a limit ordinal, $\gamma + 1 < \beta$. So by the inductive hypothesis $\theta_{\gamma+1}$ is true. That is, by the definition of $\theta_{\gamma+1}$,

$$\forall\varphi\big(\mathsf{Pr}(\varphi \to \theta_\gamma) \to (\mathsf{Con}^{\gamma}(\varphi) \to \mathsf{Con}^{\gamma}(\varphi \wedge \neg f(\varphi)))\big).$$

By instantiation, we infer that

$$\mathsf{Pr}(\zeta \to \theta_\gamma) \to (\mathsf{Con}^\gamma(\zeta) \to \mathsf{Con}^\gamma(\zeta \wedge \neg f(\zeta))).$$

Since $\zeta \vdash \theta_\gamma$ and $\mathsf{Con}^\gamma(\zeta)$, this means that $\mathsf{Con}^\gamma(\zeta \wedge \neg f(\zeta))$. Since γ was a generic ordinal less than β, we get that

$$\forall \gamma < \beta, \mathsf{Con}^\gamma(\zeta \wedge \neg f(\zeta)),$$

i.e., $\mathsf{Con}^\beta(\zeta)$. This completes the proof of the lemma. ❑

Theorem 2.3.9 (Restatement of Theorem 2.1.5)**.** *Let f be monotonic. Suppose that for all φ,*
(i) $\varphi \wedge \mathsf{Con}^\alpha(\varphi)$ implies $f(\varphi)$,
(ii) if $[f(\varphi)] \neq [\bot]$, then $f(\varphi)$ strictly implies $\varphi \wedge \mathsf{Con}^\beta(\varphi)$ for all $\beta < \alpha$.
Then for every true χ, there is a true ψ such that $\psi \vdash \chi$ and $[f(\psi)] = [\psi \wedge \mathsf{Con}^\alpha(\psi)]$.

Proof. Let χ be a true sentence. By the lemma, θ_α is true. So

$$\varphi := \chi \wedge \theta_\alpha$$

is true. We let

$$\psi := \varphi \wedge (f(\varphi) \to \mathsf{Con}^\alpha(\varphi)).$$

Note that $\psi \vdash \chi$. We now show that $[\psi \wedge \mathsf{Con}^\alpha(\psi)] = [f(\psi)]$.

Claim. $f(\psi) \vdash (\psi \wedge f(\varphi))$.

Since f is monotonic.

Claim. $(\psi \wedge f(\varphi)) \vdash (\varphi \wedge \mathsf{Con}^\alpha(\varphi))$.

By the definition of ψ.

Claim. $(\varphi \wedge \mathsf{Con}^\alpha(\varphi)) \vdash (\psi \wedge \mathsf{Con}^\alpha(\psi))$.

It is clear from the definition of ψ that $(\varphi \wedge \mathsf{Con}^\alpha(\varphi)) \vdash \psi$. So it suffices to show that $(\varphi \wedge \mathsf{Con}^\alpha(\varphi)) \vdash \mathsf{Con}^\alpha(\psi)$.

$$\begin{aligned}
\varphi \wedge \mathsf{Con}^\alpha(\varphi) &\vdash \forall \zeta \big(\big(\forall \beta < \alpha \mathsf{Pr}(\zeta \to \theta_\beta) \big) \to \big(\mathsf{Con}^\alpha(\zeta) \to \mathsf{Con}^\alpha(\zeta \wedge \neg f(\zeta)) \big) \big) \text{ by choice of } \varphi. \\
&\vdash \forall \beta < \alpha \mathsf{Pr}(\varphi \to \theta_\beta) \to \big(\mathsf{Con}^\alpha(\varphi) \to \mathsf{Con}^\alpha(\varphi \wedge \neg f(\varphi)) \big) \text{ by instantiation.} \\
&\vdash \forall \beta < \alpha \mathsf{Pr}(\varphi \to \theta_\beta) \to \mathsf{Con}^\alpha(\varphi \wedge \neg f(\varphi)) \text{ by logic.}
\end{aligned}$$

Since $\mathsf{Con}^\alpha(\varphi \wedge \neg f(\varphi)) \vdash \mathsf{Con}^\alpha(\psi)$, to prove the desired claim it suffices to show that

$$\varphi \wedge \mathsf{Con}^\alpha(\varphi) \vdash \forall \beta < \alpha \mathsf{Pr}(\varphi \to \theta_\beta).$$

We reason as follows.

$$\begin{aligned}
&\varphi \vdash \theta_\alpha \text{ by choice of } \varphi.\\
&\quad \vdash \forall \beta < \alpha(\mathsf{True}_{\Pi_3}\theta_\beta) \text{ by definition of } \theta_\alpha.\\
&\quad \vdash \mathsf{Pr}(\varphi \to \forall \beta < \alpha(\mathsf{True}_{\Pi_3}\theta_\beta)) \text{ by } \Sigma^0_1 \text{ completeness.}\\
&\quad \vdash \forall \beta < \alpha \mathsf{Pr}(\varphi \to \mathsf{True}_{\Pi_3}\theta_\beta)\\
&\quad \vdash \forall \beta < \alpha \mathsf{Pr}(\varphi \to \theta_\beta)
\end{aligned}$$

It is immediate from the preceding claims that $f(\psi) \vdash \psi \wedge \mathsf{Con}^\alpha(\psi)$. By assumption, $\psi + \mathsf{Con}^\alpha(\psi) \vdash f(\psi)$, so it follows that $[f(\psi)] = [\psi \wedge \mathsf{Con}^\alpha(\psi)]$. ❑

Corollary 2.3.10 (Restatement of Corollary 2.1.6)**.** *There is no monotonic f such that for all φ, if $[\varphi \wedge \mathsf{Con}^\alpha(\varphi)] \neq [\bot]$, then both*
(i) $\varphi \wedge \mathsf{Con}^\alpha(\varphi)$ strictly implies $f(\varphi)$ and
(ii) $f(\varphi)$ strictly implies $\varphi \wedge \mathsf{Con}^\beta(\varphi)$ for all $\beta < \alpha$.

2.4 Finite iterates of Con are inevitable

In this section and the next section we prove that the iterates of Con are, in a sense, inevitable. First we show that, for every natural number n, if a monotonic function f is always bounded by Con^n, then it is somewhere equivalent to Con^k for some $k \leqslant n$. In §5, we turn to generalizations of this result into the effective transfinite.

Theorem 2.4.1 (Restatement of Theorem 2.1.7)**.** *Let $n \in \mathbb{N}$. Let f be a monotonic function such that for every φ,*
(i) $\varphi \wedge \mathsf{Con}^n(\varphi)$ implies $f(\varphi)$ and
(ii) $f(\varphi)$ implies φ.
Then for some φ and some $k \leqslant n$, $[f(\varphi)] = [\varphi \wedge \mathsf{Con}^k(\varphi)] \neq [\bot]$.

Proof. We suppose, towards a contradiction, that there is no ψ and no $k \leqslant n$ such that $[f(\psi)] = [\psi \wedge \mathsf{Con}^k(\psi)] \neq [\bot]$. We then let φ_1 be a true statement such that

$$\begin{aligned}
&\varphi_1 \vdash \forall \zeta(\mathsf{Con}(\zeta) \to \mathsf{Con}(\zeta \wedge \neg f(\zeta)))\\
&\varphi_1 \vdash \forall k \forall \zeta\big(\mathsf{Con}^{k+1}(\zeta) \to \neg \mathsf{Pr}\big((\zeta \wedge \mathsf{Con}^k(\zeta)) \leftrightarrow f(\zeta)\big)\big).
\end{aligned}$$

The first condition is that φ_1 proves that for every consistent φ, $f(\varphi)$ strictly implies φ. The second condition is that φ_1 proves that $f(\zeta)$ never coincides with $\zeta \wedge \mathsf{Con}^k(\zeta)$, unless $[\zeta \wedge \mathsf{Con}^k(\zeta)] = [\bot]$.

We define a sequence of statements, starting with φ_1, as follows:

$$\varphi_{k+1} := \varphi_k \wedge (f(\varphi_k) \to \mathsf{Con}^k(\varphi_k)).$$

We will use our assumption to show that, for all k, $\varphi_k \wedge \mathsf{Con}^k(\varphi_k) \vdash \mathsf{Con}^k(\varphi_{k+1})$. From this we will deduce that $[f(\varphi_{n+1})] = [\varphi_{n+1} \wedge \mathsf{Con}^n(\varphi_{n+1})] \neq [\bot]$, contradicting the assumption that f and Con^n never coincide. Most of the work is contained in the proof of the following lemma.

Lemma 2.4.2. *For all k, for all $j \geqslant k$, $\big(\varphi_k \wedge \mathsf{Con}^k(\varphi_k)\big) \vdash \mathsf{Con}^k(\varphi_j)$.*

Proof. We prove the claim by a double induction. The primary induction is on k. For the **base case** $k = 1$, we prove the claim by induction on j. The *base case* $j = 1$ follows trivially. For the *inductive step* we assume that $\big(\varphi_1 \wedge \mathsf{Con}(\varphi_1)\big) \vdash \mathsf{Con}(\varphi_j)$ and show that $\big(\varphi_1 \wedge \mathsf{Con}(\varphi_1)\big) \vdash \mathsf{Con}(\varphi_{j+1})$.

$$\begin{aligned}
\varphi_1 \wedge \mathsf{Con}(\varphi_1) &\vdash \forall\zeta(\mathsf{Con}(\zeta) \to \mathsf{Con}(\zeta \wedge \neg f(\zeta))) \text{ by choice of } \varphi_1.\\
&\vdash \mathsf{Con}(\varphi_j) \to \mathsf{Con}(\varphi_j \wedge \neg f(\varphi_j)) \text{ by instantiation.}\\
\varphi_1 \wedge \mathsf{Con}(\varphi_1) &\vdash \mathsf{Con}(\varphi_j) \text{ by the inductive hypothesis.}\\
&\vdash \mathsf{Con}(\varphi_j \wedge \neg f(\varphi_j)) \text{ by logic.}\\
&\vdash \mathsf{Con}(\varphi_{j+1}) \text{ by definition of } \varphi_{j+1}.
\end{aligned}$$

For the **inductive step** we assume that the claim is true of $k-1$, i.e.,

$$\forall j \geqslant k-1\big(\big(\varphi_{k-1} \wedge \mathsf{Con}^{k-1}(\varphi_{k-1})\big) \vdash \big(\mathsf{Con}^{k-1}(\varphi_j)\big)\big).$$

We prove the claim for k. Once again, we prove the claim by induction on j. The *base case* $j = k$ follows trivially. For the *inductive step* we assume that $\varphi_k \wedge \mathsf{Con}^k(\varphi_k) \vdash \mathsf{Con}^k(\varphi_j)$. We want to prove that $\varphi_k \wedge \mathsf{Con}^k(\varphi_k) \vdash \mathsf{Con}^k(\varphi_{j+1})$.

$$\begin{aligned}
\varphi_k \wedge \mathsf{Con}^k(\varphi_k) &\vdash \forall x\forall\zeta\big(\mathsf{Con}^{x+1}(\zeta) \to \neg\mathsf{Pr}\big((\zeta \wedge \mathsf{Con}^x(\zeta)) \leftrightarrow f(\zeta)\big)\big) \text{ by choice of } \varphi_1.\\
&\vdash \mathsf{Con}^k(\varphi_j) \to \neg\mathsf{Pr}\big((\varphi_j \wedge \mathsf{Con}^{k-1}(\varphi_j)) \leftrightarrow f(\varphi_j)\big) \text{ by instantiation.}\\
\varphi_k \wedge \mathsf{Con}^k(\varphi_k) &\vdash \mathsf{Con}^k(\varphi_j) \text{ by the inner inductive hypothesis.}\\
&\vdash \neg\mathsf{Pr}\big((\varphi_j \wedge \mathsf{Con}^{k-1}(\varphi_j)) \leftrightarrow f(\varphi_j)\big) \text{ by logic.}
\end{aligned}$$

Thus, $\varphi_k \wedge \mathsf{Con}^k(\varphi_k)$ proves that one of the following cases holds.

$$(\varphi_j \wedge \mathsf{Con}^{k-1}(\varphi_j)) \nvdash f(\varphi_j)$$

$$f(\varphi_j) \nvdash (\varphi_j \wedge \mathsf{Con}^{k-1}\varphi_j)$$

We now show that $\varphi_k \wedge \mathsf{Con}^k(\varphi_k)$ refutes the second option.

Claim. $\varphi_k \wedge \mathsf{Con}^k(\varphi_k) \vdash \mathsf{Pr}\big(f(\varphi_j) \to (\varphi_j \wedge \mathsf{Con}^{k-1}\varphi_j)\big)$.

By the outer inductive hypothesis, EA proves the following conditional:

$$\theta := \big((\varphi_{j-1} \wedge \mathsf{Con}^{k-1}(\varphi_{j-1})) \to (\mathsf{Con}^{k-1}(\varphi_j))\big).$$

Thus, $f(\varphi_j)$ (which contains EA) also proves θ. We now show that $f(\varphi_j) \vdash \mathsf{Con}^{k-1}(\varphi_j)$.

$$\begin{aligned} f(\varphi_j) &\vdash \varphi_j \wedge f(\varphi_{j-1}) \text{ since } f \text{ is monotonic.} \\ &\vdash (\varphi_{j-1} \wedge (f(\varphi_{j-1}) \to \mathsf{Con}^{j-1}(\varphi_{j-1}))) \wedge f(\varphi_{j-1}) \text{ by the definition of } \varphi_j. \\ &\vdash \varphi_{j-1} \wedge \mathsf{Con}^{j-1}(\varphi_{j-1}) \text{ by logic.} \\ &\vdash \varphi_{j-1} \wedge \mathsf{Con}^{k-1}(\varphi_{j-1}) \text{ since } j \geqslant k. \\ &\vdash \mathsf{Con}^{k-1}(\varphi_j) \text{ since } f(\varphi_j) \text{ proves } \theta. \end{aligned}$$

By Σ^0_1 completeness, $(\varphi_k \wedge \mathsf{Con}^k(\varphi_k)) \vdash \mathsf{Pr}\big(f(\varphi_j) \to \mathsf{Con}^{k-1}(\varphi_j)\big)$.

Claim. $(\varphi_k \wedge \mathsf{Con}^k(\varphi_k)) \vdash \mathsf{Con}^k(\varphi_{j+1})$.

We reason as follows.

$$\begin{aligned} (\varphi_k \wedge \mathsf{Con}^k(\varphi_k)) &\vdash \neg\mathsf{Pr}\big((\varphi_j \wedge \mathsf{Con}^{k-1}(\varphi_j)) \to f(\varphi_j)\big) \text{ by the previous claim.} \\ &\vdash \mathsf{Con}(\varphi_j \wedge \neg f(\varphi_j) \wedge \mathsf{Con}^{k-1}(\varphi_j)). \\ &\vdash \mathsf{Con}(\varphi_{j+1} \wedge \mathsf{Con}^{k-1}(\varphi_j)) \text{ by the definition of } \varphi_{j+1}. \\ &\vdash \mathsf{Con}(\varphi_{j+1} \wedge \mathsf{Con}^{k-1}(\varphi_{j+1})) \text{ by the outer inductive hypothesis.} \\ &\vdash \mathsf{Con}^k(\varphi_{j+1}) \text{ by definition of } \mathsf{Con}^k. \end{aligned}$$

This concludes the proof of the lemma. ❑

As an instance of the lemma, we get that $(\varphi_n \wedge \mathsf{Con}^n(\varphi_n)) \vdash \mathsf{Con}^n(\varphi_{n+1})$. We reason as follows.

$$\begin{aligned} f(\varphi_{n+1}) &\vdash \varphi_n \wedge (f(\varphi_n) \to \mathsf{Con}^n(\varphi_n)) \text{ by the definition of } \varphi_{n+1}. \\ f(\varphi_{n+1}) &\vdash f(\varphi_n) \text{ since } f \text{ is monotonic.} \\ &\vdash \mathsf{Con}^n(\varphi_n) \text{ by logic.} \\ &\vdash \mathsf{Con}^n(\varphi_{n+1}) \text{ by the lemma.} \end{aligned}$$

On the other hand, $\varphi_{n+1} \wedge \mathsf{Con}^n(\varphi_{n+1}) \vdash f(\varphi_{n+1})$ since f is everywhere bounded by Con^n. Thus, $[f(\varphi_{n+1})] = [\varphi_{n+1} \wedge \mathsf{Con}^n(\varphi_{n+1})]$, contradicting the assumption that there is no ψ and no $k \leqslant n$ such that $[f(\psi)] = [\psi \wedge \mathsf{Con}^k(\psi)] \neq [\bot]$. ❑

2.5 Transfinite iterates of Con are inevitable.

Generalizing the proof of Theorem 2.4.1 into the transfinite poses the following difficulty. Recall that the proof of Theorem 2.4.1 makes use of a sequence of sentences starting with $\varphi_0 := \top$ where

$$\varphi_{k+1} := \varphi_k \wedge (f(\varphi_k) \to \mathsf{Con}^k(\varphi_k)).$$

It is not clear what the ωth sentence in the sequence should be. A natural idea is that for a limit ordinal λ the corresponding "limit sentence" should quantify over the sentences in the sequence beneath it and express, roughly,

$$\forall \gamma < \lambda \big(\mathsf{True}(\varphi_\gamma) \wedge (\mathsf{True}(f(\varphi_\gamma)) \rightarrow \mathsf{Con}^\gamma(\varphi_\gamma)) \big).$$

However, if the sentences in the sequence $(\varphi_\gamma)_{\gamma<\lambda}$ have unbounded syntactic complexity, then we are not guaranteed to have a truth-predicate with which we can quantify over them.

Nevertheless, we show that Theorem 2.4.1 generalizes into the transfinite given an additional assumption on complexity. Note that $\varphi \mapsto (\varphi \wedge \mathsf{Con}(\varphi))$ can be factored into two functions—the identity and $\varphi \mapsto \mathsf{Con}(\varphi)$—the latter of which always produces a Π^0_1 sentence. For the rest of this section, we will focus on monotonic functions $\varphi \mapsto \varphi \wedge f(\varphi)$ where f is monotonic and also $f(\varphi) \in \Pi^0_1$ for all φ.

Definition 2.5.1. A function f is Π^0_1 if $f(\varphi) \in \Pi^0_1$ for all φ.

For the next theorem we fix an elementary presentation Γ of a recursive well-ordering. In the statement of the theorem and throughout the proof α, β, γ, δ, etc. are names of ordinals from the notation system Γ.

Theorem 2.5.2 (Restatement of Theorem 2.1.9)**.** *Let f be a monotonic Π^0_1 function. Then either*
(i) for some $\beta \leqslant \alpha$ and some φ, $[\varphi \wedge f(\varphi)] = [\varphi \wedge \mathsf{Con}^\beta(\varphi)] \neq [\bot]$ or
(ii) for some φ, $(\varphi \wedge \mathsf{Con}^\alpha(\varphi)) \nvdash f(\varphi)$.

Proof. Let f be a monotonic Π^0_1 function such that for every φ,

$$(\varphi \wedge \mathsf{Con}^\alpha(\varphi)) \vdash (\varphi \wedge f(\varphi)).$$

We assume, for the sake of contradiction, that there is no sentence ζ and no $\beta \leqslant \alpha$ such that $[\zeta \wedge \mathsf{Con}^\beta(\zeta)] = [\zeta \wedge f(\zeta)] \neq [\bot]$. We then let φ be the conjunction of the following four sentences.

$$\forall \zeta (\mathsf{Con}(\zeta) \rightarrow \mathsf{Con}(\zeta \wedge \neg f(\zeta)))$$

$$\forall \beta \leqslant \alpha \forall \zeta \big(\mathsf{Con}^\beta(\zeta) \rightarrow \forall \delta < \beta, \neg \mathsf{Pr}\big((\zeta \wedge \mathsf{Con}^\delta(\zeta)) \leftrightarrow (\zeta \wedge f(\zeta))\big)\big)$$

$$\forall \zeta \forall \eta (\mathsf{Pr}(\zeta \rightarrow \eta) \rightarrow \mathsf{Pr}(f(\zeta) \rightarrow f(\eta)))$$

$$\forall x \big(\mathsf{Pr}(\mathsf{True}_{\Pi^0_2}(x)) \rightarrow \mathsf{True}_{\Pi^0_2}(x) \big)$$

The first expresses that for every consistent φ, $f(\varphi)$ strictly implies φ. The second sentence expresses that if $\beta < \alpha$, then $f(\zeta)$ and $\zeta \wedge \mathsf{Con}^\beta(\zeta)$ never coincide, unless $[\zeta \wedge \mathsf{Con}^\beta(\zeta)] = [\bot]$. The third sentence expresses the monotonicity of f. The fourth sentence expresses the Π^0_2 soundness of EA. Note that each of these sentences is true, so their conjunction φ is also true. Each of the four sentences is Π^0_2, whence so is φ.

We are interested in the following sequence $(\varphi_\beta)_{\beta\leqslant\Gamma}$. Note that the sentences in the sequence $(\varphi_\beta)_{\beta\leqslant\Gamma}$ all have complexity Π^0_2. Note moreover that since φ_1 is true, so is φ_β for every β.

$$\varphi_1 := \varphi.$$
$$\varphi_\gamma := \varphi_1 \wedge \forall\delta < \gamma\big(\mathsf{True}_{\Pi_1}(f(\varphi_\delta)) \to \mathsf{Con}^\delta(\varphi_\delta)\big) \text{ for } \gamma > 1.$$

Formally, we define the sequence $(\varphi_\beta)_{\beta\leqslant\Gamma}$ by Gödel's fixed point lemma as in Definition 2.3.6.

Remark 2.5.3. We may assume that the ordinal notation system Γ is provably linear in EA. Thus, $\mathsf{EA} \vdash \forall\beta \leqslant \alpha, \forall\gamma < \beta(\mathsf{True}_{\Pi_2}(\varphi_\beta) \to \mathsf{True}_{\Pi_2}(\varphi_\gamma))$.

Our goal is to show that

$$[\varphi_{\alpha+1} \wedge \mathsf{Con}^\alpha(\varphi_{\alpha+1})] = [\varphi_{\alpha+1} \wedge f(\varphi_{\alpha+1})]$$

contradicting the assumption that f and Con^α never coincide. The main lemmas needed to prove this result are the following.

Lemma 2.5.4. $\mathsf{EA} \vdash \forall\gamma \leqslant \alpha\mathsf{Pr}\big((\varphi_\gamma \wedge \neg f(\varphi_\gamma)) \to \varphi_\alpha\big)$.

Lemma 2.5.5. $\mathsf{EA} \vdash \forall\beta \leqslant \alpha\forall\gamma \leqslant \beta\mathsf{Pr}\big((\varphi_\beta \wedge \mathsf{Con}^\gamma(\varphi_\beta)) \to \mathsf{Con}^\gamma(\varphi_\beta \wedge \neg f(\varphi_\beta))\big)$.

Lemma 2.5.4 is needed to derive Lemma 2.5.5. We now show how we use Lemma 2.5.5 to derive Theorem 3.2.4. As an instance of Lemma 2.5.5, letting $\alpha = \beta = \gamma$, we infer that

$$\mathsf{EA} \vdash \mathsf{Pr}\big((\varphi_\alpha \wedge \mathsf{Con}^\alpha(\varphi_\alpha)) \to \mathsf{Con}^\alpha(\varphi_\alpha \wedge \neg f(\varphi_\alpha))\big).$$

From the soundness of EA, we infer that

$$\varphi_\alpha + \mathsf{Con}^\alpha(\varphi_\alpha) \vdash \mathsf{Con}^\alpha(\varphi_\alpha \wedge \neg f(\varphi_\alpha)). \qquad (\mp)$$

We then reason as follows.

$$\begin{aligned}
\varphi_{\alpha+1} &\vdash \varphi_\alpha \wedge (f(\varphi_\alpha) \to \mathsf{Con}^\alpha(\varphi_\alpha)) \text{ by the definition of } \varphi_{\alpha+1}.\\
f(\varphi_{\alpha+1}) &\vdash f(\varphi_\alpha) \text{ since } f \text{ is monotonic.}\\
\varphi_{\alpha+1} + f(\varphi_{\alpha+1}) &\vdash \varphi_\alpha \wedge \mathsf{Con}^\alpha(\varphi_\alpha) \text{ by logic.}\\
&\vdash \mathsf{Con}^\alpha(\varphi_{\alpha+1}) \text{ by } \mp.
\end{aligned}$$

On the other hand, $\varphi_{\alpha+1} + \mathsf{Con}^\alpha(\varphi_{\alpha+1}) \vdash f(\varphi_{\alpha+1})$ since f is everywhere bounded by Con^α. Since φ_1 is true, so too is $\varphi_{\alpha+1}$, whence we infer that

$$[\varphi_{\alpha+1} \wedge \mathsf{Con}^\alpha(\varphi_{\alpha+1})] = [\varphi_{\alpha+1} \wedge f(\varphi_{\alpha+1})] \neq [\bot],$$

contradicting the claim that there is no sentence ζ and no $\beta \leqslant \alpha$ such that $[\zeta \wedge \mathsf{Con}^\beta(\zeta)] = [\zeta \wedge f(\zeta)] \neq [\bot]$. ❑

It remains to prove Lemma 2.5.4 and Lemma 2.5.5. We devote one subsection to each.

Proof of Lemma 2.5.4

In this subsection we prove Lemma 2.5.4. First we recall the statement of the lemma.

Lemma 2.5.6 (Restatement of Lemma 2.5.4)**.**

$$\mathsf{EA} \vdash \forall \gamma \leqslant \alpha \big(\mathsf{Pr}(\varphi_\gamma \wedge \neg f(\varphi_\gamma)) \rightarrow \varphi_\alpha \big).$$

Proof. We reason in EA. Let $\gamma \leqslant \alpha$. We assume that

$$\mathsf{True}_{\Pi_2}(\varphi_\gamma) \wedge \neg \mathsf{True}_{\Pi_1}(f(\varphi_\gamma)). \tag{η}$$

We we want to derive φ_α, i.e.

$$\varphi_1 \wedge \forall \sigma < \alpha (\mathsf{True}_{\Pi_1}(f(\varphi_\sigma)) \rightarrow \mathsf{Con}^\sigma(\varphi_\sigma)).$$

The first conjunct follows trivially from the assumption that $\mathsf{True}_{\Pi_2}(\varphi_\gamma)$. We now prove the second conjunct of φ_α in two parts, first for all σ such that $\alpha > \sigma \geqslant \gamma$ and then for all $\sigma < \gamma$.

$\alpha > \sigma \geqslant \gamma$: From the assumption that $\mathsf{True}_{\Pi^0_2}(\varphi_\gamma)$ we infer that φ_1, whence we infer that f is monotonic. Thus, for all $\delta \geqslant \gamma$, $f(\varphi_\delta) \vdash f(\varphi_\gamma)$, i.e., $\mathsf{EA} \vdash \big(f(\varphi_\delta) \rightarrow f(\varphi_\gamma)\big)$. From φ_1 we also infer that EA is Π^0_2 sound, and so we infer that for all $\delta \geqslant \gamma$, $\mathsf{True}_{\Pi_1}(f(\varphi_\delta)) \rightarrow \mathsf{True}_{\Pi_1}(f(\varphi_\gamma))$. From the assumption that $\neg \mathsf{True}_{\Pi_1}(f(\varphi_\gamma))$ we then infer that for all $\delta \geqslant \gamma$, $\neg \mathsf{True}_{\Pi_1}(f(\varphi_\delta))$, whence for all $\delta \geqslant \gamma$, $\mathsf{True}_{\Pi_1}(f(\varphi_\delta)) \rightarrow \mathsf{Con}^\delta(\varphi_\delta)$.

$\sigma < \gamma$: By Remark 2.5.3, η implies that

$$\forall \sigma < \gamma (\mathsf{True}_{\Pi_1}\big(f(\varphi_\sigma)) \rightarrow \mathsf{Con}^\sigma(\varphi_\sigma)\big).$$

This completes the proof of Lemma 2.5.4. ❑

Proof of Lemma 2.5.5

In this subsection we prove Lemma 2.5.5. We recall the statement of Lemma 2.5.5.

Lemma 2.5.7 (Restatement of Lemma 2.5.5)**.**

$$\mathsf{EA} \vdash \forall \beta \leqslant \alpha \forall \gamma \leqslant \beta \mathsf{Pr}\big(\varphi_\beta + \mathsf{Con}^\gamma(\varphi_\beta) \rightarrow \mathsf{Con}^\gamma(\varphi_\beta \wedge \neg f(\varphi_\beta))\big).$$

The proof of this lemma is importantly different from the proof of Lemma 2.4.2. In particular, to push the induction through limit stages we need to know not only that the inductive hypothesis is true but also that it is provable in EA. We resolve this issue by using Schmerl's technique of *reflexive transfinite induction* (see Proposition 2.3.4).

In the proof of the lemma, we let $\mathcal{C}(\gamma, \delta)$ abbreviate the claim that

$$\varphi_\delta + \mathsf{Con}^\gamma(\varphi_\delta) \vdash \mathsf{Con}^\gamma(\varphi_\delta \wedge \neg f(\varphi_\delta)).$$

Proof. We want to show that

$$\mathsf{EA} \vdash \forall \beta \leqslant \alpha (\forall \gamma \leqslant \beta (\mathcal{C}(\gamma, \beta))).$$

By Proposition 2.3.4 it suffices to show that

$$\mathsf{EA} \vdash \forall \alpha (\mathsf{Pr}(\forall \beta \leqslant \alpha \forall \gamma \leqslant \beta \mathcal{C}(\gamma, \beta)) \to \forall \gamma \leqslant \alpha \mathcal{C}(\gamma, \alpha)).^2$$

Thus, we **reason in** EA and fix α. We assume that

$$\mathsf{Pr}(\forall \beta \leqslant \alpha, \forall \gamma \leqslant \beta, \mathcal{C}(\gamma, \beta)). \tag{$\triangle$}$$

We let $\gamma \leqslant \alpha$ and we want to show that $\mathcal{C}(\gamma, \alpha)$.

Since $\varphi_\alpha \vdash \varphi$ we infer that

$$\varphi_\alpha + \mathsf{Con}^\gamma(\varphi_\alpha) \vdash \forall \delta < \gamma, \neg \mathsf{Pr}\big((\varphi_\alpha \wedge \mathsf{Con}^\delta(\varphi_\alpha)) \leftrightarrow (\varphi_\alpha \wedge f(\varphi_\alpha))\big). \tag{$\sharp$}$$

We first note that both

$$\begin{aligned}
\varphi_\alpha &\vdash \forall \delta < \gamma (\mathsf{True}_{\Pi_1}(f(\varphi_\delta)) \to \mathsf{Con}^\delta(\varphi_\delta)) \text{ by the definition of } \varphi_\alpha \text{ and also} \\
\varphi_\alpha + f(\varphi_\alpha) &\vdash \forall \delta < \gamma \mathsf{Pr}(f(\varphi_\alpha) \to f(\varphi_\delta)) \text{ since } \varphi_1 \text{ proves the monotonicity of } f. \\
&\vdash \forall \delta < \gamma (f(\varphi_\alpha) \to \mathsf{True}_{\Pi_1}(f(\varphi_\delta))) \text{ since } \varphi_1 \text{ proves the } \Pi^0_2 \text{ soundness of } \mathsf{EA}. \\
&\vdash \forall \delta < \gamma, \mathsf{True}_{\Pi_1}(f(\varphi_\delta)) \text{ by logic.}
\end{aligned}$$

Thus, we may reason as follows.

$$\begin{aligned}
\varphi_\alpha + f(\varphi_\alpha) &\vdash \forall \delta < \gamma, \mathsf{Con}^\delta(\varphi_\delta) \\
&\vdash \forall \delta < \gamma, \mathsf{Con}^\delta(\varphi_\delta \wedge \neg f(\varphi_\delta))) \text{ since } (\triangle) \text{ delivers } \mathcal{C}(\delta, \delta). \\
&\vdash \forall \delta < \gamma, \mathsf{Con}^\delta(\varphi_\alpha) \text{ by Lemma 2.5.4.}
\end{aligned}$$

Thus, by Σ^0_1 completeness,

$$\mathsf{EA} \vdash \forall \delta < \gamma \mathsf{Pr}\big((\varphi_\alpha \wedge f(\varphi_\alpha)) \to \mathsf{Con}^\delta(\varphi_\alpha)\big).$$

Combined with $(\sharp)$, this delivers

$$\begin{aligned}
\varphi_\alpha + \mathsf{Con}^\gamma(\varphi_\alpha) &\vdash \forall \delta < \gamma \neg \mathsf{Pr}\big((\varphi_\alpha \wedge \mathsf{Con}^\delta(\varphi_\alpha)) \to f(\varphi_\alpha)\big). \\
&\vdash \forall \delta < \gamma, \mathsf{Con}(\varphi_\alpha \wedge \neg f(\varphi_\alpha) \wedge \mathsf{Con}^\delta(\varphi_\alpha)). \\
&\vdash \forall \delta < \gamma, \mathsf{Con}\big(\varphi_\alpha \wedge \neg f(\varphi_\alpha) \wedge \mathsf{Con}^\delta(\varphi_\alpha \wedge \neg f(\varphi_\alpha))\big) \text{ since } (\triangle) \text{ delivers } \mathcal{C}(\delta, \alpha). \\
&\vdash \mathsf{Con}^\gamma(\varphi_\alpha \wedge \neg f(\varphi_\alpha)).
\end{aligned}$$

This completes the proof of Lemma 2.5.5. ❑

[2]The reader might expect that we need to write "$\beta < \alpha$" instead of "$\beta \leqslant \alpha$" in the antecedent for this to match the statement of Proposition 2.3.4. However, it is clear from the proof of Proposition 2.3.4 that this suffices.

Theorem 3.2.4 shows the inevitability of the consistency operator. For a sufficiently constrained monotonic function f, f must coincide with an iterate of Con on some non-trivial sentence. However, it is not clear from the proofs of Theorem 2.4.1 or Theorem 3.2.4 that f must coincide with Con on a *true* sentence.

Question 2.5.8. *Let f be a monotonic Π^0_1 function. Suppose that for every φ,*

$$(\varphi \wedge \mathsf{Con}^\alpha(\varphi)) \vdash f(\varphi).$$

*Must there be some $\beta \leqslant \alpha$ and some **true** φ such that*

$$[\varphi \wedge f(\varphi)] = [\varphi \wedge \mathsf{Con}^\beta(\varphi)]?$$

2.6 1-consistency and iterated consistency

Just as the Π^0_1 fragments of natural theories can often be approximated by iterated consistency statements, the Π^0_2 fragments of natural theories can often be approximated by iterated 1-consistency statements. A theory T is *1-consistent* if $T + \mathrm{Th}_{\Pi^0_1}(\mathbb{N})$ is consistent. The 1-consistency of $\mathsf{EA} + \varphi$ can be expressed by the following Π^0_2 sentence, $1\mathsf{Con}(\varphi)$:

$$\forall x(\mathsf{True}_{\Pi^0_1}(x) \to \mathsf{Con}(\varphi \wedge \mathsf{True}_{\Pi^0_1}(x))).$$

In this section, we investigate the relationship between 1-consistency and iterated consistency. First, we show that $1\mathsf{Con}$ majorizes every iterate of Con^α.

Proposition 2.6.1 (Restatement of Proposition 2.1.11)**.** *For any elementary presentation α of a recursive well ordering, there is a true sentence φ such that for every ψ, if $\psi \vdash \varphi$, then $(\psi \wedge 1\mathsf{Con}(\psi))$ implies $(\psi \wedge \mathsf{Con}^\alpha(\psi))$. Moreover, if $[\psi \wedge \mathsf{Con}^\alpha(\psi)] \neq [\bot]$ then $(\psi \wedge 1\mathsf{Con}(\psi))$ strictly implies $(\psi \wedge \mathsf{Con}^\alpha(\psi))$.*

Proof. Let α be an elementary presentation of a recursive well-ordering. Let φ be a true sentence such that $\varphi \vdash \mathsf{TI}^\alpha_{\Pi^0_1}$, i.e., φ implies the validity of transfinite induction along α for Π^0_1 predicates. We prove that

$$(\varphi \wedge 1\mathsf{Con}(\varphi)) \vdash \mathsf{Con}^{\alpha+1}(\varphi).$$

Since $\varphi \wedge 1\mathsf{Con}(\varphi) \vdash \mathsf{TI}^\alpha_{\Pi^0_1}$, it suffices to show that:
Base case: $(\varphi \wedge 1\mathsf{Con}(\varphi)) \vdash \mathsf{Con}(\varphi)$
Successor case: $(\varphi \wedge 1\mathsf{Con}(\varphi)) \vdash \forall \beta < \alpha(\mathsf{Con}^\beta(\varphi) \to \mathsf{Con}^{\beta+1}(\varphi))$
Limit case: $(\varphi \wedge 1\mathsf{Con}(\varphi)) \vdash \forall \lambda\Big(lim(\lambda) \to \big((\forall \beta < \lambda \mathsf{Con}^\beta(\varphi)) \to \mathsf{Con}^\lambda(\varphi)\big)\Big)$

The **base case** and the **limit case** are both trivial. For the **successor case** we first note that by the definition of $1\mathsf{Con}(\varphi)$,

$$1\mathsf{Con}(\varphi) \vdash \forall x(\mathsf{True}_{\Pi^0_1}(x) \to \mathsf{Con}(\varphi \wedge \mathsf{True}_{\Pi^0_1}(x))),$$

and so by substituting $\mathsf{Con}^\beta(\varphi)$ in for x,

$$1\mathsf{Con}(\varphi) \vdash \mathsf{True}_{\Pi^0_1}(\mathsf{Con}^\beta(\varphi)) \to \mathsf{Con}(\varphi \wedge \mathsf{True}_{\Pi^0_1}(\mathsf{Con}^\beta(\varphi))). \qquad (\oplus)$$

Thus, we reason as follows.

$$\begin{aligned} 1\mathsf{Con}(\varphi) \vdash \mathsf{Con}^\beta(\varphi) &\to \mathsf{Con}(\varphi \wedge \mathsf{True}_{\Pi^0_1}(\mathsf{Con}^\beta(\varphi))) \text{ by } (\oplus). \\ &\to \mathsf{Con}(\varphi \wedge \mathsf{Con}^\beta(\varphi)). \\ &\to \mathsf{Con}^{\beta+1}(\varphi) \text{ by the definition of } \mathsf{Con}^{\beta+1}. \end{aligned}$$

It is clear that the implication $\varphi \wedge 1\mathsf{Con}(\varphi) \vdash \varphi \wedge \mathsf{Con}^\alpha(\varphi)$ is strict as long as $[\varphi \wedge \mathsf{Con}^\alpha(\varphi)] \neq [\bot]$. This completes the proof of the proposition. ❑

In light of the previous proposition, one might conjecture that $1\mathsf{Con}$ is the weakest monotonic function majorizing every function of the form Con^α for some recursive well-ordering α on true sentences. However, this is not so. To demonstrate this, we use a recursive linear order that has no hyperarithmetic infinite descending sequences. Harrison [43] introduced such an ordering with order-type $\omega_1^{CK} \times (1 + \mathbb{Q})$; see also Feferman and Spector [28] who consider such orderings in the context of iterated reflection principles. We use a presentation $\mathcal{H}$ of Harrison's ordering such satisfying the conditions explicated in Definition 2.3.1. We note that since $\mathcal{H}$ has no hyperarithmetic descending sequences, transfinite induction along $\mathcal{H}$ for Π^0_1 properties is valid. Our idea is to produce a function stronger than each Con^α but weaker than $1\mathsf{Con}$ by iterating Con along the Harrison linear order.

Theorem 2.6.2 (Restatement of Theorem 2.1.12)**.** *There are infinitely many monotonic functions f such that for every recursive ordinal α, there is an elementary presentation a of α such that f majorizes Con^a on a true ideal but also $1\mathsf{Con}$ majorizes f on a true ideal.*

Proof. In Definition 2.3.2, we used Gödel's fixed point lemma to produce iterates of Con along an elementary well-ordering. We similarly use Gödel's fixed point lemma to define sentences $\mathbf{Con}^\star(\varphi, \beta)$ for $\beta \in \mathcal{H}$ as follows.

$$\mathsf{EA} \vdash \mathbf{Con}^\star(\varphi, \beta) \leftrightarrow \forall \gamma <_{\mathcal{H}} \beta, \mathsf{Con}(\varphi \wedge \mathbf{Con}^\star(\varphi, \gamma)).$$

We use the notation $\mathsf{Con}^\beta(\varphi)$ for $\mathbf{Con}^\star(\varphi, \beta)$. Recall that we are assuming that it is elementarily calculable whether an element of $\mathcal{H}$ is zero or a successor or a limit. Thus, the following clauses are provable in EA.

- $\mathsf{Con}^0(\varphi) \leftrightarrow \top$
- $\mathsf{Con}^{\gamma+1}(\varphi) \leftrightarrow \mathsf{Con}(\varphi \wedge \mathsf{Con}^\gamma(\varphi))$
- $\mathsf{Con}^\lambda(\varphi) \leftrightarrow \forall \gamma <_{\mathcal{H}} \lambda, \mathsf{Con}^\gamma(\varphi)$ for λ a limit.

Claim. *For $\gamma \in \mathcal{H}$, the function $\varphi \mapsto Con^\gamma(\varphi)$ is monotonic.*

This follows immediately from Proposition 2.3.5. Note that in the statement of Lemma 2.3.4 we assume only that $<$ is an elementary *linear* ordering, not a well-ordering.

Claim. *There are infinitely many monotonic functions f such that for every recursive well-ordering α, there is an elementary presentation a of α such that f majorizes Con^a on true sentences.*

If $x <_{\mathcal{H}} y$ then $\mathsf{Con}^y(\varphi)$ strictly implies $\mathsf{Con}^x(\varphi)$ for every φ such that $\mathsf{Con}^x(\varphi) \neq [\bot]$. Given the order type of $\mathcal{H}$, this means that for infinitely many γ, for every recursive well-ordering α, Con^γ majorizes Con^a where a represents α in $\mathcal{H}$.

Claim. $1\mathsf{Con}$ *majorizes* Con^a *on true sentences for each* $a \in \mathcal{H}$.

Since every Π^0_1 definable subset of ω has an $\mathcal{H}$-least element, the sentence $\mathsf{TI}^{\mathcal{H}}_{\Pi^0_1}$, which expresses the validity of transfinite induction along $\mathcal{H}$ for Π^0_1 predicates, is true. But then if $\varphi \vdash \mathsf{TI}^{\mathcal{H}}_{\Pi^0_1}$, then for any $\gamma \in \mathcal{H}$, $(\varphi \wedge 1\mathsf{Con}(\varphi))$ strictly implies $(\varphi \wedge \mathsf{Con}^\gamma(\varphi))$ as long as $[(\varphi \wedge \mathsf{Con}^\gamma(\varphi))] \neq [\bot]$, as in Proposition 2.6.1. ❑

2.7 An unbounded recursively enumerable set that contains no true ideals

In this section we prove a limitative result. Theorem 4.1.2 demonstrates that if f is monotonic and that for all consistent φ, (i) $\varphi \wedge \mathsf{Con}(\varphi)$ implies $f(\varphi)$ and (ii) $f(\varphi)$ strictly implies φ, then for cofinally many true φ, $[f(\varphi)] = [\varphi \wedge \mathsf{Con}(\varphi)]$. It is natural to conjecture that cofinal equivalence with Con be strengthened to equivalence to Con *in the limit*, i.e., on a true ideal. One strategy to strengthen Theorem 4.1.2 in this way would be to show that every recursively enumerable set that contains arbitrarily strong true sentences and that is closed under provable equivalence contains a true ideal.

We now show that the aforementioned strategy fails. To this end, we define a recursively enumerable set $\mathcal{A}$ that contains arbitrarily strong true sentences and that is closed under provable equivalence but does not contain any true ideals. We are grateful to Matthew Harrison-Trainor for simplifying the proof of the following proposition.

Proposition 2.7.1 (Restatement of Proposition 2.1.13)**.** *There is a recursively enumerable set $\mathcal{A}$ that contains arbitrarily strong true sentences and that is closed under EA provable equivalence but does not contain any true ideals.*

Proof. Let $\{\varphi_0, \varphi_1, ...\}$ be an effective Gödel numbering of the language of arithmetic. We describe the construction of $\mathcal{A}$ in stages. During a stage n we may *activate* a sentence ψ, in which case we say that ψ is *active* until it is *deactivated* at some later stage $n + k$. After describing the construction of $\mathcal{A}$ we verify that A has the desired properties.

Stage 0: Numerate φ_0 and $\neg\varphi_0$ into $\mathcal{A}$. Activate the sentences $(\varphi_0 \wedge \mathsf{Con}(\varphi_0))$ and $(\neg\varphi_0 \wedge \mathsf{Con}(\neg\varphi_0))$.

Stage n+1: There are finitely many active sentences. For each such sentence ψ, numerate $\theta_0 := (\psi \wedge \varphi_{n+1})$ and $\theta_1 := (\psi \wedge \neg\varphi_{n+1})$ into $\mathcal{A}$. Deactivate the sentence ψ and activate the sentences $(\theta_0 \wedge \mathsf{Con}(\theta_0))$ and $(\theta_1 \wedge \mathsf{Con}(\theta_1))$.

We dovetail the construction with a search through EA proofs. If we ever see that $\mathsf{EA} \vdash \varphi \leftrightarrow \psi$ for some φ that we have already numerated into $\mathcal{A}$, then we numerate ψ into $\mathcal{A}$.

Now we check that $\mathcal{A}$ has the desired properties. It is clear that $\mathcal{A}$ is recursively enumerable and that $\mathcal{A}$ is closed under EA provable equivalence.

Claim. *$\mathcal{A}$ contains arbitrarily strong true sentences. That is, for each true sentence φ, there is a true sentence ψ such that $\psi \vdash \varphi$ and $\psi \in \mathcal{A}$.*

At any stage in the construction of $\mathcal{A}$, there are finitely many active sentences, ψ_0, ..., ψ_k. An easy induction shows that exactly one of $\psi_0, ..., \psi_k$ is true. Indeed, exactly one of φ_0 or $\neg\varphi_0$ is true, and hence so is exactly one of $\varphi_0 \wedge \mathsf{Con}(\varphi_0)$ and $\neg\varphi_0 \wedge \mathsf{Con}(\neg\varphi_0)$. And if θ is true, then so is exactly one of $\zeta_0 := \theta \wedge \varphi_k$ and $\zeta_1 := \theta \wedge \neg\varphi_k$, and hence so too is exactly one of $\zeta_0 \wedge \mathsf{Con}(\zeta_0)$ and $\zeta_1 \wedge \mathsf{Con}(\zeta_1)$.

Let φ_k be a true sentence. At stage k in the construction of $\mathcal{A}$ there are only finitely many active sentences $\psi_0, ..., \psi_n$. We have already seen that exactly one of ψ_i is true. But then $\varphi_k \wedge \psi_i$ is true, $(\varphi_k \wedge \psi_i \vdash \varphi_k)$, and $(\varphi_k \wedge \psi_i)$ is numerated into $\mathcal{A}$.

Claim. *$\mathcal{A}$ contains no true ideals.*

An easy induction shows that if ψ_0 and ψ_1 are both active at the same stage, then for any θ, if θ implies both ψ_0 and ψ_1 then $\theta \in [\bot]$.

Let φ be a true sentence in $\mathcal{A}$. By the previous remark, the only sentences in $\mathcal{A}$ that strictly imply φ are (i) EA refutable sentences and (ii) sentences that imply $\varphi \wedge \mathsf{Con}(\varphi)$. Since the Lindenbaum algebra of EA is dense, this means there is some ψ such that $(\varphi \wedge \mathsf{Con}(\varphi))$ strictly implies ψ strictly implies φ but $\psi \notin \mathcal{A}$. ❑

The following questions remain.

Question 2.7.2. *Is the relation of cofinal agreement on true sentences an equivalence relation on recursive monotonic operators?*

Question 2.7.3. *Let f be recursive and monotonic. Suppose that for all consistent φ,*
(i) $\varphi \wedge \mathsf{Con}(\varphi)$ implies $f(\varphi)$ and
(ii) $f(\varphi)$ implies φ.
Must f be equivalent to the identity or to Con on a true ideal?

Chapter 3

A note on the consistency operator

3.1 Introduction

Gödel's second incompleteness theorem states that no consistent sufficiently strong effectively axiomatized theory T proves its own consistency statement Con_T. Using ad hoc proof-theoretic techniques (namely, Rosser-style self-reference) one can construct Π_1 sentences φ that are not provable in T such that $T + \varphi$ is a strictly weaker theory than $T + \mathsf{Con}_T$. Nevertheless, Con_T seems to be the weakest *natural* Π_1 sentence that is not provable in T. Without a mathematical definition of "natural," however, it is difficult to formulate a precise conjecture that would explain this phenomenon. This is a special case of the well known empirical observation that natural axiomatic theories are pre-well-ordered by consistency strength, which S. Friedman, Rathjen, and Weiermann [33] call one of the "great mysteries in the foundations of mathematics."

Recursion theorists have observed a similar phenomenon in Turing degree theory. One can use ad hoc recursion-theoretic methods like the priority method to construct non-recursive Σ_1 definable sets whose Turing degree is strictly below that of $0'$. Nevertheless, $0'$ seems to be the weakest *natural* non-recursive r.e. degree. Once again, without a mathematical definition of "natural," however, it is difficult to formulate a precise conjecture that would explain this phenomenon.

A popular approach to studying natural Turing degrees is to focus on degree-invariant functions; a function f on the reals is *degree-invariant* if, for all reals A and B, $A \equiv_T B$ implies $f(A) \equiv_T f(B)$. The definitions of natural Turing degrees tend to relativize to arbitrary degrees, yielding degree invariant functions on the reals; for instance, the construction of $0'$ relativizes to yield the Turing Jump. Sacks [85] asked whether there is a degree invariant solution to Post's Problem. Recall that a function $W : 2^\omega \to 2^\omega$ is a *recursively enumerable operator* if there is an $e \in \omega$ such that, for each A, $W(A) = W_e^A$, the e^{th} set recursively enumerable in A.

Question 3.1.1 (Sacks)**.** *Is there a degree-invariant recursively enumerable operator W such that for every real A, $A <_T W_e^A <_T A'$?*

Though the question remains open, Slaman and Steel [94] proved that there is no order-preserving solution to Post's Problem. Recall that a function f on the reals is *order-preserving* if, for all reals A and B, $A \leqslant_T B$ implies $f(A) \leqslant_T f(B)$.

In [68], Montalbán and the author proved a proof-theoretic analogue of a negative answer to Sacks' question for order-preserving functions. Let T be a sound, sufficiently strong effectively axiomatized theory in the language of arithmetic, e.g., EA.[1] A function $\mathfrak{g}$ is *monotone* if, for all sentences φ and ψ, $T \vdash \varphi \to \psi$ implies $T \vdash \mathfrak{g}(\varphi) \to \mathfrak{g}(\psi)$ (this is just to say that $\mathfrak{g}$ induces a monotone function on the Lindenbaum algebra of T). Let $[\varphi]$ denote the equivalence class of φ modulo T provable equivalence, i.e., $[\varphi] := \{\psi : T \vdash \varphi \leftrightarrow \psi\}$. One of the main theorems of [68] is the following.

Theorem 3.1.2 (Montalbán–W.)**.** *Let $\mathfrak{g}$ be recursive and monotone such that:*

- *for all φ, $T + \mathsf{Con}_T(\varphi) \vdash \mathfrak{g}(\varphi)$*
- *for all consistent φ, $T + \varphi \nvdash \mathfrak{g}(\varphi)$*

Then for every true φ, there is a true ψ such that $T + \psi \vdash \varphi$ and

$$[\psi \wedge \mathfrak{g}(\psi)] = [\psi \wedge \mathsf{Con}_T(\psi)].$$

To state a corollary of this theorem, we recall that φ *strictly implies* ψ if one of the following holds:

(i) $T + \varphi \vdash \psi$ and $T + \psi \nvdash \varphi$.

(ii) $[\varphi] = [\psi] = [\bot]$.

Corollary 3.1.3. *There is no recursive monotone $\mathfrak{g}$ such that for every φ, $\big(\varphi \wedge \mathsf{Con}_T(\varphi)\big)$ strictly implies $\big(\varphi \wedge \mathfrak{g}(\varphi)\big)$ and $\big(\varphi \wedge \mathfrak{g}(\varphi)\big)$ strictly implies φ.*

The Slaman–Steel theorem suggests a strengthening of these results. Recall that a *cone* in the Turing degrees is any set of the form $\{B : B \geqslant_T A\}$ where A is a Turing degree. The following is a special case of a theorem due to Slaman and Steel.

Theorem 3.1.4 (Slaman–Steel)**.** *Let $f : 2^\omega \to 2^\omega$ be Borel and order-preserving. Then one of the following holds:*

1. *$f(A) \leqslant_T A$ on a cone.*
2. *$A' \leqslant_T f(A)$ on a cone.*

[1] EA is a theory in the language of arithmetic (with exponentiation) axiomatized by the axioms of Robinson's Q, recursive axioms for exponentiation, and induction for bounded formulas.

Montalbán and the author asked whether Theorem 3.1.2 could be strengthened in the style of the Slaman–Steel theorem, i.e., by showing that all increasing monotone recursive functions that are no stronger than the consistency operator are equivalent to the consistency operator in the limit. In this note we provide a positive answer to this question.

To sharpen the notion of the "limit behavior" of a function, we introduce the notion of a true cone. A *cone* is any set $\mathfrak{C}$ of the form $\{\psi : T+\psi \vdash \varphi\}$ where φ is a sentence. A *true cone* is a cone that contains a true sentence. In §2 we prove that all recursive monotone operators that produce sentences of some bounded arithmetical complexity are bounded from below by the consistency operator on a true cone.

Theorem 3.1.5. *Let $\mathfrak{g}$ be recursive and monotone such that, for some $k \in \mathbb{N}$, for all φ, $\mathfrak{g}(\varphi)$ is Π_k. Then one of the following holds:*

1. *There is a true cone $\mathfrak{C}$ such that for all $\varphi \in \mathfrak{C}$,*

$$T + \varphi \vdash \mathfrak{g}(\varphi).$$

2. *There is a true cone $\mathfrak{C}$ such that for all $\varphi \in \mathfrak{C}$,*

$$T + \varphi + \mathfrak{g}(\varphi) \vdash \mathsf{Con}_T(\varphi).$$

In §3 we prove that the condition that $\mathfrak{g}$ is recursive cannot be weakened. More precisely, we exhibit a monotone $0'$ recursive function which vacillates between behaving like the identity operator and behaving like the consistency operator.

Theorem 3.1.6. *There is a $0'$ recursive monotone function $\mathfrak{g}$ such that, for every φ, $\mathfrak{g}(\varphi)$ is Π_1, yet for arbitrarily strong true sentences*

$$[\varphi \wedge \mathfrak{g}(\varphi)] = [\varphi \wedge \mathsf{Con}_T(\varphi)]$$

and for arbitrarily strong true sentences

$$[\varphi \wedge \mathfrak{g}(\varphi)] = [\varphi].$$

Though Theorem 3.1.5 is a considerable strengthening of the result in [68], we conjecture that it admits of a dramatic improvement. We remind the reader that the aforementioned theorem of Slaman and Steel is a special case of a sweeping classification of increasing Borel order-preserving function. We say that a function f is *increasing* if, for all A, $A \leqslant_T f(A)$.

Theorem 3.1.7 (Slaman–Steel)**.** *Let $f : 2^\omega \to 2^\omega$ be increasing, Borel, order-preserving. Suppose that for some $\alpha < \omega_1$, $f(A) \leqslant_T A^{(\alpha)}$ for every A. Then for some $\beta \leqslant \alpha$, $f(A) \equiv_T A^{(\beta)}$ on a cone.*

We conjecture that a similar classification of monotone proof-theoretic operators is possible. Our conjecture is stated in terms of iterated consistency statements. Let $\prec$ be a nice elementary presentation of a recursive well-ordering.[2] We define the iterates of the consistency operator by appealing to Gödel's fixed point lemma.

$$T \vdash \mathsf{Con}^{\alpha}_{T}(\varphi) \leftrightarrow \forall \beta \prec \alpha \mathsf{Con}_{T}\big(\varphi \wedge \mathsf{Con}^{\beta}_{T}(\varphi)\big)$$

For true φ, the iterations of Con_T form a proper hierarchy of true sentences by Gödel's second incompleteness theorem. We make the following conjecture.

Conjecture 3.1.8. *Suppose $\mathfrak{g}$ is monotone, non-constant, and recursive such that, for every φ, $\mathfrak{g}(\varphi) \in \Pi_1$. Let $\prec$ be a nice elementary presentation of well-ordering and α an ordinal notation. Suppose that, for every φ,*

$$T + \varphi + \mathsf{Con}^{\alpha}_{T}(\varphi) \vdash \mathfrak{g}(\varphi).$$

Then for some $\beta \preceq \alpha$, for all φ in a true cone,

$$[\varphi \wedge \mathfrak{g}(\varphi)] = [\varphi \wedge \mathsf{Con}^{\beta}_{T}(\varphi)].$$

According to the conjecture, if an increasing monotone recursive function $\mathfrak{g}$ that produces only Π_1 sentences is no stronger than $\mathsf{Con}^{\alpha}_{T}$, it is equivalent on a true cone to Con^{β}_{T} for some $\beta \preceq \alpha$. This would provide a classification of a large class of monotone proof-theoretic operators in terms of their limit behavior.

3.2 The main theorem

Let T be a sound, recursively axiomatized extension of EA in the language of arithmetic. We want to show that $T + \mathsf{Con}_T(\varphi)$ is the weakest natural theory that results from adjoining a Π_1 sentence to T. A central notion in our approach is that of a monotone operator on finite extensions of T.

Definition 3.2.1. $\mathfrak{g}$ is *monotone* if, for every φ and ψ,

$$T \vdash \varphi \rightarrow \psi \text{ implies } T \vdash \mathfrak{g}(\varphi) \rightarrow \mathfrak{g}(\psi).$$

Remark 3.2.2. We will switch quite frequently using the notation $T + \varphi \vdash \psi$ and $T \vdash \varphi \rightarrow \psi$, trusting that no confusion arises. The two claims are equivalent, by the Deduction Theorem.

Our goal is to prove that the consistency operator is, roughly, the weakest operator for uniformly strengthening theories. Our strategy is to show that any uniform method for extending theories that is as weak as the consistency operator must be equivalent to the consistency operator in the limit. We sharpen the notion "in the limit" with the following definitions.

[2] Nice elementary presentations of well-orderings are defined in [6], see §2.3, Definition 1.

Definition 3.2.3. Given a sentence φ, the *cone generated by* φ is the set of all sentences ψ such that $T \vdash \psi \to \varphi$. A *cone* is any set $\mathfrak{C}$ such that, for some φ, $\mathfrak{C}$ is the cone generated by φ. A *true cone* is a cone that is generated by a sentence that is true in the standard model $\mathbb{N}$.

We are now ready to state and prove the main theorem. Note that the following is a restatement of Theorem 3.1.5.

Theorem 3.2.4. *Let T be a sound, effectively axiomatized extension of EA. Let $\mathfrak{g}$ be recursive and monotone such that, for some $k \in \mathbb{N}$, for all φ, $\mathfrak{g}(\varphi)$ is Π_k. Then one of the following holds:*

1. *There is a true cone $\mathfrak{C}$ such that for all $\varphi \in \mathfrak{C}$,*
$$T + \varphi \vdash \mathfrak{g}(\varphi).$$
2. *There is a true cone $\mathfrak{C}$ such that for all $\varphi \in \mathfrak{C}$,*
$$T + \varphi + \mathfrak{g}(\varphi) \vdash \mathsf{Con}_T(\varphi).$$

Proof. Since $\mathfrak{g}$ is recursive, its graph is defined by a Σ_1 formula $\mathcal{G}$, i.e., for any φ and ψ,
$$\mathfrak{g}(\varphi) = \psi \iff \mathbb{N} \vDash \mathcal{G}(\ulcorner\varphi\urcorner, \ulcorner\psi\urcorner).$$
Since T is sound and Σ_1 complete, this implies that for any φ and ψ,
$$\mathfrak{g}(\varphi) = \psi \iff T \vdash \mathcal{G}(\ulcorner\varphi\urcorner, \ulcorner\psi\urcorner) \qquad (\star)$$
From now on we drop the corner quotes and write $\mathcal{G}(\varphi, \psi)$ instead of $\mathcal{G}(\ulcorner\varphi\urcorner, \ulcorner\psi\urcorner)$, trusting that no confusion will arise.

We consider the following sentence in the language of arithmetic:
$$\forall x \forall y \Big(\big(\mathcal{G}(x, y) \wedge \mathsf{True}_{\Pi_k}(y)\big) \to \mathsf{Con}_T(x) \Big). \qquad (A)$$

Informally, A says that, for every φ, the truth of $\mathfrak{g}(\varphi)$ implies the consistency of $T + \varphi$. Note that we need to use a partial truth predicate in the statement A since we are formalizing a uniform claim about the outputs of the function $\mathfrak{g}$. For any specific output ψ of the function $\mathfrak{g}$, T will be able to reason about ψ without relying on the partial truth predicate.

We break into cases based on whether A is true or false.

Case 1: A is true in the standard model $\mathbb{N}$. We claim that in this case
$$\mathfrak{C} := \{\varphi : T + \varphi \vdash A\}$$
satisfies condition (2) from the statement of the theorem. Clearly $\mathfrak{C}$ is a true cone. It suffices to show that for any $\varphi \in \mathfrak{C}$,
$$T + \varphi + \mathfrak{g}(\varphi) \vdash \mathsf{Con}_T(\varphi).$$

So let $\varphi \in \mathfrak{C}$ and let $\psi = \mathfrak{g}(\varphi)$. We reason as follows.

$$\begin{aligned}
T + \varphi &\vdash \forall x \forall y \Big(\big(\mathcal{G}(x,y) \wedge \mathsf{True}_{\Pi_k}(y) \big) \to \mathsf{Con}_T(x) \Big) \text{ by choice of } \mathfrak{C}. \\
T + \varphi &\vdash \big(\mathcal{G}(\varphi, \psi) \wedge \mathsf{True}_{\Pi_k}(\psi) \big) \to \mathsf{Con}_T(\varphi) \text{ by instantiation.} \\
T + \varphi &\vdash \mathcal{G}(\varphi, \psi) \text{ by observation } (\star). \\
T + \varphi &\vdash \mathsf{True}_{\Pi_k}(\psi) \to \mathsf{Con}_T(\varphi) \text{ from the previous two lines by logic.} \\
T + \varphi + \psi &\vdash \mathsf{Con}_T(\varphi) \text{ trivially from the previous line.} \\
T + \varphi + \mathfrak{g}(\varphi) &\vdash \mathsf{Con}_T(\varphi) \text{ since } \psi = \mathfrak{g}(\varphi).
\end{aligned}$$

Case 2: A is false in the standard model $\mathbb{N}$. We infer that

$$\exists \varphi \exists \psi \Big(\mathcal{G}(\varphi, \psi) \wedge \mathsf{True}_{\Pi_k}(\psi) \wedge \neg \mathsf{Con}_T(\varphi) \Big).$$

Thus, there is an inconsistent sentence φ such that $\mathfrak{g}(\varphi)$ is a true Π_k sentence. This is to say that $\mathfrak{g}(\bot)$ is true. We claim that in this case

$$\mathfrak{C} := \{ \varphi : T + \varphi \vdash \mathfrak{g}(\bot) \}$$

satisfies condition (1) from the statement of the theorem. Clearly $\mathfrak{C}$ is a true cone. It suffices to show that for any $\varphi \in \mathfrak{C}$,

$$T + \varphi \vdash \mathfrak{g}(\varphi).$$

So let $\varphi \in \mathfrak{C}$. By the definition of $\mathfrak{C}$, we infer that

$$T + \varphi \vdash \mathfrak{g}(\bot). \tag{$\dagger$}$$

We reason as follows.

$$\begin{aligned}
T &\vdash \bot \to \varphi \text{ by logic.} \\
T &\vdash \mathfrak{g}(\bot) \to \mathfrak{g}(\varphi) \text{ by the monotonicity of } \mathfrak{g}. \\
T + \varphi &\vdash \mathfrak{g}(\varphi) \text{ from the previous line and } \dagger, \text{ by logic.}
\end{aligned}$$

This completes the proof. ❑

Remark 3.2.5. Theorem 3.2.4 is stated about operators $\mathfrak{g}$ that produce sentences of bounded arithmetical complexity, i.e., for some $k \in \mathbb{N}$, for all φ, $\mathfrak{g}(\varphi)$ is Π_k. The reason for this restriction is to invoke the partial truth-predicate for Π_k sentences when providing the sentence A.

3.3 Recursiveness is a necessary condition

In the proof of Theorem 3.2.4 we appealed to the recursiveness of $\mathfrak{g}$ to show that T correctly calculates the values of $\mathfrak{g}$, i.e., that for every φ and ψ,

$$\mathfrak{g}(\varphi) = \psi \iff T \vdash \mathcal{G}(\ulcorner\varphi\urcorner, \ulcorner\psi\urcorner).$$

In this section we show that recursiveness is a necessary condition for the proof of Theorem 3.2.4. In particular, we exhibit a monotone operator $\mathfrak{g}$ which is recursive in $0'$ and produces only Π_1 sentences such that for arbitrarily strong true sentences φ,

$$[\varphi \wedge \mathfrak{g}(\varphi)] = [\varphi \wedge \mathsf{Con}_T(\varphi)]$$

and for arbitrarily strong true sentences φ,

$$[\varphi \wedge \mathfrak{g}(\varphi)] = [\varphi].$$

Our proof makes use of a recursive set $\mathfrak{A}$ that contains arbitrarily strong true sentences and omits arbitrarily strong true sentences. A very similar set is constructed in [68]. We now present the construction of the set $\mathfrak{A}$, which is necessary to understand the proof of the theorem. After describing the construction of $\mathfrak{A}$ we will verify some of its basic properties.

> Let $\{\varphi_0, \varphi_1, ...\}$ be an effective Gödel numbering of the language of arithmetic. We describe the construction of $\mathfrak{A}$ in stages. During a stage n we may *activate* a sentence ψ, in which case we say that ψ is *active* until it is *deactivated* at some later stage.
>
> **Stage 0:** Numerate φ_0 and $\neg\varphi_0$ into $\mathfrak{A}$. Activate the sentences $(\varphi_0 \wedge \mathsf{Con}_T(\varphi_0))$ and $(\neg\varphi_0 \wedge \mathsf{Con}_T(\neg\varphi_0))$.
>
> **Stage n+1:** There are finitely many active sentences. For each such sentence ψ, numerate $\theta_0 := (\psi \wedge \varphi_{n+1})$ and $\theta_1 := (\psi \wedge \neg\varphi_{n+1})$ into $\mathfrak{A}$. Deactivate the sentence ψ and activate the sentences $(\theta_0 \wedge \mathsf{Con}_T(\theta_0))$ and $(\theta_1 \wedge \mathsf{Con}_T(\theta_1))$.

Remark 3.3.1. It can be useful to visualize, along with the construction of $\mathfrak{A}$, the construction of an upwards growing tree that is (at most) binary branching. The nodes in the tree are the *consistent* sentences that are numerated into $\mathfrak{A}$. The immediate successors in this tree of a sentence φ have the form $(\varphi \wedge \mathsf{Con}_T(\varphi) \wedge \theta)$ and $(\varphi \wedge \mathsf{Con}_T(\varphi) \wedge \neg\theta)$. Thus, the successors of any two points are inconsistent with each other. Observe that for any two distinct sentences φ and ψ in the tree, φ is below ψ (i.e., φ and ψ belong to the same path and φ is below ψ) if and only if $T + \psi \vdash \varphi$. It follows from the previous two observations that any two sentences that are incompatible with each other in the tree ordering are inconsistent with each other.

Lemma 3.3.2. *At any stage in the construction of $\mathfrak{A}$, (i) exactly one of the active sentences is true in the standard model $\mathbb{N}$ and (ii) exactly one of the sentences numerated into $\mathfrak{A}$ is true in the standard model $\mathbb{N}$.*

Proof. We proceed by induction on the stages in the construction of $\mathfrak{A}$.

Stage 0: Exactly one of φ_0 or $\neg\varphi_0$ is true (these are the numerated sentences), and hence so is exactly one of $\varphi_0 \wedge \mathsf{Con}(\varphi_0)$ and $\neg\varphi_0 \wedge \mathsf{Con}(\neg\varphi_0)$ (these are the activated sentences).

Stage n+1: At the end of stage n there is exactly one true activated sentence θ. Then so exactly one of $\zeta_0 := \theta \wedge \varphi_n$ and $\zeta_1 := \theta \wedge \neg\varphi_n$ is true (these are the numerated sentences). Hence exactly one of $\zeta_0 \wedge \mathsf{Con}(\zeta_0)$ and $\zeta_1 \wedge \mathsf{Con}(\zeta_1)$ is true (these are the activated sentences). ❑

Corollary 3.3.3. *There is a unique branch through the tree described in Remark 3.3.1 that contains only true sentences. We will call it the* true branch.

Lemma 3.3.4. $\mathfrak{A}$ *contains arbitrarily strong true sentences.*

Proof. Let ψ be a true sentence. ψ appears at some point in our Gödel numbering of the language of arithmetic, i.e., for some n, ψ is φ_n. Going into stage n of the construction of $\mathfrak{A}$, there is exactly one true active sentence θ by Lemma 3.3.2. Then $\theta \wedge \varphi_n$ is numerated into $\mathfrak{A}$. So $\mathfrak{A}$ contains a true sentence that implies ψ. ❑

Our proof also makes use of iterated consistency statements. Let $\prec$ be an elementary presentation of ω. For the sake of convenience, we reiterate the definition of the iterates of the consistency operator. We define these iterates by appealing to Gödel's fixed point lemma:

$$T \vdash \mathsf{Con}^{\alpha}_{T}(\varphi) \leftrightarrow \forall\beta \prec \alpha \mathsf{Con}_{T}\big(\varphi \wedge \mathsf{Con}^{\beta}_{T}(\varphi)\big)$$

For true φ, the iterates of Con_T form a proper hierarchy of true sentences by Gödel's second incompleteness theorem.

Definition 3.3.5. For a true sentence ψ numerated into $\mathfrak{A}$ at stage n, let θ_ψ be a true sentence that is either the $(n+1)^{th}$ sentence in the Gödel numbering of the language or the negation thereof (depending on which is true). The point of the definition is this: if ψ is a true sentence numerated into $\mathfrak{A}$, then the next true sentence numerated into $\mathfrak{A}$ is $\psi \wedge \mathsf{Con}_T(\psi) \wedge \theta_\psi$.

Lemma 3.3.6. *For arbitrarily strong true sentences* $\psi \in \mathfrak{A}$, $T \nvdash (\psi \wedge \mathsf{Con}_T(\psi)) \rightarrow \theta_\psi$.

Proof. Suppose not, i.e., suppose that there is a true φ such that for all true ψ, if both $T \vdash \psi \rightarrow \varphi$ and $\psi \in \mathfrak{A}$, then:

$$T \vdash (\psi \wedge \mathsf{Con}_T(\psi)) \rightarrow \theta_\psi. \qquad (\oplus)$$

By Lemma 3.3.4, $\mathfrak{A}$ contains arbitrarily strong true sentences, so we know there is at least one such sentence ψ_0 in $\mathfrak{A}$ that implies φ. By the construction of $\mathfrak{A}$, the true sentences numerated into $\mathfrak{A}$ after ψ_0 are:

- $\psi_1 := \psi_0 \wedge \mathsf{Con}_T(\psi_0) \wedge \theta_{\psi_0}$

- $\psi_2 := \psi_1 \wedge \mathsf{Con}_T(\psi_1) \wedge \theta_{\psi_1}$

and so on. Each ψ_n implies φ and is in $\mathfrak{A}$. Thus, each ψ_n satisfies condition ($\oplus$), i.e., for each ψ_n, $T \vdash (\psi_n \wedge \mathsf{Con}_T(\psi_n)) \rightarrow \theta_{\psi_n}$. This means that for all $n \geqslant 1$ the final conjunct of ψ_n is superfluous. It follows that for each n, ψ_n is T provably equivalent to $\psi_0 \wedge \mathsf{Con}_T^n(\psi_0)$. But then no sentence in $\mathfrak{A}$ is stronger than $\psi_0 \wedge \mathsf{Con}_T^\omega(\psi_0)$, contradicting the fact proved in Lemma 3.3.4, i.e., that $\mathfrak{A}$ contains arbitrarily strong true sentences. ❑

We are now ready to state and prove the theorem. Note that the following is a restatement of Theorem 3.1.6.

Theorem 3.3.7. *Let T be a sound, effectively axiomatized extension of* EA*. There is a $0'$ recursive monotone function $\mathfrak{g}$ such that, for every φ, $\mathfrak{g}(\varphi)$ is Π_1, yet for arbitrarily strong true sentences φ*

$$[\varphi \wedge \mathfrak{g}(\varphi)] = [\varphi \wedge \mathsf{Con}_T(\varphi)]$$

and for arbitrarily strong true sentences φ

$$[\varphi \wedge \mathfrak{g}(\varphi)] = [\varphi].$$

Proof. We define the function $\mathfrak{g}$ as follows:

$$\mathfrak{g}(\varphi) = \begin{cases} \bigwedge\{\mathsf{Con}_T(\zeta) : \zeta \in \mathfrak{A} \text{ and } T + \varphi \vdash \zeta\} & \text{if } [\varphi] \neq [\bot] \\ \bot & \text{otherwise} \end{cases}$$

We will check one-by-one that $\mathfrak{g}$ satisfies the properties ascribed to it in the statement of the theorem. We start by checking that $\mathfrak{g}$ is $0'$ recursive. In so doing, we will also demonstrate that $\mathfrak{g}$ is well defined, i.e., always produces a finitary sentence.

Claim. *$\mathfrak{g}$ is $0'$ recursive.*

To verify that $\mathfrak{g}$ is $0'$ recursive, we informally describe an algorithm for calculating $\mathfrak{g}$ using $0'$ as an oracle. Here is the algorithm: Given an input φ, first use $0'$ to determine whether $[\varphi] = [\bot]$. If so, output $\bot$. Otherwise, we have to find all sentences $\psi \in \mathfrak{A}$ such that $T + \varphi \vdash \zeta$. Let's say that φ is the n^{th} sentence in our Gödel numbering of the language of arithmetic. By the construction of $\mathfrak{A}$, the only sentences in $\mathfrak{A}$ that $T + \varphi$ proves must have been numerated into $\mathfrak{A}$ by stage n. So find each of the finitely many sentences that were activated by stage n in the construction of $\mathfrak{A}$. For any such sentence ζ, use $0'$ to determine whether $T + \varphi \vdash \zeta$.[3] Once all sentences $\psi \in \mathfrak{A}$ such that $T + \varphi \vdash \zeta$ have been found, output the conjunction of their consistency statements.

It is now routine to verify that the following claim is true:

Claim. *$\mathfrak{g}$ is monotone and always produces a Π_1 sentence.*

[3]Querying $0'$ is not strictly necessary here. If we already know that two sentences φ and ζ are consistent, we can determine effectively whether $T + \varphi \vdash \zeta$ by paying attention to details of the construction of $\mathfrak{A}$.

We now work towards showing that $\mathfrak{g}$ behaves like the consistency operator for arbitrarily strong inputs.

Claim. *For arbitrarily strong true sentences* φ, $[\varphi \wedge \mathfrak{g}(\varphi)] = [\varphi \wedge \mathsf{Con}_T(\varphi)]$.

To see why the claim is true, note that whenever $\varphi \in \mathfrak{A}$, it follows that $[\mathfrak{g}(\varphi)] = [\mathsf{Con}_T(\varphi)]$, whence

$$[\varphi \wedge \mathfrak{g}(\varphi)] = [\varphi \wedge \mathsf{Con}_T(\varphi)].$$

Since $\mathfrak{A}$ contains arbitrary strong true sentences, it follows immediately that for arbitrarily strong true sentences φ,

$$[\varphi \wedge \mathfrak{g}(\varphi)] = [\varphi \wedge \mathsf{Con}_T(\varphi)].$$

Thus, to prove the theorem, it suffices to see that $\mathfrak{g}$ behaves like the identity operator on arbitrarily strong inputs.

Claim. *For arbitrarily strong true sentences* φ, $[\varphi \wedge \mathfrak{g}(\varphi)] = [\varphi]$.

To this end, we will assume only that ψ is a true sentence satisfying the following claim:

$$T \nvdash (\psi \wedge \mathsf{Con}_T(\psi)) \rightarrow \theta_\psi. \tag{$\sharp$}$$

Recall that if ψ was numerated into $\mathfrak{A}$ at stage n, then θ_ψ is a true sentence that is either the $(n+1)^{th}$ sentence in the Gödel numbering of the language or the negation thereof (depending on which is true). By Lemma 3.3.6, we know that for arbitrarily strong true sentences ψ, ψ satisfies ($\sharp$). We will then show that for any such ψ, where φ is the sentence $(\psi \wedge \mathsf{Con}_T(\psi))$, the following identity holds:

$$[\varphi \wedge \mathfrak{g}(\varphi)] = [\varphi],$$

thus certifying the truth of Claim 3.3.

So let ψ be a true sentence in $\mathfrak{A}$ satisfying ($\sharp$). By Lemma 3.3.2, there is a unique next true sentence numerated into $\mathfrak{A}$, and that sentence is $\big(\psi \wedge \mathsf{Con}_T(\psi) \wedge \theta_\psi\big)$. We assert the following claim:

Claim. *For all sentences* $\zeta \in \mathfrak{A}$, *if* $T + \big(\psi \wedge \mathsf{Con}_T(\psi)\big) \vdash \zeta$ *then also* $T + \psi \vdash \zeta$.

Let's see why Claim 3.3 is true. Since $(\psi \wedge \mathsf{Con}_T(\psi))$ is true, any sentence ζ in $\mathfrak{A}$ that is implied by $\big(\psi \wedge \mathsf{Con}_T(\psi)\big)$ belongs to the true branch (see Corollary 3.3.3). By assumption ($\sharp$), $(\psi \wedge \mathsf{Con}_T(\psi))$ has strength strictly intermediate between ψ and $(\psi \wedge \mathsf{Con}_T(\psi) \wedge \theta_\psi)$. Accordingly, any sentence ζ in $\mathfrak{A}$ that is implied by $(\psi \wedge \mathsf{Con}_T(\psi))$ must have been numerated into $\mathfrak{A}$ before $(\psi \wedge \mathsf{Con}_T(\psi) \wedge \theta_\psi)$. Recall that ψ is the sentence in the true branch numerated into $\mathfrak{A}$ immediately before $(\psi \wedge \mathsf{Con}_T(\psi) \wedge \theta_\psi)$. So ζ either is ψ or was numerated into $\mathfrak{A}$ earlier than ψ. Either way, ψ implies ζ. This certifies the truth of Claim 3.3.

We now introduce the formula $\varphi := \big(\psi \wedge \mathsf{Con}_T(\psi)\big)$. We make the following claim:

Claim. $[\varphi \wedge \mathfrak{g}(\varphi)] = [\varphi \wedge \bigwedge\{\mathsf{Con}_T(\zeta) : \zeta \in \mathfrak{A} \text{ and } T + \psi \vdash \zeta\}]$.

We argue for Claim 3.3 as follows:

$$\begin{aligned}[\varphi \wedge \mathfrak{g}(\varphi)] &= [\varphi \wedge \bigwedge\{\mathsf{Con}_T(\zeta) : \zeta \in \mathfrak{A} \text{ and } T + \varphi \vdash \zeta\}] \text{ by definition of } \mathfrak{g} \\ &= [\varphi \wedge \bigwedge\{\mathsf{Con}_T(\zeta) : \zeta \in \mathfrak{A} \text{ and } T + (\psi \wedge \mathsf{Con}_T(\psi)) \vdash \zeta\}] \text{ by choice of } \varphi \\ &= [\varphi \wedge \bigwedge\{\mathsf{Con}_T(\zeta) : \zeta \in \mathfrak{A} \text{ and } T + \psi \vdash \zeta\}] \text{ by Claim 3.3.}\end{aligned}$$

With Claim 3.3 on board, we are now ready to prove that $[\varphi] = [\varphi \wedge \mathfrak{g}(\varphi)]$. We reason as follows:

$T + \varphi + \mathsf{Con}_T(\psi) \vdash \bigwedge\{\mathsf{Con}_T(\zeta) : \zeta \in \mathfrak{A} \text{ and } T + \psi \vdash \zeta\}$ by the monotonity of Con_T.

$T + \varphi + \mathsf{Con}_T(\psi) \vdash \mathfrak{g}(\varphi)$ by Claim 3.3.

$T + \varphi \vdash \mathfrak{g}(\varphi)$ since $T + \varphi \vdash \mathsf{Con}_T(\psi)$, by the choice of φ.

This trivially implies that:

$$[\varphi] = [\varphi \wedge \mathfrak{g}(\varphi)].$$

This completes the proof of the theorem. ❑

Part II

Second-Order Arithmetic

Chapter 4

Reflection ranks and ordinal analysis

This chapter contains joint work with Fedor Pakhomov.

4.1 Introduction

It is a well-known empirical phenomenon that *natural* axiomatic theories are well-ordered[1] according to many popular metrics of proof-theoretic strength, such as consistency strength. This phenomenon is manifest in *ordinal analysis*, a research program wherein recursive ordinals are assigned to theories to measure their proof-theoretic strength. However, these metrics of proof-theoretic strength do *not* well-order axiomatic theories *in general*. For instance, there are descending chains of sound theories, each of which proves the consistency of the next. However, all such examples of ill-foundedness make use of unnatural, artificial theories. Without a mathematical definition of "natural," it is unclear how to provide a general mathematical explanation of the apparent well-orderedness of the hierarchy of natural theories.

In this paper we introduce a metric of proof-theoretic strength and prove that it is immune to these pathological instances of ill-foundedness. Recall that a theory T is Π^1_1 sound just in case every Π^1_1 theorem of T is true. The Π^1_1 soundness of T is expressible in the language of second-order arithmetic by a formula $\mathsf{RFN}_{\Pi^1_1}(T)$. The formula $\mathsf{RFN}_{\Pi^1_1}(T)$ is also known as the *uniform* Π^1_1 *reflection principle for* T.

Definition 4.1.1. For theories T and U in the language of second-order arithmetic we say that $T \prec_{\Pi^1_1} U$ if U proves the Π^1_1 soundness of T.

This metric of proof-theoretic strength is coarser than consistency strength, but, as we noted, it is also more robust. In practice, when one shows that U proves the consistency of T, one often also establishes the stronger fact that U proves the Π^1_1 soundness of T. Our first main theorem is the following.

[1] Of course, by *well-ordered* here we mean *pre-well-ordered*.

Theorem 4.1.2. *The restriction of $\prec_{\Pi^1_1}$ to the Π^1_1-sound extensions of ACA_0 is well-founded.*

Accordingly, we can attach a well-founded rank—*reflection rank*—to Π^1_1 sound extensions of ACA_0 in the $\prec_{\Pi^1_1}$ ordering.

Definition 4.1.3. The *reflection rank* of T is the rank of T in the ordering $\prec_{\Pi^1_1}$ restricted to Π^1_1 sound extensions of ACA_0. We write $|T|_{\mathsf{ACA}_0}$ to denote the reflection rank of T.

What is the connection between the reflection rank of T and the Π^1_1 proof-theoretic ordinal of T? Recall that the Π^1_1 *proof-theoretic ordinal* $|T|_{\mathsf{WO}}$ of a theory T is the supremum of the order-types of T-provably well-founded primitive recursive linear orders. We will show that the reflection ranks and Π^1_1 proof-theoretic ordinals of theories are closely connected. Recall that ACA^+_0 is axiomatized over ACA_0 by the statement "for every X, the ω^{th} jump of X exists."

Theorem 4.1.4. *For any Π^1_1-sound extension T of ACA^+_0, $|T|_{\mathsf{ACA}_0} = |T|_{\mathsf{WO}}$.*

In general, if $|T|_{\mathsf{ACA}_0} = \boldsymbol{\alpha}$ then $|T|_{\mathsf{WO}} \geqslant \boldsymbol{\varepsilon}_{\boldsymbol{\alpha}}$. We provide examples of theories such that $|T|_{\mathsf{ACA}_0} = \boldsymbol{\alpha}$ and $|T|_{\mathsf{WO}} > \boldsymbol{\varepsilon}_{\boldsymbol{\alpha}}$. Nevertheless for many theories T with $|T|_{\mathsf{ACA}_0} = \boldsymbol{\alpha}$ we have $|T|_{\mathsf{WO}} = \boldsymbol{\varepsilon}_{\boldsymbol{\alpha}}$.

To prove these results, we extend techniques from the proof theory of iterated reflection principles to the second-order context. In particular, we focus on iterated Π^1_1 reflection. Roughly speaking, the theories $\mathbf{R}^{\alpha}_{\Pi^1_1}(T)$ of α-iterated Π^1_1-reflection over T are defined as follows

$$\begin{aligned}
&\mathbf{R}^0_{\Pi^1_1}(T) := T \\
&\mathbf{R}^{\alpha}_{\Pi^1_1}(T) := T + \bigcup_{\beta<\alpha} \mathsf{RFN}_{\Pi^1_1}\big(\mathbf{R}^{\beta}_{\Pi^1_1}(T)\big) \text{ for } \alpha \succ 0.
\end{aligned}$$

The formalization of this definition in arithmetic requires some additional efforts; see §2 for details.

Iterated reflection principles have been used previously to calculate proof-theoretic ordinals. For instance, Schmerl [87] used iterated reflection principles to establish bounds on provable arithmetical transfinite induction principles for fragments of PA. Beklemishev [7] has also calculated proof-theoretic ordinals of subsystems of PA via iterated reflection. These results differ from ours in two important ways. First, these results concern only theories in the language of first-order arithmetic, and hence do not engender calculations of Π^1_1 proof-theoretic ordinals. Second, these results are notation-dependent, i.e., they involve the calculation of proof-theoretic ordinals *modulo* the choice of a particular (natural) ordinal notation system. We are concerned with Π^1_1 reflection. Hence, in light of Theorem 4.1.2, we are able to calculate proof-theoretic ordinals in a manner that is not sensitive to the choice of a particular ordinal notation system.

Theorem 4.1.5. *Let α be an ordinal notation system with the order type $|\alpha| = \boldsymbol{\alpha}$. Then $|\mathbf{R}^{\alpha}_{\Pi^1_1}(\mathsf{ACA}_0)|_{\mathsf{ACA}_0} = \boldsymbol{\alpha}$ and $|\mathbf{R}^{\alpha}_{\Pi^1_1}(\mathsf{ACA}_0)|_{\mathsf{WO}} = \boldsymbol{\varepsilon}_{\boldsymbol{\alpha}}$.*

It is possible to prove Theorem 4.1.4 and Theorem 4.1.5 by formalizing infinitary derivations in ACA_0 and appealing to cut-elimination, and in an early draft of this paper we did just that. Lev Beklemishev suggested that it might be possible to prove these results with methods from the proof theory of iterated reflection principles, namely conservation theorems in the style of Schmerl [87]. Though these methods have become quite polished for studying subsystems of first-order arithmetic, they have not yet been extended to Π^1_1 ordinal analysis. Thus, we devote a section of the paper to developing these techniques in the context of second-order arithmetic. We thank Lev for encouraging us to pursue this approach. Our main result in this respect is the following conservation theorem, where $\Pi^1_1(\Sigma^0_2)$ denotes the complexity class consisting of formulas of the form $\forall X\, \mathsf{F}$ where $\mathsf{F} \in \Pi^0_3$.

Theorem 4.1.6. $\mathbf{R}^{\alpha}_{\Pi^1_1}(\mathsf{ACA}_0)$ *is* $\Pi^1_1(\Sigma^0_2)$ *conservative over* $\mathbf{R}^{\varepsilon_\alpha}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0)$.

To prove this result, we establish connections between Π^1_1 reflection over second-order theories and reflection over arithmetical theories with free set variables.

Finally, we demonstrate that Theorem 4.1.2 could be used for straightforward well-foundedness proofs for certain ordinal notation systems. A recent development in ordinal analysis is the use of ordinal notation systems that are based on reflection principles. Roughly, the elements of such notation systems are reflection principles and they are ordered by proof-theoretic strength. Such notation systems have been extensively studied since Beklemishev [8] endorsed their use as an approach to the canonicity problem for ordinal notations. See [29] for a survey of such notation systems. We prove the well-foundedness of Beklemishev's reflection notation system for ε_0 using the well-foundedness of the $\prec_{\Pi^1_1}$-order. Previously, Beklemishev proved the well-foundedness of this system by constructing the isomorphism with Cantor's ordinal notation system for ε_0. We expect that our techniques—or extensions thereof—could be used to prove the well-foundedness of ordinal notation systems for stronger axiomatic theories.

Here is our plan for the rest of the paper. In §4.2 we fix our notation and introduce some key definitions. In §4.3 we present our technique for showing that certain classes of theories are well-founded (or nearly well-founded) according to various notions of proof-theoretic strength. Our first application of this technique establishes Theorem 4.1.2. In §4.3 we prove various conservation results that connect iterated reflection principles with transfinite induction. The theorems in §4.3 extend results of Schmerl from first-order theories to pseudo Π^1_1 theories, i.e., to theories axiomatized by formulas with at most free set variables, and to second-order theories. We conclude with a proof of Theorem 4.1.6. In §4.4 we establish connections between the reflection ranks and proof-theoretic ordinals of theories, including proofs of Theorem 4.1.4 and Theorem 4.1.5. In §4.5 we show how to use our results to prove the well-foundedness of ordinal notation systems based on reflection principles. In §4.6 we present an explicit example by proving the well-foundedness of Beklemishev's notation system for ε_0.

4.2 Definitions and notation

In this section we describe and justify our choice of meta-theory. We then fix some notation and present some key definitions. Finally, we describe a proof-technique that we will use repeatedly throughout the paper, namely, Schmerl's technique of reflexive induction.

Treatment of theories

Recall that EA is a finitely axiomatizable theory in the language of arithmetic with the exponential function, i.e., in the signature $(0, 1, +, \times, 2^x, \leqslant)$. EA is characterized by the standard recursive axioms for addition, multiplication, and exponentiation as well as the induction schema for Δ_0 formulas. Note that by Δ_0 formulas we mean Δ_0 formulas in the language with exponentiation. EA is strong enough to facilitate typical approaches to arithmetization of syntax. Moreover, EA proves its own Σ_1 completeness.

We will also be interested in EA^+. EA^+ is a theory in the language of EA. EA^+ extends EA by the additional axiom "superexponentiation is total." By superexponentiation, we mean the function 2^x_x where $2^x_0 = x$ and $2^x_{y+1} = 2^{2^x_y}$. EA^+ is the weakest extension of EA in which the cut-elimination theorem is provable. Indeed, the cut-elimination theorem is equivalent to the totality of superexponentiation over EA. See [9] for details on EA and EA^+; see also [41] for details on EA and EA^+ in a slightly different formalism without an explicit symbol for exponentiation. We will use EA^+ as a meta-theory for proving many of our results.

In this paper we will examine theories in three different languages. First the language of first-order arithmetic, i.e., the language of EA. Second the language of first-order arithmetic extended with one additional free set variable X; we also call this the *pseudo*-Π^1_1 *language*. And finally the language of second-order arithmetic. The language of first-order arithmetic of course is a sublanguage of the other two languages. And we consider the pseudo-Π^1_1 language to be a sublanguage of the language of second-order arithmetic by identifying each pseudo-Π^1_1 sentence F with the second-order sentence $\forall X\, \mathsf{F}$.

In the first-order context we are interested in the standard arithmetical complexity classes Π_n and Σ_n. We write Π_∞ to denote the class of all arithmetical formulas. We write $\mathbf{\Pi}^0_n$ to denote the class of formulas that are just like Π_n formulas except that their formulas (potentially) contain a free set variable X. Formulas in the complexity class $\mathbf{\Pi}^0_n$ *cannot* have set quantifiers, and so contain *only* free set variables. Of course, the class $\mathbf{\Sigma}^0_n$ is defined dually to the class $\mathbf{\Pi}^0_n$. We write $\mathbf{\Pi}^0_\infty$ to denote the class of boldface arithmetical formulas, i.e., the class of arithmetical formulas (potentially) with a free set variable.

In the second-order context we are mostly interested in the standard analytical complexity classes Π^1_1 and Σ^1_1. However, we will also use other complexity classes. Suppose $\mathcal{C} \subset \mathcal{L}_2$ is one of the following classes of formulas: Π^0_m or Σ^0_m, for $m \geqslant 1$. Then we denote by $\Pi^1_n(\mathcal{C})$ the class of all the formulas of the form $\forall X_1 \exists X_2 \ldots Q X_n\, \mathsf{F}$, where $\mathsf{F} \in \mathcal{C}$. We define $\Sigma^1_n(\mathcal{C})$ dually.

For a first-order theory T, we use $T(X)$ to denote the pseudo Π^1_1 pendant of T. For example, the theory $\mathsf{PA}(X)$ contains (i) the axioms of PA and (ii) induction axioms for

all formulas in the language, including those with free set variables. The theories $\mathsf{EA}(X)$, $\mathsf{EA}^+(X)$, and $\mathsf{I\Sigma}_1(X)$ are defined analogously, i.e., their induction axioms are extended to include formulas with the free set variable X.

Formulas in any of the three languages we are working with can naturally be identified with words in a suitable finite alphabet, which, in turn, are naturally one-to-one encoded by numbers. Accordingly, we can fix a Gödel numbering of these languages. We denote the Gödel number of an expression τ by $\ulcorner\tau\urcorner$. Many natural syntactic relations (x is a logical axiom, z the result of applying Modus Ponens to x and y, x encodes a Π_n formula, etc.) are elementary definable and their simplest properties can be verified within EA. We also fix a one-to-one elementary coding of finite sequences of natural numbers. $\langle x_1, ..., x_n\rangle$ denotes the code of a sequence $x_1, ..., x_n$ and, for any fixed n, is an elementary function of $x_1, ..., x_n$.

We are concerned with recursively enumerable theories. Officially, a theory T is a Σ_1 formula $\mathsf{Ax}_T(x)$ that is understood as a formula defining the (Gödel numbers of) axioms of T in the standard model of arithmetic, i.e., the set of axioms of T is $\{\varphi : \mathbb{N} \vDash \mathsf{Ax}_T(\varphi)\}$. Thus, we are considering theories *intensionally*, via their axioms, rather than as deductively closed sets of formulas.

Since our base theory EA is fairly weak, we have to be careful with our choice of formalizations of proof predicates. Namely, we want our provability predicate to be Σ_1. And due to this we can't use the straightforwardly defined predicates $\mathsf{PrfNat}_T(x, y)$: x is a Hilbert-style proof of y, where all axioms are either axioms of first-order logic or axioms of T. The predicates $\mathsf{PrfNat}_T(x, y)$ are equivalent to $\forall^b\Sigma_1$-formulas over EA ($\forall^b\Sigma_1$-formulas are the formulas starting with a bounded universal quantifier followed by Σ_1-formula). However, EA is too weak to equivalently transform $\forall^b\Sigma_1$-formulas to Σ_1-formulas; for this one needs the collection scheme $\mathsf{B}\Sigma_1$, which isn't provable in EA. We note that this doesn't affect most natural theories T, in particular, for any T with Δ_0 formula Ax_T, the predicate $\mathsf{PrfNat}_T(x, y)$ is equivalent to a Σ_1 formula over EA.

Nevertheless, to avoid this issue, we work with proof predicates that are forced to be Σ_1 in EA, which are sometimes called *smooth proof* predicates. In the definition of the smooth proof predicate, a "proof" is a pair consisting of an actual Hilbert style proof and a uniform bound for witnesses to the facts that axioms in the proof indeed are axioms. We simply write $\mathsf{Prf}_T(x, y)$ to formalize that x is a "smooth proof" of y in theory T. The predicates $\mathsf{Prf}_T(x, y)$ are Δ_0-formulas. The predicate $\mathsf{Pr}_T(y)$ is shorthand for $\exists x\mathsf{Prf}_T(x, y)$. We use the predicate $\mathsf{Con}(T)$ as shorthand for $\neg\mathsf{Pr}_T(\bot)$, where we fix $\bot$ to be some contradictory sentence.

The closed term $1 + 1 + ... + 1$ (n times) is the numeral of n and is denoted $\underline{n}$. We often omit the bar when no confusion can occur. We also often omit the corner quotes from Gödel numbers when no confusion can occur. For instance, we can encode the notion of a formula φ being provable in a theory T, by saying that there is a T-proof (a sequence subject to certain constraints) the last element of which is the numeral of the Gödel number of φ. However, instead of writing $\mathsf{Pr}_T(\underline{\ulcorner\varphi\urcorner})$ to say that φ is provable we simply write $\mathsf{Pr}_T(\varphi)$.

Suppose T and U are recursively enumerable theories in the same language. We write $T \sqsubseteq U$ if T is a subtheory of U; we can formalize the claim that $T \sqsubseteq U$ in arithmetic with the formula $\forall\varphi\big(\mathsf{Pr}_T(\varphi) \to \mathsf{Pr}_U(\varphi)\big)$. We write $T \equiv U$ if $T \sqsupseteq U$ and $U \sqsupseteq T$. For a class

$\mathcal{C}$ of sentences of the language of T we write $T \sqsubseteq_{\mathcal{C}} U$ if the set of $\mathcal{C}$-theorems of T is a subset of $\mathcal{C}$-theorems of U; this could be naturally formalized in arithmetic with the formula $\forall \varphi \in \mathcal{C}\big(\mathsf{Pr}_T(\varphi) \to \mathsf{Pr}_U(\varphi)\big)$. We write $T \equiv_{\mathcal{C}} U$ if $T \sqsubseteq_{\mathcal{C}} U$ and $U \sqsubseteq_{\mathcal{C}} T$.

We will be interested in partial truth-definitions for various classes of formulas for which we could prove Tarski's bi-conditionals. For a class $\mathcal{C}$ of formulas we call a formula $\mathsf{Tr}_{\mathcal{C}}(x)$ a partial truth definition for $\mathcal{C}$ over a theory T, if $\mathsf{Tr}_{\mathcal{C}}(x)$ is from the class $\mathcal{C}$ and

$$T \vdash \varphi(\vec{x}) \leftrightarrow \mathsf{Tr}_{\mathcal{C}}(\varphi(\vec{x})), \text{ for all } \varphi(\vec{x}) \text{ from } \mathcal{C}.$$

Moreover, we will work only with truth definitions such that the above property is provable in EA.

In the book by Hájek and Pudlák [41, §I.1(d)] there is a construction of partial truth definitions for classes Π_n and Σ_n, $n \geqslant 1$, over $\mathsf{I}\Sigma_1$. However, we will use a sharper construction of partial truth definitions for classes Π_n and Σ_n, $n \geqslant 1$, over EA which could be found in [14, Appendix A]. And we will use truth definitions for classes $\mathbf{\Pi}_n$ and $\mathbf{\Sigma}_n$, $n \geqslant 1$, over $\mathsf{EA}(X)$ that as well were constructed in [14, Appendix A].

In the case of second-order arithmetic there are partial truth definitions for classes $\Pi^1_n(\Pi^0_m)$, $\Sigma^1_n(\Sigma^0_m)$, $\Sigma^1_n(\Pi^0_m)$, and $\Pi^1_n(\Sigma^0_m)$, where $m \geqslant 1$, over RCA_0. One could easily construct this partial truth definitions from the partial truth definitions for classes $\mathbf{\Pi}_n$ and $\mathbf{\Sigma}_n$ over $\mathsf{EA}(X)$. However, over ACA_0 it is possible to construct partial truth definitions for the classes Π^1_n and Σ^1_n, $n \geqslant 1$. Let $\mathbf{\Sigma}^1_1$ be the class of Σ^1_1-formulas with a set parameter X. It is easy to construct partial truth definitions for classes Π^1_n and Σ^1_n, $n \geqslant 1$, from a partial truth definition for $\mathbf{\Sigma}^1_1$. Simpson [92, Lemma V.1.4] proves that for each Σ^1_1 formula $\varphi(X)$ there exists a Δ^0_0 formula $\theta(x, y)$ such that

$$\mathsf{ACA}_0 \vdash \forall X\big(\varphi(X) \leftrightarrow (\exists f\colon \mathbb{N} \to \mathbb{N})\ \forall m\ \theta_\varphi(X \upharpoonright m, f \upharpoonright m)\big).$$

Here $X \upharpoonright m$ is the natural number encoding the finite set $X \cap \{0, \dots, m-1\}$ and $f \upharpoonright m$ is the code of the finite sequence $\langle f(0), \dots, f(m-1)\rangle$. From Simpson's proof it is easy to extract a Kalmar elementary algorithm for constructing the formula θ_φ from a formula φ. And by the same argument as Simpson we show that the Σ^1_1-formula $(\exists f\colon \mathbb{N} \to \mathbb{N})\ \forall m\ \mathsf{Tr}_{\Pi^0_1}(\theta_x(X \upharpoonright m, f \upharpoonright m))$ is a partial truth definition $\mathsf{Tr}_{\mathbf{\Sigma}^1_1}(X, x)$ for the class $\mathbf{\Sigma}^1_1$ over ACA_0.

Ordinal notations

There are many ways of treating ordinal notations in arithmetic. We choose one specific method that will be suitable when we work in the theory EA^+ (and its extensions). Our results will be valid for other natural choices of treatment of ordinal notations, but some of the proofs would have to be tweaked slightly.

Often we will use ordinal notation systems within formal theories that couldn't prove (or even express) the well-foundedness of the notation systems. Also, most of our results are intensional in nature and don't require the notation system to be well-founded from an

external point of view. Due to this, our definition of an ordinal notation system does not require it to be well-founded.

Officially, an ordinal notation α is a tuple $\langle \varphi(x), \psi(x,y), p, n\rangle$ where $\varphi, \psi \in \Delta_0$, $\varphi(n)$ is true according to Tr_{Σ_1}, and p is an EA proof of the fact that on the set $\{x \mid \varphi(x)\}$ the order

$$x \prec_\alpha y \stackrel{\text{def}}{\iff} \psi(x,y)$$

is a strict linear order. More formally p is an EA proof of the conjunction of the following sentences:

1. $\forall x, y, z\big(\varphi(x) \wedge \varphi(y) \wedge \varphi(z) \wedge \psi(x,y) \wedge \psi(y,z) \to \psi(x,z)\big)$ (Transitivity);
2. $\forall x\big(\varphi(x) \to \neg\psi(x,x)\big)$ (Irreflexivity);
3. $\forall x, y\Big(\varphi(x) \wedge \varphi(y) \wedge x \neq y \to \big(\psi(x,y) \wedge \neg\psi(y,x)\big) \vee \big(\psi(y,x) \wedge \neg\psi(x,y)\big)\Big)$ (Antisymmetry).

We now define a partial order $\prec$ on the set of all notation systems. Any tuples $\alpha = \langle \varphi, \psi, p, n\rangle$ and $\alpha' = \langle \varphi', \psi', p', n\rangle$ are $\prec$-incomparable if either $\varphi \neq \varphi'$, or $\psi \neq \psi'$, or $p \neq p'$. If α, β are of the form $\alpha = \langle \varphi, \psi, p, n\rangle$ and $\beta = \langle \varphi, \psi, p, m\rangle$, we put $\alpha \prec \beta$ if $\mathsf{Tr}_{\Sigma_1}\big(\psi(n,m)\big)$ but $\mathsf{Tr}_{\Sigma_1}\big(\neg\psi(m,n)\big)$.

Clearly the relation $\prec$ and the property of being an ordinal notation system are expressible by Σ_1-formulas. In EA^+ we could expand the language by a definable superexponentiation function 2^x_y. Since the superexponentiation function is EA^+ provably monotone, by a standard technique one could show that EA^+ proves induction for the class $\Delta_0(2^x_y)$ of formulas with bounded quantifiers in the expanded language. It is easy to show that over EA^+ the truth of Δ_0-formulas according to the Σ_1-truth predicate could be expressed by a $\Delta_0(2^x_y)$ formula. Thus, the order $\prec$ and the property of being an ordinal notation system are expressible by $\Delta_0(2^x_y)$ formulas, which allows us to reason about them in EA^+ is a straightforward manner.

Let us show that EA^+ proves that $\prec$ is a disjoint union of linear orders. First we note that the theory EA^+ proves the Π_2 soundness of EA (i.e. $\mathsf{RFN}_{\Pi_2}(\mathsf{EA})$, see section below). And we note that for any $\alpha = \langle \varphi(x), \psi(x,y), p, n\rangle$ the conclusion of p (conjunction of sentences (1)–(3)) is EA-provably equivalent to a Π_1 sentence. Hence for any notation system $\alpha = \langle \varphi(x), \psi(x,y), p, n\rangle$ the theory EA^+ proves that the corresponding conjunction of sentences (1)–(3) is true. Using this we easily prove in EA^+ that $\prec$ is a linear ordering, when restricted to the tuples that share the same first three components.

For an ordinal notation α the value of $|\alpha|$ is either an ordinal or ∞. If the lower cone $(\{\beta \mid \beta \prec \alpha\}, \prec)$ is well-founded, then $|\alpha|$ is the ordinal isomorphic to the well-ordering $(\{\beta \mid \beta \prec \alpha\}, \prec)$. Otherwise, $|\alpha| = \infty$. In other words, $|\alpha|$ is the well-founded rank of α in the $\prec$-order.

An alternative (more standard) approach to treating ordinal notations in arithmetic is to fix an elementary ordinal notation up to some ordinal $\boldsymbol{\alpha}$. This is a fixed linear order

$\mathbf{L} = (\mathcal{D}_{\mathbf{L}}, \prec_{\mathbf{L}})$, where both $\mathcal{D}_{\mathbf{L}} \subseteq \mathbb{N}$ and $\prec_{\mathbf{L}} \subseteq \mathbb{N} \times \mathbb{N}$ are given by Δ_0 formulas such that (i) $\mathbf{L}$ is provably linear in EA, (ii) $\mathbf{L}$ is well-founded, and (iii) the order type of $\mathbf{L}$ is $\boldsymbol{\alpha}$. It has been empirically observed that the ordinal notation systems that arise in ordinal analysis results in proof theory are of this kind; see, e.g., [80]. Note that from any $\mathbf{L}$ of this sort we could easily form an ordinal notation (in our sense) α such that there is a Kalmar elementary isomorphism f between $\mathbf{L}$ and $(\{\beta \mid \beta \prec \alpha\}, \prec)$; moreover, the latter is provable in EA^+.

Further we will work with ordinal notation systems that are given by some combinatorially defined system of terms and order on them. The standard example of such a system is the Cantor ordinal notation system up to ε_0. For the notations that we will consider it will be always possible to formalize in EA the definition and proof that the order is linear. Thus, as described above, we will be able to form an ordinal notation α such that there will be a natural isomorphism between $(\{\beta \mid \beta \prec \alpha\}, \prec)$ and the initial combinatorially defined ordinal notation system. We will make transitions from combinatorial definitions of notation systems to ordinal notation systems in our sense without any further comments.

Moreover, we will use expressions like ω^α and ε_α, where α is some ordinal notation system. Let us consider a notation system $\alpha = \langle \varphi(x), \psi(x,y), p, n\rangle$ and define the notation system $\omega^\alpha = \langle \varphi'(x), \psi'(x,y), p', n'\rangle$. We want the order $\prec_{\omega^\alpha}$ to be the order on the terms $\omega^{a_1} + \ldots + \omega^{a_k}$, where $a_1 \succeq_\alpha \ldots \succeq_\alpha a_k$. And the order $\prec_{\omega^\alpha}$ is defined as the usual order on Cantor normal forms, where we compare a_i by the order $\prec_\alpha$. By arithmetizing this definition of $\prec_{\omega^\alpha}$ we get φ', ψ', and p'. We put n' to be the number encoding the term ω^n. Note that, according to this definition, α and ω^α are $\prec$-incomparable. However, if $\alpha \prec \beta$, then $\omega^\alpha \prec \omega^\beta$.

The definition of the notation system ε_α is similar to that of ω^α. The system of terms for ε_α consists of nested Cantor normal forms built up from 0 and elements ε_a, for $a \in \mathsf{dom}(\prec_\alpha)$. The comparission of nested Cantor normal forms is defined in the standard fashion, where we compare elements ε_a and ε_b as $a \prec_\alpha b$.

Reflection principles

Suppose $\mathcal{C}$ is some class of formulas in one of the languages that we consider and T is a theory in the same language. The uniform $\mathcal{C}$ reflection principle $\mathsf{RFN}_{\mathcal{C}}(T)$ over T is the schema

$$\forall \vec{x}\big(\mathsf{Pr}_T(\varphi(\vec{x})) \to \varphi(\vec{x})\big)$$

for all $\varphi \in \mathcal{C}$, where $\vec{x}$ are free number variables and $\varphi(\vec{x})$ contains no other variables.

In those cases for which we have a truth-definition for $\mathcal{C}$ in T the scheme $\mathsf{RFN}_{\mathcal{C}}(T)$ can be axiomatized by the single sentence

$$\forall \varphi \in \mathcal{C}\big(\mathsf{Pr}_T(\varphi) \to \mathsf{Tr}_{\mathcal{C}}(\varphi)\big).$$

Given an ordinal notation system $\prec$, we informally define the operation $\mathbf{R}^{\cdot}_{\mathcal{C}}(\cdot)$ of iterated

$\mathcal{C}$ reflection along $\prec$ as follows.

$$\mathbf{R}^0_{\mathcal{C}}(T) := T$$
$$\mathbf{R}^\alpha_{\mathcal{C}}(T) := T + \bigcup_{\beta\prec\alpha} \mathsf{RFN}_{\mathcal{C}}\big(\mathbf{R}^\beta_{\mathcal{C}}(T)\big) \text{ for } \alpha \succ 0.$$

More formally, we appeal to Gödel's fixed point lemma in EA. We fix a formula $\mathsf{RFN\text{-}Inst}_{\mathcal{C}}(U, x)$, where U and x are first-order variables, that formalizes the fact that x is an instance of the scheme $\mathsf{RFN}_{\mathcal{C}}(U)$. We now want to define a Σ_1 formula $\mathsf{Ax}_{\mathbf{R}^\alpha_{\mathcal{C}}(T)}(x)$ (note that α, T, and x are arguments of the formula) that defines the set of axioms of the theories $\mathbf{R}^\alpha_{\mathcal{C}}(T)$. We define the formula as a fixed point:

$$\mathsf{EA} \vdash \mathsf{Ax}_{\mathbf{R}^\alpha_{\mathcal{C}}(T)}(x) \leftrightarrow \big(\mathsf{Ax}_T(x) \vee \exists\beta \prec \alpha\ \mathsf{RFN\text{-}Inst}_{\mathcal{C}}(\mathbf{R}^\beta_{\mathcal{C}}(T), x))\big),$$

note that when we substitute $\mathbf{R}^\beta_{\mathcal{C}}(T)$ in $\mathsf{RFN\text{-}Inst}_{\mathcal{C}}$ we actually substitute (the Gödel number of) $\mathsf{Ax}_{\mathbf{R}^\beta_{\mathcal{C}}(T)}$.

Beklemishev introduced this approach to defining progressions of iterated reflection in [6]; the reader can find a more modern version of this approach in [13]. It is easy to prove that this definition of progressions of iterated reflection provides a unique (up to EA provable deductive equivalence) definition of the theories $\mathbf{R}^\alpha_{\mathcal{C}}(T)$.

Reflexive induction

We often employ Schmerl's technique of *reflexive induction*. Reflexive induction is a way of simulating large amounts of transfinite induction in weak theories. The technique is facilitated by the following theorem; we include the proof of the theorem, which is very short.

Theorem 4.2.1 (Schmerl)**.** *Let T be a recursively axiomatized theory (in one of the languages that we consider) that contains EA. Suppose*

$$T \vdash \forall\alpha\Big(\mathsf{Pr}_T\big(\forall\beta \prec \alpha\ \varphi(\beta)\big) \to \varphi(\alpha)\Big).$$

Then $T \vdash \forall\alpha\ \varphi(\alpha)$.[2]

Proof. Suppose that $T \vdash \forall\alpha\Big(\mathsf{Pr}_T\big(\forall\beta \prec \alpha\ \varphi(\beta)\big) \to \varphi(\alpha)\Big)$. We infer that

$$T \vdash \forall\alpha\mathsf{Pr}_T\big(\forall\beta \prec \alpha\ \varphi(\beta)\big) \to \forall\alpha\ \varphi(\alpha),$$

whence it follows that

$$T \vdash \mathsf{Pr}_T\big(\forall\alpha\ \varphi(\alpha)\big) \to \forall\alpha\ \varphi(\alpha).$$

Löb's theorem then yields $T \vdash \forall\alpha\ \varphi(\alpha)$. ❑

[2]Schmerl proved this result over the base theory PRA. Beklemishev [7] weakened the base theory to EA.

Accordingly, to prove claims of the form $T \vdash \forall\alpha\ \varphi(\alpha)$, we often prove that $T \vdash \forall\alpha\Big(\mathsf{Pr}_T\big(\forall\beta \prec \alpha\ \varphi(\beta)\big) \to \varphi(\alpha)\Big)$ and infer the desired claim by Schmerl's Theorem. While working inside T, we refer to the assumption $\mathsf{Pr}_T\big(\forall\beta \prec \alpha\ \varphi(\beta)\big)$ as the *reflexive induction hypothesis*.

4.3 Well-foundedness and reflection principles

In this section we develop a technique for showing that certain orders on axiomatic theories exhibit a well-foundeness like properties. The coarsest order that we will consider is Π^1_1 reflection order for which we will prove that its restriction to Π^1_1 sound theories is well-founded. For weaker reflection and consistency orders we will prove only some well-foundedness like properties. Also we note that the same technique is used in [24, Theorem 3.2] to prove certain facts about axiomatic theories of truth and in [60, Theorem 1.1] to prove a recursion-theoretic result concerning the hyper-degrees.

Our technique is inspired by H. Friedman's [31] proof of the following result originally due to Steel [98]; recall that $\leqslant_T$ denotes Turing reducibility.

Theorem 4.3.1. *Let $P \subset \mathbb{R}^2$ be arithmetic. Then there is no sequence $(x_n)_{n<\omega}$ of reals such that for every n, both $x_n \geqslant_T x'_{n+1}$ and also x_{n+1} is the unique real y such that $P(x_n, y)$.*

Friedman and Steel were not directly investigating the well-foundedness of axiomatic systems, but rather an analogous phenomenon from recursion theory, namely, the well-foundedness of natural Turing degrees under Turing reducibility. The adaptability of Friedman's proof arguably strengthens the analogy between these phenomena.

In this section we study both first and second order theories. The first theory that we treat with our technique is ACA_0, a subsystem of second-order arithmetic that has been widely studied in reverse mathematics. ACA_0 is arithmetically conservative over PA. We then turn to other applications of our technique. We consider RCA_0, another subsystem of second-order arithmetic and familiar base theory from reverse mathematics. RCA_0 is conservative over $\mathsf{I\Sigma}_1$. We then turn to first-order theories, and we study elementary arithmetic EA as our object theory.

Π^1_1-Reflection

In this subsection we examine the ordering $\prec_{\Pi^1_1}$ on r.e. extensions of ACA_0, where

$$T \prec_{\Pi^1_1} U \iff^{\text{def}} U \vdash \mathsf{RFN}_{\Pi^1_1}(T).$$

We will show that there are no infinite $\prec_{\Pi^1_1}$ descending sequences of Π^1_1 sound extensions of ACA_0. We recall that, provably in ACA_0, a theory T is Π^1_1 sound if and only if T is consistent with any true Σ^1_1 statement.

Theorem 4.3.2. (ACA_0) *The restriction of the order $\prec_{\Pi^1_1}$ to Π^1_1-sound r.e. extensions of ACA_0 is well-founded.*

Proof. In order to prove the result in ACA_0 we show the inconsistency of the theory ACA_0 plus the following statement DS, which says that there *is* a descending sequence of Π^1_1 sound extensions of ACA_0 in the $\prec_{\Pi^1_1}$ ordering:

$$\mathsf{DS} := \exists E\colon \langle T_i \mid i \in \mathbb{N}\rangle(\mathsf{RFN}_{\Pi^1_1}(T_0) \wedge \forall x\, \mathsf{Pr}_{T_x}(\mathsf{RFN}_{\Pi^1_1}(T_{x+1})) \wedge \forall x(T_x \sqsupseteq \mathsf{ACA}_0))$$

Note that $E\colon \langle T_i \mid i \in \mathbb{N}\rangle$ is understood to mean that E is a set encoding a sequence $\langle T_0, T_1, T_2, \ldots\rangle$ of r.e. theories.

If we prove that $\mathsf{ACA}_0{+}\mathsf{DS}$ proves its own consistency, then the inconsistency of $\mathsf{ACA}_0{+}\mathsf{DS}$ follows from Gödel's second incompleteness theorem. We reason in $\mathsf{ACA}_0 + \mathsf{DS}$ to to prove consistency of $\mathsf{ACA}_0 + \mathsf{DS}$.

Let $E\colon \langle T_i \mid i \in \mathbb{N}\rangle$ be a sequence of theories witnessing the truth of DS. Let us consider the sentence F

$$\exists U\colon \langle S_i \mid i \in \mathbb{N}\rangle(S_0 = T_1 \wedge \forall x\, \mathsf{Pr}_{S_x}(\mathsf{RFN}_{\Pi^1_1}(S_{x+1})) \wedge \forall x(S_x \sqsupseteq \mathsf{ACA}_0)).$$

The sentence F is true since we could take $\langle T_{i+1} : i \in \mathbb{N}\rangle$ as U. It is easy to observe that F is Σ^1_1.

From $\mathsf{RFN}_{\Pi^1_1}(T_0)$ we get that T_0 is consistent with any true Σ^1_1 statement. Thus, we infer that

$$\mathsf{Con}(T_0 + \mathsf{F}).$$

Now using the fact that $\mathsf{Pr}_{T_0}\big(\mathsf{RFN}_{\Pi^1_1}(T_1)\big)$ and that $T_0 \sqsupseteq \mathsf{ACA}_0$ we conclude,

$$\mathsf{Con}(\mathsf{ACA}_0 + \mathsf{RFN}_{\Pi^1_1}(T_1) + \mathsf{F}).$$

But it is easy to see that $\mathsf{RFN}_{\Pi^1_1}(T_1) + \mathsf{F}$ implies DS in ACA_0. In particular, we may take $\langle T_1, T_2, ...\rangle$ as our new witness to DS. Thus, we conclude that $\mathsf{Con}(\mathsf{ACA}_0 + \mathsf{DS})$. ❑

We now observe that a similar result holds over RCA_0. To do so, we consider formulas from the complexity class $\Pi^1_1(\Sigma^0_2)$ (see §2.4). It is easy to see that the proof of Theorem 4.3.3 remains valid if we replace the theory ACA_0 with RCA_0, the complexity class Π^1_1 with $\Pi^1_1(\Sigma^0_2)$, and the complexity class Σ^1_1 with $\Sigma^1_1(\Pi^0_2)$. Thus, we also infer the following.

Theorem 4.3.3. (RCA_0) *The restriction of the order $\prec_{\Pi^1_1(\Sigma^0_2)}$ to $\Pi^1_1(\Sigma^0_2)$-sound r.e. extensions of RCA_0 theories is well-founded.*

Π_3 soundness

In this subsection we study the complexity of descending sequences of r.e. theories with respect to Π_3 soundness. We recall that (provably in EA) a theory T is Π_3 sound just in case T is 2-consistent, i.e., just in case T is consistent with any true Π_2 sentence.

Theorem 4.3.4. *There is no recursively enumerable sequence $(T_n)_{n<\omega}$ of r.e. extensions of* EA *such that T_0 is Π_3 sound and such that for every n, $T_n \vdash \mathsf{RFN}_{\Pi_3}(T_{n+1})$.*

Proof. If the theorem fails, then the following sentence is true,

$$\mathsf{DS} := \exists e\colon \langle T_i \mid i \in \mathbb{N}\rangle \Big(\mathsf{RFN}_{\Pi_3}(T_0) \wedge \forall x\ \mathsf{Pr}_{T_x}\big(\mathsf{RFN}_{\Pi_3}(T_{x+1})\big)\Big)$$

where $\exists e : \langle T_i : i \in \mathbb{N}\rangle$ is understood to mean that e is an index for a Turing machine enumerating the sequence $\langle T_0, T_1, ...\rangle$.

We show that $\mathsf{EA} + \mathsf{DS}$ proves its own consistency, whence, by Gödel's second incompleteness theorem, $\mathsf{EA} + \mathsf{DS}$ is inconsistent and hence DS is false.

Work in $\mathsf{EA} + \mathsf{DS}$. Since DS is true, it has some witness $e\colon \langle T_i \mid i \in \mathbb{N}\rangle$. We now consider the sequence e' that results from omitting T_0 from e. More formally, we consider the sequence $e'\colon \langle T'_i \mid i \in \mathbb{N}\rangle$, which is numerated by the Turing functional $\{e'\} : x \mapsto \{e\}(x+1)$. That is, for each i, $T'_i = T_{i+1}$.

From DS we infer that for all x, $T_{x+1} \vdash \mathsf{RFN}_{\Pi_3}(T_{x+2})$. Thus, for every x, $T'_x \vdash \mathsf{RFN}_{\Pi_3}(T'_{x+1})$ by the definition of e'.

From the first conjunct of DS we infer that $\mathsf{RFN}_{\Pi_3}(T_0)$. That is, T_0 is consistent with any Π_2 truth. Thus, we infer that

$$T_0 + \forall x\ \mathsf{Pr}_{T'_x}\big(\mathsf{RFN}_{\Pi_3}(T'_{x+1})\big)$$

is consistent.

On the other hand, from DS we infer that T_0 *proves* the Π_3 soundness of T'_0. So it is consistent that e' witnesses DS. ❑

Consistency

In this subsection we provide a new proof of a theorem independently due to H. Friedman, Smorynski, and Solovay (see [59, 96]). Before stating the theorem we recall that, EA proves the equivalence of, the consistency sentences $\mathsf{Con}(T)$ and the Π_1-reflection principle $\mathsf{RFN}_{\Pi_1}(T)$.

Theorem 4.3.5. *There is no recursively enumerable sequence $(T_n)_{n<\omega}$ of r.e. extensions of* EA *such that T_0 is consistent and such that* $\mathsf{EA} \vdash \forall x\ \mathsf{Pr}_{T_x}\big(\mathsf{Con}(T_{x+1})\big)$.

Proof. Suppose, toward a contradiction, that there is a recursively enumerable sequence $(T_n)_{n<\omega}$ of r.e. extensions of EA such that T_0 is consistent and such that

$$\mathsf{EA} \vdash \forall x\ \mathsf{Pr}_{T_x}\big(\mathsf{Con}(T_{x+1})\big).$$

Since EA is sound, we also infer that for every n, $T_n \vdash \mathsf{Con}(T_{n+1})$. Thus the following sentence is true.

$$\mathsf{DS} := \exists e\colon \langle T_i \mid i \in \mathbb{N}\rangle \Big(\mathsf{Con}(T_0) \wedge \mathsf{Pr}_{\mathsf{EA}}\big(\forall x\ \mathsf{Pr}_{T_x}(\mathsf{Con}(T_{x+1}))\big) \wedge \forall x\ \mathsf{Pr}_{T_x}\big(\mathsf{Con}(T_{x+1})\big)\Big)$$

where $\exists e : \langle T_i : i \in \mathbb{N}\rangle$ is understood to mean that e is an index for a Turing machine enumerating the sequence $\langle T_0, T_1, ...\rangle$.

We show that $\mathsf{EA} + \mathsf{DS}$ proves its own consistency, whence, by Gödel's second incompleteness theorem, $\mathsf{EA} + \mathsf{DS}$ is inconsistent and hence DS is false.

Work in $\mathsf{EA} + \mathsf{DS}$. Since DS is true, it has some witness $e\colon \langle T_i \mid i \in \mathbb{N}\rangle$. We consider the sequence $e'\colon \langle T'_i \mid i \in \mathbb{N}\rangle$ that results from dropping T_0 from the sequence produced by e. More formally, we consider the sequence e' which is numerated by the Turing functional $\{e'\} : x \mapsto \{e\}(x+1)$.

Claim. *e' is provably a witness to DS in T_0.*

To see that e' provably witnesses the third conjunct of DS in T_0, we reason as follows.

$$\begin{aligned}
&\mathsf{EA} \vdash \forall x\ \mathsf{Pr}_{T_{x+1}}\mathsf{Con}(T_{x+2}) \text{ by } \mathsf{DS}.\\
&\mathsf{EA} \vdash \forall x\ \mathsf{Pr}_{T'_x}\mathsf{Con}(T'_{x+1}) \text{ since } T'_x = T_{x+1} \text{ by definition of } e'.\\
&T_0 \vdash \forall x\ \mathsf{Pr}_{T'_x}\mathsf{Con}(T'_{x+1}) \text{ since } T_0 \text{ extends } \mathsf{EA}.
\end{aligned}$$

To see that e' provably witnesses the second conjunct of DS in T_0, we reason as follows.

$$\begin{aligned}
&\mathsf{EA} \vdash \forall x\ \mathsf{Pr}_{T'_x}\mathsf{Con}(T'_{x+1}) \text{ as above.}\\
&\mathsf{EA} \vdash \mathsf{Pr}_{\mathsf{EA}}\big(\forall x\ \mathsf{Pr}_{T'_x}\mathsf{Con}(T'_{x+1})\big) \text{ by the } \Sigma_1 \text{ completeness of } \mathsf{EA}.\\
&T_0 \vdash \mathsf{Pr}_{\mathsf{EA}}\big(\forall x\ \mathsf{Pr}_{T'_x}\mathsf{Con}(T'_{x+1})\big) \text{ since } T_0 \text{ extends } \mathsf{EA}.
\end{aligned}$$

We now show that e' provably witnesses the first conjunct of DS in T_0. From the first conjunct of DS we infer that $\mathsf{Con}(T_0)$. It follows that T_0 is Π_1 sound. We reason as follows.

$$\begin{aligned}
&T_0 \vdash \mathsf{Con}(T_1) \text{ by } \mathsf{DS}.\\
&T_0 \vdash \mathsf{Con}(T'_0) \text{ since provably } T'_0 = T_1.
\end{aligned}$$

We then infer that $\mathsf{Con}(T'_0)$ by the Π_1 soundness of T_0. So e' is provably a witness to DS in a consistent theory. Therefore $\mathsf{EA} + \mathsf{DS}$ is consistent. ❑

Remark 4.3.6. Note that we just proved the non-existence of EA-*provably* descending r.e. sequences. Without the condition of EA provability such descending sequences *do* exist. H. Friedman, Smorynski, and Solovay independently proved that there is a recursive sequence $\langle T_0, T_1, ...\rangle$ of consistent extensions of EA such that for all n, $T_n \vdash \mathsf{Con}(T_{n+1})$, answering a question of Gaifman; see [96] for details.

Π_2 soundness

We now know that there are *no* recursive descending sequences of Π_3 sound theories with respect to the Π_3 reflection order, but there *are* recursive descending sequences of consistent theories with respect to consistency strength. In this subsection we treat the remaining case, namely, Π_2 soundness. We prove that there *is* an infinite sequences $\langle T_0, T_1, ...\rangle$ of Π_2 sound

extensions of EA such that for all n, $T_n \vdash \mathsf{RFN}_{\Pi_2}(T_{n+1})$. In this sense, Theorem 4.3.4 is best possible.

In the section, for technical reasons it will be useful for us to impose some natural conditions on our proof predicate. We make sure that any proof in our proof system has only one conclusion, whence

$$\mathsf{EA} \vdash \forall x, y_1, y_2 \Big(\big(\mathsf{Prf}_T(x, y_1) \wedge \mathsf{Prf}_T(x, y_2) \big) \to y_1 = y_2 \Big).$$

Moreover, we arrange the proof system so that indices for statements are less than or equal to the indices for their proofs, i.e.,

$$\mathsf{EA} \vdash \forall x, y (\mathsf{Prf}_T(x, y) \to y \leqslant x). \tag{4.1}$$

Note that the conclusions of the theorems in our paper are not sensitive to the choice of proof predicate as long as the resulting provability predicates are EA-provably equivalent. And it is easy to see that even if our initial choice of $\mathsf{Prf}_T(x, \varphi)$ didn't satisfy the mentioned conditions, it is easy to modify it to satisfy the conditions, while preserving the provability predicate $\mathsf{Pr}_T(\varphi)$ up to EA-provable equivalence.

Before proving the theorem we make a few more remarks preliminary remarks. We use the symbol $\dot{-}$ to denote the truncated subtraction function, i.e., $n \dot{-} m = n - m$ if $n > m$ and 0 otherwise. We remind the reader that, provably in EA, a theory is Σ_1 sound if and only if it is Π_2 sound. We also pause to make the following remark, which will invoke in the proof of the theorem.

Remark 4.3.7. For any Π_2 sound extension T of EA, the theory $T + \neg\mathsf{RFN}_{\Pi_2}(T)$ is Π_2 sound. This is actually an instance Gödel's second incompleteness theorem that is applied to 1-provability rather than the ordinary provability. Recall that 1-provability predicate $1\text{-}\mathsf{Pr}_T(\varphi)$ for a theory T is

$$\exists \psi \in \Sigma_2 \big(\mathsf{Tr}_{\Sigma_2}(\psi) \wedge \mathsf{Pr}_T(\psi \to \varphi) \big). \tag{4.2}$$

The consistency notion that corresponds to 1-provability is precisely Π_2-soundness:

$$\mathsf{EA} \vdash \forall \varphi \, \big(\neg 1\text{-}\mathsf{Pr}_T(\neg\varphi) \leftrightarrow \mathsf{RFN}_{\Pi_2}(T + \varphi) \big). \tag{4.3}$$

It is easy to see that 1-provability predicate for a theory T satisfies the usual Hilbrt-Bernays-Löb derivability conditions. Thus Gödel's second incompleteness theorem for it states that if a theory $T \sqsupseteq \mathsf{EA}$ is Π_2-sound, then $\mathsf{RFN}_{\Pi_2}(T)$ is not 1-provable in T. And the latter is equivalent to Π_2-soundness of $T + \neg\mathsf{RFN}_{\Pi_2}(T)$.

We are now ready for the proof of the theorem.

Theorem 4.3.8. *There is a recursive sequence $(\varphi_n)_{n<\omega}$ of Π_2-sound sentences such that, for each n, $\mathsf{EA} + \varphi_n \vdash \mathsf{RFN}_{\Pi_2}(\mathsf{EA} + \varphi_{n+1})$.*

Proof. For each $n \in \mathbb{N}$, we define the sentence φ_n as follows:

$$\varphi_n := \exists \psi \in \Sigma_1 \exists p \Big(\mathsf{Prf}_{\mathsf{I}\Sigma_2}(p, \psi) \wedge \neg \mathsf{True}_{\Sigma_1}(\psi) \wedge \mathsf{RFN}_{\Pi_2}\big(\mathbf{R}^{p \dot{-} n}_{\Pi_2}(\mathsf{EA})\big) \Big)$$

That is, φ_n expresses "$\mathsf{I}\Sigma_2$ proves a false Σ_1 sentence via a proof p, and Π_2 reflection for EA can be iterated up to $p \dot{-} n$."

The motivation for picking that individual formula is as follows: To find a descending sequence, we will iterate Π_2 reflection up to some non-standard number. So we need to make sure that our formula forces a certain number to be non-standard but without implying any false Π_2 sentences. The way we do that is by saying that $\mathsf{I}\Sigma_2$ proves a false Σ_1 sentence. This has (we will show) no false Π_2 consequences. However, (the code of) any proof witnessing a failure of Σ_1 soundness in $\mathsf{I}\Sigma_2$ must be non-standard. We find our descending sequence by iterating Π_2 reflection up to this non-standard number.

Now the formal details start. We need to check that φ_n is Π_2 sound for each n, and that $\mathsf{EA} + \varphi_n \vdash \mathsf{RFN}_{\Pi_2}(\mathsf{EA} + \varphi_{n+1})$.

Claim. $\mathsf{EA} + \varphi_n$ *is* Π_2 *sound for each* n.

The first thing to note is that

$$\mathsf{I}\Sigma_2 \vdash \forall x \; \mathsf{RFN}_{\Pi_2}\big(\mathbf{R}^{x}_{\Pi_2}(\mathsf{EA})\big). \tag{4.4}$$

Indeed, $\mathsf{I}\Sigma_2$ can prove the latter claim by induction on x. Recall that $\mathsf{I}\Sigma_2 \equiv I\Pi_2$ and $\forall x \; \mathsf{RFN}_{\Pi_2}\big(\mathbf{R}^{x}_{\Pi_2}(\mathsf{EA})\big)$ is a Π_2-formula, hence $\mathsf{I}\Sigma_2$ could formalize the necessary induction. Also it is known that $\mathsf{I}\Sigma_2 \supseteq \mathsf{I}\Sigma_1 \equiv \mathsf{EA} + \mathsf{RFN}_{\Pi_3}(\mathsf{EA})$ and that

$$\mathsf{EA} + \mathsf{RFN}_{\Pi_3}(\mathsf{EA}) \vdash \psi \to \mathsf{RFN}_{\Pi_2}(\mathsf{EA} + \psi),$$

for any Π_2-formula ψ. This allows us to verify the base and step of the induction in $\mathsf{I}\Sigma_2$.

The second thing to note is that, since Π_2 reflection is provably equivalent (in EA) to Σ_1 reflection, it follows that:

$$\mathsf{I}\Sigma_2 + \neg \mathsf{RFN}_{\Pi_2}(\mathsf{I}\Sigma_2) \vdash \exists \psi \in \Sigma_1 \exists p \big(\mathsf{Prf}_{\mathsf{I}\Sigma_2}(p, \psi) \wedge \neg \mathsf{True}_{\Sigma_1}(\psi) \big) \tag{4.5}$$

Putting these two observations together, we infer that, for each standard $n \in \mathbb{N}$,

$$\mathsf{I}\Sigma_2 + \neg \mathsf{RFN}_{\Pi_2}(\mathsf{I}\Sigma_2) \vdash \exists \psi \in \Sigma_1 \exists p \big(\mathsf{Prf}_{\mathsf{I}\Sigma_2}(p, \psi) \wedge \neg \mathsf{True}_{\Sigma_1}(\psi) \wedge \mathsf{RFN}^{p \dot{-} n}_{\Pi_2}(\mathsf{EA}) \big) \tag{4.6}$$

which is just to say that for each standard $n \in \mathbb{N}$, $\mathsf{I}\Sigma_2 + \neg \mathsf{RFN}_{\Pi_2}(\mathsf{I}\Sigma_2) \vdash \varphi_n$. Thus, to see that $\mathsf{EA} + \varphi_n$ is Π_2 sound, it suffices to observe that $\mathsf{I}\Sigma_2 + \neg \mathsf{RFN}_{\Pi_2}(\mathsf{I}\Sigma_2)$ is Π_2 sound. The latter claim follows immediately from Remark 4.3.7.

Before checking that $\mathsf{EA} + \varphi_n \vdash \mathsf{RFN}_{\Pi_2}(\mathsf{EA} + \varphi_{n+1})$, we will establish the following lemma:

Lemma 4.3.9. *For all standard* $n \in \mathbb{N}$,

$$\mathsf{EA} \vdash \forall p \forall \psi \in \Sigma_1 \Big(\big(\mathsf{Prf}_{\mathsf{I}\Sigma_2}(p, \psi) \wedge \neg \mathsf{True}_{\Sigma_1}(\psi) \big) \to p > n \Big).$$

Proof. The first thing to note is that (by the Σ_1 soundness of $\mathsf{I}\Sigma_2$ and the Σ_1 completeness of EA) for any $\psi \in \Sigma_1$, if $\mathsf{I}\Sigma_2 \vdash \psi$ then also $\mathsf{EA} \vdash \psi$. Now, for any standard $p \in \mathbb{N}$, EA can check whether p constitutes an $\mathsf{I}\Sigma_2$ proof of a Σ_1 sentence ψ, and if p does constitute such a proof, then EA will prove ψ as well. That is, for each standard $p \in \mathbb{N}$:

$$\mathsf{EA} \vdash \forall \psi \in \Sigma_1 \Big(\mathsf{Prf}_{\mathsf{I}\Sigma_2}(p, \psi) \to \mathsf{True}_{\Sigma_1}(\psi) \Big)$$

It follows that for each standard $n \in \mathbb{N}$:

$$\mathsf{EA} \vdash \forall p \leqslant n \forall \psi \in \Sigma_1 \Big(\mathsf{Prf}_{\mathsf{I}\Sigma_2}(p, \psi) \to \mathsf{True}_{\Sigma_1}(\psi) \Big)$$

Whence for each standard $n \in \mathbb{N}$:

$$\mathsf{EA} \vdash \forall p \forall \psi \in \Sigma_1 \Big(\big(\mathsf{Prf}_{\mathsf{I}\Sigma_2}(p, \psi) \wedge \neg \mathsf{True}_{\Sigma_1}(\psi) \big) \to p > n \Big)$$

This completes the proof of the lemma. ❑

With the lemma on board, we are now ready to verify the following claim:

Claim. *For each $n \in \mathbb{N}$,*

$$\mathsf{EA} + \varphi_n \vdash \mathsf{RFN}_{\Pi_2}(\mathsf{EA} + \varphi_{n+1}).$$

Let's fix an $n \in \mathbb{N}$ and **reason in** $\mathsf{EA} + \varphi_n$**:**

According to φ_n, there is an $\mathsf{I}\Sigma_2$ proof p of a false Σ_1 sentence ψ and $\mathsf{RFN}_{\Pi_2}(\mathbf{R}_{\Pi_2}^{p \dot{-} n}(\mathsf{EA}))$ is Π_2-sound. From Lemma 4.3.9 we infer that $p > n$. It follows that $p \dot{-} n > 0$, whence $p \dot{-} n = (p \dot{-} (n+1)) + 1$. Hence

$$\mathbf{R}_{\Pi_2}^{p \dot{-} n}(\mathsf{EA}) \equiv \mathbf{R}_{\Pi_2}^{(p \dot{-} (n+1)) + 1}(\mathsf{EA}) \equiv \mathsf{EA} + \mathsf{RFN}_{\Pi_2}\big(\mathbf{R}_{\Pi_2}^{p \dot{-} (n+1)}(\mathsf{EA})\big). \tag{4.7}$$

Thus

$$\mathsf{RFN}_{\Pi_2}\Big(\mathsf{EA} + \mathsf{RFN}_{\Pi_2}\big(\mathbf{R}_{\Pi_2}^{p \dot{-} (n+1)}(\mathsf{EA})\big)\Big)$$

Since $\mathsf{Prf}_{\mathsf{I}\Sigma_2}(p, \psi)$ is a true Σ_1 sentence and ψ is a false Σ_1 sentence we infer that

$$\mathsf{RFN}_{\Pi_2}\Big(\mathsf{EA} + \mathsf{Prf}_{\mathsf{I}\Sigma_2}(p, \psi) + \neg \mathsf{True}_{\Sigma_1}(\psi) + \mathsf{RFN}_{\Pi_2}\big(\mathbf{R}_{\Pi_2}^{p \dot{-} (n+1)}(\mathsf{EA})\big)\Big)$$

Which straightforwardly implies $\mathsf{RFN}_{\Pi_2}(\mathsf{EA} + \varphi_{n+1})$. This completes the proof of the theorem. ❑

Question 4.3.10. *In Theorem 4.3.4 and Theorem 4.3.8 we studied how strong reflection principles should be to guarantee that there are no* recursive *descending sequences in the corresponding reflection order. It is natural to ask how this result could be generalized to higher Turing degrees.*

Let n be a natural number. For which m is there a sequence $\langle T_i \mid i \in \mathbb{N} \rangle$ recursive in $0^{(n)}$ such that all T_i are Π_m sound extensions of EA and $T_i \vdash \mathsf{RFN}_{\Pi_m}(T_{i+1})$, for all i? The same question for Σ_m?

4.4 Iterated reflection and conservation

In this section we prove a number of conservation theorems relating iterated reflection and transfinite induction. These results are inspired by the following theorem, which is often known as *Schmerl's formula* [87]. For an ordinal notation system α, ω_n^α is the result of n-applications of ω-exponentiation (see §4.2), starting with α, i.e., $\omega_0^\alpha = \alpha$ and $\omega_{n+1}^\alpha = \omega^{\omega_n^\alpha}$.

Theorem 4.4.1 (Schmerl)**.** *Let n, m be natural numbers. Provably in EA^+, for any notation system α,*

$$\mathbf{R}^\alpha_{\Pi^0_{n+m}}(\mathsf{EA}^+) \equiv_{\Pi^0_n} \mathbf{R}^{\omega_m(\alpha)}_{\Pi^0_n}(\mathsf{EA}^+).$$

Schmerl's formula is a useful tool for calculating the proof-theoretic ordinals of first-order theories. In this section we will develop tools in the mold of Schmerl's formula for calculating the Π^1_1 proof-theoretic ordinals of second-order theories. Throughout this section we will rely on the following analogue of Theorem 4.4.1 that is also due to Schmerl [88].

Theorem 4.4.2 (Schmerl)**.** *Let n be a natural number. Provably in EA^+, for any ordinal notation α,*

$$\mathbf{R}^\alpha_{\mathbf{\Pi}^0_\infty}(\mathsf{PA}(X)) \equiv_{\mathbf{\Pi}^0_n} \mathbf{R}^{\varepsilon_\alpha}_{\mathbf{\Pi}^0_n}(\mathsf{EA}^+(X)).$$

Note that the versions of Schmerl's formulas that we give above aren't exactly what Schmerl proved, but rather versions of the formulas that are natural given the notation of our paper. And they could be proved by either application of Schmerl's technique or Beklemishev's technique [7]. In fact in an early preprint of this paper [71, §6.2] we provided a proof of Theorem 4.4.2, however since the technique that we used wasn't new and the result is just a slight variation of [88] we removed it from the paper.

Here is a roadmap for the rest of this section. In §4.4 we prove Theorem 4.4.9 that states that

$$\mathbf{R}^\alpha_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0) \equiv_{\mathbf{\Pi}^0_\infty} \mathbf{R}^{1+\alpha}_{\mathbf{\Pi}^0_3}(\mathsf{EA}^+(X)).$$

In §4.3 we use this result to prove Theorem 4.1.6, i.e., that

$$\mathbf{R}^\alpha_{\Pi^1_1}(\mathsf{ACA}_0) \equiv_{\Pi^1_1(\Sigma^0_2)} \mathbf{R}^{\varepsilon_\alpha}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0).$$

In §4.5 we will combine Theorem 4.1.6 with the results from §4.3 (especially Theorem 4.3.2 and Theorem 4.3.3) to establish connections between iterated reflection and ordinal analysis. In particular, we will use iterated reflection principles to calculate the Π^1_1 proof-theoretic ordinals of a wide range of theories.

Before continuing, we alert the reader that many of the proofs in this section use Schmerl's technique of reflexive induction. For a description of this technique, please see §2.4.

Iterated reflection and recursive comprehension

Recall that there are no descending chains in the $\mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}$ ordering of $\Pi^1_1(\Sigma^0_2)$ sound extensions of RCA_0 (this is Theorem 4.3.3). In this subsection we investigate iterated $\Pi^1_1(\Sigma^0_2)$ reflection over the theory RCA_0. The main result of this subsection is that $\mathbf{R}^{\alpha}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0)$ is Π^1_1 conservative over $\mathbf{R}^{1+\alpha}_{\mathbf{\Pi}^0_3}(\mathsf{EA}^+(X))$. This result will be used in the next section to calculate Π^1_1 proof-theoretic ordinals of subsystems of second-order arithmetic.

Before proving the theorem we prove a few lemmas. These lemmas concern proof-theoretic properties of theories that are closed under an inference rule that we call the $\mathbf{\Delta}^0_1$ *substitution rule*.

Definition 4.4.3. Suppose φ and $\theta(x)$ are $\mathbf{\Pi}^0_\infty$ formulas that may have other free variables. We denote by $\varphi[\theta(x)]$ the result of substituting the formula $\theta(x)$ in for the free set variable X, i.e. to obtain $\varphi[\theta(x)]$ we first rename all the bounded variables of φ in order to ensure that there are no clashes with free variables of θ and then replace each atomic subformula of φ of the form $t \in X$ with $\theta(t)$.

Definition 4.4.4. We write $\mathsf{Subst}_{\mathbf{\Delta}^0_1}[\varphi]$ to denote the formula

$$\forall\theta_1(x)\forall\theta_2(x)\Big(\forall y\big(\mathsf{Tr}_{\mathbf{\Pi}^0_1}(\theta_1(y)) \leftrightarrow \mathsf{Tr}_{\mathbf{\Sigma}^0_1}(\theta_2(y))\big) \rightarrow \varphi[\mathsf{Tr}_{\mathbf{\Pi}^0_1}(\theta_1(x))]\Big).$$

A theory T is closed under the $\mathbf{\Delta}^0_1$ *substitution rule* if, for any formula $\psi(X)$, whenever $T \vdash \psi(X)$ then $T \vdash \mathsf{Subst}_{\mathbf{\Delta}^0_1}[\psi]$.

Recall that there is a translation $\varphi(X) \longmapsto \forall X\ \varphi(X)$ from the set of $\mathbf{\Pi}^0_\infty$ sentences to the set of sentences of the language of second order arithmetic. Recall also that we are regarding the pseudo-Π^1_1 language as a sublanguage of the language of second order arithmetic by identifying each pseudo Π^1_1 sentence with its translation.

Lemma 4.4.5. (EA^+) *For each $\mathbf{\Pi}^0_\infty$ sentence $\varphi(X)$ the following are equivalent.*

1. $\mathsf{RCA}_0 + \forall X\ \varphi(X)$ *is $\mathbf{\Pi}^0_\infty$ conservative over $\mathsf{I\Sigma}_1(X) + \varphi(X)$.*
2. $\mathsf{I\Sigma}_1(X) + \varphi(X)$ *is closed under the $\mathbf{\Delta}^0_1$ substitution rule.*
3. $\mathsf{I\Sigma}_1(X) + \varphi(X)$ *proves* $\mathsf{Subst}_{\mathbf{\Delta}^0_1}[\varphi]$.

Proof. We work in EA^+ and consider a $\mathbf{\Pi}^0_\infty$ sentence $\varphi(X)$.

(1) $\rightarrow$ (2): Suppose that $\mathsf{RCA}_0 + \forall X\ \varphi(X)$ is $\mathbf{\Pi}^0_\infty$ conservative over $\mathsf{I\Sigma}_1(X) + \varphi(X)$. Suppose that $\mathsf{I\Sigma}_1(X) + \varphi(X) \vdash \psi(X)$. Then $\mathsf{RCA}_0 + \forall X\ \varphi(X) \vdash \psi(X)$. Applying recursive comprehension, we derive $\mathsf{RCA}_0 + \forall X\ \varphi(X) \vdash \mathsf{Subst}_{\mathbf{\Delta}^0_1}[\psi]$. Hence, by $\mathbf{\Pi}^0_\infty$ conservativity, $\mathsf{I\Sigma}_1(X) + \varphi(X) \vdash \mathsf{Subst}_{\mathbf{\Delta}^0_1}[\psi]$.

(2) $\rightarrow$ (3): By application of the $\mathbf{\Delta}^0_1$ substitution rule to φ.

(3) → (1): Suppose that $\mathsf{I\Sigma}_1(X) + \varphi(X)$ proves $\mathsf{Subst}_{\mathbf{\Delta}^0_1}[\varphi]$. We recall the well-known ω-interpretation of RCA_0 into $\mathsf{I\Sigma}_1(X)$ wherein we interpret sets by indices for X-recursive sets; see, e.g., [92, §IX.1]. The image of the sentence $\forall X\ \varphi(X)$ under this interpretation is the sentence $\mathsf{Subst}_{\mathbf{\Delta}^0_1}[\varphi]$. This latter sentence is provable in $\mathsf{I\Sigma}_1(X) + \varphi(X)$ by assumption. Thus, this interpretation actually interprets $\mathsf{RCA}_0 + \forall X\ \varphi(X)$ in $\mathsf{I\Sigma}_1(X) + \varphi(X)$. Therefore, for any sentence $\psi(X)$, if $\mathsf{RCA}_0 + \forall X\ \varphi(X)$ proves $\forall X\ \psi(X)$, then $\mathsf{I\Sigma}_1(X) + \varphi(X)$ proves $\mathsf{Subst}_{\mathbf{\Delta}^0_1}[\psi]$, which is the image of $\forall X\ \psi(X)$ under the interpretation. Obviously, $\mathsf{I\Sigma}_1(X) + \varphi(X) \vdash \mathsf{Subst}_{\mathbf{\Delta}^0_1}[\psi] \to \psi(X)$, for any $\mathbf{\Pi}^0_\infty$ formula $\psi(X)$. Therefore, $\mathsf{RCA}_0 + \forall X\ \varphi(X)$ is $\mathbf{\Pi}^0_\infty$ conservative over $\mathsf{I\Sigma}_1(X) + \varphi(X)$. ❑

Question 4.4.6. *Combining Theorem 4.3.3 and Lemma 4.4.5 it is easy to observe that the restriction of the order $\prec_{\mathbf{\Pi}^0_3}$ to $\mathbf{\Pi}^0_3$-sound r.e. extensions of $\mathsf{I\Sigma}_1(X)$ that are closed under the $\mathbf{\Delta}^0_1$-substitution rule is well-founded. Could we drop the condition on closure under the $\mathbf{\Delta}^0_1$-substitution rule? For which n is the restriction of the order $\prec_{\mathbf{\Pi}^0_n}$ to $\mathbf{\Pi}^0_n$-sound r.e. extensions of $\mathsf{I\Sigma}_1(X)$ well-founded?*

Remark 4.4.7. We recall that $\mathsf{I\Sigma_1} \equiv \mathsf{EA}^+ + \mathsf{RFN}_{\Pi_3}(\mathsf{EA}^+)$. See, e.g., [8]. The same argument could be used to show that $\mathsf{I\Sigma_1}(X) \equiv \mathsf{EA}^+(X) + \mathsf{RFN}_{\mathbf{\Pi}^0_3}(\mathsf{EA}^+(X))$.

Lemma 4.4.8. (EA^+) *For a fixed theory T, if EA^+ proves "$T \supseteq \mathsf{I\Sigma}_1(X)$ and T is closed under the $\mathbf{\Delta}^0_1$ substitution rule," then $\mathsf{EA}^+(X) + \mathsf{RFN}_{\mathbf{\Pi}^0_3}(T)$ is closed under the $\mathbf{\Delta}^0_1$ substitution rule.*

Proof. Suppose that EA^+ proves "$T \supseteq \mathsf{I\Sigma}_1(X)$ is closed under the $\mathbf{\Delta}^0_1$ substitution rule." Let us use the name U for the theory $\mathsf{EA}^+(X) + \mathsf{RFN}_{\mathbf{\Pi}^0_3}(T)$. We want to show that U is closed under the $\mathbf{\Delta}^0_1$ substitution rule. Note that, by Remark 4.4.7, U contains $\mathsf{I\Sigma}_1(X)$. That is, $U \equiv \mathsf{I\Sigma}_1(X) + \mathsf{RFN}_{\mathbf{\Pi}^0_3}(T)$. Over $\mathsf{EA}(X)$, the reflection schema $\mathsf{RFN}_{\mathbf{\Pi}^0_3}(T)$ is equivalent to

$$\forall \varphi \in \mathbf{\Pi}^0_3 \Big(\mathsf{Pr}_T\big(\mathsf{Tr}_{\mathbf{\Pi}^0_3}(\varphi)\big) \to \mathsf{Tr}_{\mathbf{\Pi}^0_3}(\varphi) \Big).$$

Thus, by Lemma 4.4.5, it suffices to show that U proves

$$\mathsf{Subst}_{\mathbf{\Delta}^0_1}[\forall \varphi \in \mathbf{\Pi}^0_3 \Big(\mathsf{Pr}_T\big(\mathsf{Tr}_{\mathbf{\Pi}^0_3}(\varphi)\big) \to \mathsf{Tr}_{\mathbf{\Pi}^0_3}(\varphi) \Big)].$$

But since the formula $\mathsf{Pr}_T(\mathsf{Tr}_{\mathbf{\Pi}^0_3}(\varphi))$ doesn't contain occurences of X, we could push $\mathsf{Subst}_{\mathbf{\Delta}^0_1}$ under the quantifier, i.e., it will be sufficient to show that

$$U \vdash \forall \varphi \in \mathbf{\Pi}^0_3 \Big(\mathsf{Pr}_T\big(\mathsf{Tr}_{\mathbf{\Pi}^0_3}(\varphi)\big) \to \mathsf{Subst}_{\mathbf{\Delta}^0_1}[\mathsf{Tr}_{\mathbf{\Pi}^0_3}(\varphi)] \Big).$$

Observe that $\mathsf{Subst}_{\mathbf{\Delta}^0_1}[\mathsf{Tr}_{\mathbf{\Pi}^0_3}(\varphi)]$ is equivalent to a $\mathbf{\Pi}^0_3$ formula over $\mathsf{EA}(X)$. We reason as follows.

$$\begin{aligned}
&U \vdash \text{ "}T \text{ is closed under the } \mathbf{\Delta}^0_1 \text{ substitution rule," by assumption.}\\
&U \vdash \forall \varphi \in \mathbf{\Pi}^0_3 \Big(\mathsf{Pr}_T\big(\mathsf{Tr}_{\mathbf{\Pi}^0_3}(\varphi)\big) \to \mathsf{Pr}_T(\mathsf{Subst}_{\mathbf{\Delta}^0_1}[\mathsf{Tr}_{\mathbf{\Pi}^0_3}(\varphi)]) \Big)\\
&U \vdash \forall \varphi \in \mathbf{\Pi}^0_3 \Big(\mathsf{Pr}_T\big(\mathsf{Tr}_{\mathbf{\Pi}^0_3}(\varphi)\big) \to \mathsf{Subst}_{\mathbf{\Delta}^0_1}[\mathsf{Tr}_{\mathbf{\Pi}^0_3}(\varphi)] \Big) \text{ by } \mathsf{RFN}_{\mathbf{\Pi}^0_3}(T).
\end{aligned}$$

This concludes the proof of the lemma. ❑

With these lemmas on board we are ready for the proof of the main theorem of this subsection.

Theorem 4.4.9. (EA^+) *For any ordinal notation α,*

$$\mathbf{R}^{\alpha}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0) \equiv_{\mathbf{\Pi}^0_\infty} \mathbf{R}^{1+\alpha}_{\mathbf{\Pi}^0_3}(\mathsf{EA}^+(X)).$$

Proof. We prove the claim by reflexive induction. We reason in EA^+ and assume the reflexive induction hypothesis: provably in EA^+, for any $\beta \prec \alpha$,

$$\mathbf{R}^{\beta}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0) \equiv_{\mathbf{\Pi}^0_\infty} \mathbf{R}^{1+\beta}_{\mathbf{\Pi}^0_3}(\mathsf{EA}^+(X)).$$

Of course, since RCA_0 contains EA^+, this also implies that,

$$\mathsf{RCA}_0 \vdash \forall \beta \prec \alpha \Big(\mathbf{R}^{\beta}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0) \equiv_{\mathbf{\Pi}^0_\infty} \mathbf{R}^{1+\beta}_{\mathbf{\Pi}^0_3}\big(\mathsf{EA}^+(X)\big) \Big)$$

If RCA_0 proves mutual Γ conservation of two theories T and U, then $\mathsf{RFN}_\Gamma(T)$ and $\mathsf{RFN}_\Gamma(U)$ are equivalent over RCA_0. Thus, we immediately infer

$$\mathsf{RCA}_0 \vdash \forall \beta \prec \alpha \Big(\mathsf{RFN}_{\mathbf{\Pi}^0_3}\big(\mathbf{R}^{\beta}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0)\big) \leftrightarrow \mathsf{RFN}_{\mathbf{\Pi}^0_3}\big(\mathbf{R}^{1+\beta}_{\mathbf{\Pi}^0_3}\big(\mathsf{EA}^+(X)\big)\big) \Big) \tag{4.8}$$

We now reason as follows.

$$\begin{aligned}
\mathbf{R}^{\alpha}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0) &\equiv \mathsf{RCA}_0 + \bigcup_{\beta<\alpha} \mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}\big(\mathbf{R}^{\beta}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0)\big) \text{ by definition.} \\
&\equiv_{\mathbf{\Pi}^0_\infty} \mathsf{RCA}_0 + \bigcup_{\beta<\alpha} \mathsf{RFN}_{\mathbf{\Pi}^0_3}\big(\mathbf{R}^{\beta}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0)\big) \\
&\equiv \mathsf{RCA}_0 + \bigcup_{\beta<\alpha} \mathsf{RFN}_{\mathbf{\Pi}^0_3}\Big(\mathbf{R}^{1+\beta}_{\mathbf{\Pi}^0_3}\big(\mathsf{EA}^+(X)\big)\Big) \text{ by (4.8).}
\end{aligned}$$

Since $\mathbf{R}^1_{\mathbf{\Pi}^0_3}(\mathsf{EA}^+(X)) \equiv \mathsf{I}\mathbf{\Sigma}_1(X)$, we are able to show that

$$\mathbf{R}^{1+\alpha}_{\mathbf{\Pi}^0_3}\big(\mathsf{EA}^+(X)\big) \equiv \mathsf{I}\mathbf{\Sigma}_1(X) + \bigcup_{\beta<\alpha} \mathsf{RFN}_{\mathbf{\Pi}^0_3}\Big(\mathbf{R}^{1+\beta}_{\mathbf{\Pi}^0_3}\big(\mathsf{EA}^+(X)\big)\Big),$$

by the following argument:

$$\begin{aligned}
\mathbf{R}^{1+\alpha}_{\mathbf{\Pi}^0_3}\big(\mathsf{EA}^+(X)\big) &\equiv \mathbf{R}^{1}_{\mathbf{\Pi}^0_3}\big(\mathsf{EA}^+(X)\big) + \mathbf{R}^{1+\alpha}_{\mathbf{\Pi}^0_3}\big(\mathsf{EA}^+(X)\big) \text{ since } 1 \leqslant 1+\alpha. \\
&\equiv \mathsf{I}\mathbf{\Sigma}_1(X) + \mathbf{R}^{1+\alpha}_{\mathbf{\Pi}^0_3}\big(\mathsf{EA}^+(X)\big) \text{ since } \mathbf{R}^{1}_{\mathbf{\Pi}^0_3}\big(\mathsf{EA}^+(X)\big) \equiv \mathsf{I}\mathbf{\Sigma}_1(X). \\
&\equiv \mathsf{I}\mathbf{\Sigma}_1(X) + \bigcup_{\beta<\alpha} \mathsf{RFN}_{\mathbf{\Pi}^0_3}\Big(\mathbf{R}^{1+\beta}_{\mathbf{\Pi}^0_3}\big(\mathsf{EA}^+(X)\big)\Big) \text{ by definition.}
\end{aligned}$$

Hence in order to finish the proof of the lemma it will be enough to show that

$$\mathsf{I\Sigma}_1(X) + \bigcup_{\beta<\alpha} \mathsf{RFN}_{\mathbf{\Pi}^0_3}(\mathbf{R}^{1+\beta}_{\mathbf{\Pi}^0_3}(\mathsf{EA}^+(X))) \equiv_{\mathbf{\Pi}^0_\infty} \mathsf{RCA}_0 + \bigcup_{\beta<\alpha} \mathsf{RFN}_{\mathbf{\Pi}^0_3}(\mathbf{R}^{1+\beta}_{\mathbf{\Pi}^0_3}(\mathsf{EA}^+(X))),$$

which, by Lemma 4.4.5, can be achieved by proving that

$$\mathsf{I\Sigma}_1(X) + \bigcup_{\beta<\alpha} \mathsf{RFN}_{\mathbf{\Pi}^0_3}(\mathbf{R}^{1+\beta}_{\mathbf{\Pi}^0_3}(\mathsf{EA}^+(X)))$$

is closed under the $\mathbf{\Delta}^0_1$ substitution rule. We will prove this closedness in the rest of the proof.

By a usual compactness argument, it will be enough to show that $\mathsf{I\Sigma}_1(X)$ is closed under the $\mathbf{\Delta}^0_1$ substitution rule and that for each $\beta \prec \alpha$ the theories $\mathsf{I\Sigma}_1(X)+\mathsf{RFN}_{\mathbf{\Pi}^0_3}(\mathbf{R}^{1+\beta}_{\mathbf{\Pi}^0_3}(\mathsf{EA}^+(X)))$ are closed under the $\mathbf{\Delta}^0_1$ substitution rule. Closure of $\mathsf{I\Sigma}_1(X)$ under the $\mathbf{\Delta}^0_1$ substitution rule follows directly from Lemma 4.4.5.

By Lemma 4.4.5, we infer that, for each $\beta \prec \alpha$, $\mathbf{R}^{1+\beta}_{\mathbf{\Pi}^0_3}(\mathsf{EA}^+(X))$ is EA^+ provably closed under the $\mathbf{\Delta}^0_1$ substitution rule. Thus, by Lemma 4.4.8, we infer that for each $\beta \prec \alpha$,

$$\mathsf{EA}^+(X) + \mathsf{RFN}_{\mathbf{\Pi}^0_3}(\mathbf{R}^{1+\beta}_{\mathbf{\Pi}^0_3}(\mathsf{EA}^+(X)))$$

is closed under the $\mathbf{\Delta}^0_1$ substitution rule. Since $\mathsf{EA}^+(X)+\mathsf{RFN}_{\mathbf{\Pi}^0_3}(\mathbf{R}^{1+\beta}_{\mathbf{\Pi}^0_3}(\mathsf{EA}^+(X))) \sqsupseteq \mathsf{I\Sigma}_1(X)$, the theory $\mathsf{I\Sigma}_1(X) + \mathsf{RFN}_{\mathbf{\Pi}^0_3}(\mathbf{R}^{1+\beta}_{\mathbf{\Pi}^0_3}(\mathsf{EA}^+(X)))$ is closed under the $\mathbf{\Delta}^0_1$ substitution rule. This concludes the proof of the lemma. ❑

Iterated reflection and arithmetical comprehension

In this subsection we investigate the relationship between iterated Π^1_1 reflection over ACA_0 and iterated $\Pi^1_1(\Sigma^0_2)$ reflection over RCA_0. The main theorem of this subsection is that $\mathbf{R}^{\alpha}_{\Pi^1_1}(\mathsf{ACA}_0)$ is $\Pi^1_1(\Sigma^0_2)$ conservative over $\mathbf{R}^{\varepsilon_\alpha}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0)$. The proof of the main theorem of this subsection is similar to the proof of Theorem 4.4.9. For our first step towards this result, we establish a conservation theorem relating extensions of ACA_0 with extensions of $\mathsf{PA}(X)$.

There is a standard semantic argument that ACA_0 is conservative over PA (see, e.g., [92, Section IX.1]). We will present a version of this argument for extensions of ACA_0 by Π^1_1 sentences. Moreover we ensure that this conservation result is provable in ACA_0. Before presenting the argument, we will say a bit about how we will formalize model theory within ACA_0 for the purposes of our argument.

We will reason in ACA_0 and use the formalization of model theory from [92, Section II.8]. Recall that according to formalization from [92, Section II.8] a model $\mathfrak{M}$ essentially is a set that encodes the domain of $\mathfrak{M}$ (which is by necessity a subset of $\mathbb{N}$) and the full satisfaction relation for $\mathfrak{M}$ (the latter essentially is the elementary diagram of the model $\mathfrak{M}$). Note that if one would require $\mathfrak{M}$ contain information only about the satisfaction of atomic formulas,

rather then all formulas, the resulting notion of a model would be weaker. This is due to the fact that in ACA_0, unlike in stronger theories, it is not always possible to recover the elementary diagram of a model from its atomic diagram.

Due to this limitation, in ACA_0 it is sometimes (including in our proof) useful to employ weak models [92, Definition II.8.9]. A *weak model* $\mathfrak{M}$ of a theory T is a set that encodes the domain of $\mathfrak{M}$ and a partial satisfaction relation for $\mathfrak{M}$ that is defined only on Boolean combinations of subformulas of formulas used in axioms of T such that all the axioms of T are according to this satisfaction relation. The key fact that we use is that ACA_0 proves that any theory that has a weak model is consistent [92, Theorem II.8.10].

Lemma 4.4.10. (ACA_0) *Let $\varphi(X), \psi(X)$ be $\mathbf{\Pi}^0_\infty$. If $\mathsf{ACA}_0 + \forall X\ \varphi(X) \vdash \forall X\ \psi(X)$ then $\mathsf{PA}(X) + \{\varphi[\theta] : \theta(x)$ is $\mathbf{\Pi}^0_\infty\} \vdash \psi(X)$, where θ could contain additional variables.*

Proof. We reason in ACA_0. We denote by U the theory $\mathsf{PA}(X) + \{\varphi[\theta] : \theta$ is $\mathbf{\Pi}^0_\infty\}$. Let us consider any $\psi(X)$ such that $U \nvdash \psi(X)$. To prove the lemma we need to show that $\mathsf{ACA}_0 + \forall X\ \varphi(X) \nvdash \forall X\ \psi(X)$.

There is a model $\mathfrak{M}$ of $U + \neg\psi(X)$. Note that here X is just a unary predicate. We enrich $\mathfrak{M}$ by adding, as the family $\mathcal{S}$ of second-order objects, all the sets defined in $\mathfrak{M}$ by $\mathbf{\Pi}^0_\infty$ formulas that may contain additional parameters from the model.

Let us first show how we could finish the proof without ensuring that our argument could be formalized in ACA_0 and only then indicate how to carry out the formalization. Indeed, it is easy to see that the second-order structure $(\mathfrak{M}, \mathcal{S})$ satisfies $\mathsf{ACA}_0 + \forall X\ \varphi(X)$: the presence of the full induction schema in U guarantees that $(\mathfrak{M}, \mathcal{S})$ satisfies set induction, our definition of $\mathcal{S}$ guarantees that arithmetical comprehension holds in $(\mathfrak{M}, \mathcal{S})$, and the fact that we had axioms $\{\varphi[\theta] : \theta$ is $\mathbf{\Pi}^0_\infty\}$ in U guarantees that $\forall X\ \varphi(X)$ holds in $(\mathfrak{M}, \mathcal{S})$. And since $\psi(X)$ failed in $\mathfrak{M}$, the sentence $\forall X\ \psi(X)$ fails in $(\mathfrak{M}, \mathcal{S})$. Therefore, $\mathsf{ACA}_0 + \forall X\ \varphi(X) \nvdash \forall X\ \psi(X)$.

Now let us show how to formalize the latter argument in ACA_0. We want to show that we could extend $(\mathfrak{M}, \mathcal{S})$ to a weak model of $\mathsf{ACA}_0 + \forall X\ \varphi(X)$. From the satisfaction relation for $\mathfrak{M}$ we can trivially construct the partial satisfaction relation for $(\mathfrak{M}, \mathcal{S})$ that covers all Π^0_∞ formulas with parameters from $(\mathfrak{M}, \mathcal{S})$. And since we are working in ACA_0, using arithmetical comprehension for every (externally) fixed n we could expand the latter partial satisfaction relation to all the formulas constructed from Π^0_∞ formulas by arbitrary use of propositional connectives and with introduction of at most n quantifier alternations. For $n = 2$ this expanded partial satisfaction relation covers all the axioms of $\mathsf{ACA}_0 + \forall X\ \varphi(X) + \neg\forall X\ \psi(X)$. Now after we constructed this satisfaction relation we could proceed as in the paragraph above and show that in this partial satisfaction realtion all the axioms of $\mathsf{ACA}_0 + \forall X\ \varphi(X) + \neg\forall X\ \psi(X)$ are true. Hence we have a weak model of $\mathsf{ACA}_0 + \forall X\ \varphi(X)$. Therefore, $\mathsf{ACA}_0 + \forall X\ \varphi(X) \nvdash \forall X\ \psi(X)$. ❑

Remark 4.4.11. Although we don't provide a proof here, we note that with some additional care it is possible to establish Lemma 4.4.10 in EA^+ by appealing to the Π_2-conservativity of $\mathsf{WKL}^\star_0$ + "super-exponentiation is total" over EA^+, see [93] for the Π_2-conservativity of $\mathsf{WKL}^\star_0$

over EA. But it isn't possible to prove this result in EA since ACA_0 enjoys non-elementary speed-up over PA.

Definition 4.4.12. We say that a pseudo Π^1_1 theory $T(X)$ is *closed under substitution* if whenever $T \vdash \varphi(X)$ then also $T \vdash \varphi[\theta(x)]$ for any $\mathbf{\Pi}^0_\infty$ formula θ.

Lemma 4.4.13. *If a theory T proves every substitution variant of its own axioms, then T is closed under substitution.*

Proof. Suppose that T proves every substitution variant of its own axioms. Let θ be a $\mathbf{\Pi}^0_\infty$ formula and let $\varphi(X)$ be a theorem of T. Since $\varphi(X)$ is a theorem of T, there is some finite conjunction $A_T(X)$ of axioms of T such that the sentence

$$A_T(X) \to \varphi(X)$$

is a theorem of pure logic. Since pure logic is closed under substitution, the sentence

$$A_T[\theta(x)] \to \varphi[\theta(x)]$$

is also a theorem of pure logic. Since T proves every substitution variant of its own axioms, T proves $A_T[\theta(x)]$, whence T proves $\varphi[\theta(x)]$. ❑

Lemma 4.4.14. $\mathsf{PA}(X) + \mathbf{R}^\alpha_{\mathbf{\Pi}^0_\infty}(\mathsf{PA}(X))$ *is closed under substitution.*

Proof. We prove the claim by reflexive induction. We reason within EA^+ and assume the reflexive induction hypothesis: provably in EA^+, for all $\beta \prec \alpha$, $\mathsf{PA}(X) + \mathbf{R}^\beta_{\mathbf{\Pi}^0_\infty}(\mathsf{PA}(X))$ is closed under substitution. First we note that

$$\mathbf{R}^\alpha_{\mathbf{\Pi}^0_\infty}\big(\mathsf{PA}(X)\big) \equiv \mathsf{PA}(X) + \bigcup_{\beta<\alpha} \mathsf{RFN}_{\mathbf{\Pi}^0_\infty}\Big(\mathbf{R}^\beta_{\mathbf{\Pi}^0_\infty}\big(\mathsf{PA}(X)\big)\Big).$$

For $\beta \prec \alpha$ let us denote by S_β the theory

$$\mathsf{PA}(X) + \mathsf{RFN}_{\mathbf{\Pi}^0_\infty}\Big(\mathbf{R}^\beta_{\mathbf{\Pi}^0_\infty}\big(\mathsf{PA}(X)\big)\Big).$$

To prove that $\mathbf{R}^\alpha_{\mathbf{\Pi}^0_\infty}\big(\mathsf{PA}(X)\big)$ is closed under substitution it suffices to prove that, for every $\beta \prec \alpha$, S_β is closed under substitution.

By Lemma 4.4.13, to prove that S_β is closed under substitution, it suffices to show that S_β proves every substitution-variant of its own axioms. Let us use the name U_β to denote the theory $\mathbf{R}^\beta_{\mathbf{\Pi}^0_\infty}\big(\mathsf{PA}(X)\big)$. An axiom of the theory S_β is either an axiom of $\mathsf{PA}(X)$ or is a sentence of the form $\forall\vec{y}\big(Pr_{U_\beta}\big(\varphi(X,\vec{y})\big) \to \varphi(X,\vec{y})\big)$. Already the theory $\mathsf{PA}(X)$ proves every substitutional instance of its own axioms. By the reflexive induction hypothesis, U_β is provably closed under substitution. So S_β proves $\forall\vec{y}\big(Pr_{U_\beta}\big(\varphi(X,\vec{y})\big) \to \varphi(\theta,\vec{y})\big)$ for any formula θ. This is to say that S_β proves every substitution instance of its axioms. ❑

Remark 4.4.15. It follows from the lemma that the theories $\mathsf{PA}(X) + \{\mathbf{R}^{\alpha}_{\mathbf{\Pi}^0_\infty}(\mathsf{PA}(X))[\theta] : \theta \in \mathbf{\Pi}^0_\infty\}$ and $\mathsf{PA}(X) + \mathbf{R}^{\alpha}_{\mathbf{\Pi}^0_\infty}(\mathsf{PA}(X))$ are equivalent. We will make use of this observation in the proof of Lemma 4.4.16.

Most of the work towards proving the main theorem of this section is contained in the proof of the following key lemma.

Lemma 4.4.16. $\mathbf{R}^{\alpha}_{\Pi^1_1}(\mathsf{ACA}_0)$ *is* $\mathbf{\Pi}^0_\infty$ *conservative over* $\mathbf{R}^{\alpha}_{\mathbf{\Pi}^0_\infty}(\mathsf{PA}(X))$.

Proof. We prove the claim by reflexive induction. We reason within ACA_0 and assume the reflexive induction hypothesis: provably in ACA_0, for all $\beta \prec \alpha$, $\mathbf{R}^{\beta}_{\Pi^1_1}(\mathsf{ACA}_0)$ is $\mathbf{\Pi}^0_\infty$ conservative over $\mathbf{R}^{\beta}_{\mathbf{\Pi}^0_\infty}(\mathsf{PA}(X))$. This means that, provably in ACA_0, for any $\beta \prec \alpha$, $\mathbf{\Pi}^0_\infty$ reflection over $\mathbf{R}^{\beta}_{\Pi^1_1}(\mathsf{ACA}_0)$ is equivalent to $\mathbf{\Pi}^0_\infty$ reflection over $\mathbf{R}^{\beta}_{\mathbf{\Pi}^0_\infty}(\mathsf{PA}(X))$. That is:

$$\mathsf{ACA}_0 \vdash \forall \beta \prec \alpha \Big(\mathsf{RFN}_{\mathbf{\Pi}^0_\infty}\big(\mathbf{R}^{\beta}_{\Pi^1_1}(\mathsf{ACA}_0)\big) \leftrightarrow \mathsf{RFN}_{\mathbf{\Pi}^0_\infty}\big(\mathbf{R}^{\beta}_{\mathbf{\Pi}^0_\infty}(\mathsf{PA}(X))\big)\Big) \tag{4.9}$$

We reason as follows.

$$\begin{aligned}
\mathbf{R}^{\alpha}_{\Pi^1_1}(\mathsf{ACA}_0) &\equiv \mathsf{ACA}_0 + \bigcup_{\beta<\alpha} \mathsf{RFN}_{\Pi^1_1}\big(\mathbf{R}^{\beta}_{\Pi^1_1}(\mathsf{ACA}_0)\big) \text{ by definition.} \\
&\equiv_{\mathbf{\Pi}^0_\infty} \mathsf{ACA}_0 + \bigcup_{\beta<\alpha} \mathsf{RFN}_{\mathbf{\Pi}^0_\infty}\big(\mathbf{R}^{\beta}_{\Pi^1_1}(\mathsf{ACA}_0)\big) \\
&\equiv \mathsf{ACA}_0 + \bigcup_{\beta<\alpha} \mathsf{RFN}_{\mathbf{\Pi}^0_\infty}\Big(\mathbf{R}^{\beta}_{\mathbf{\Pi}^0_\infty}\big(\mathsf{PA}(X)\big)\Big) \text{ by (4.9).} \\
&\equiv \mathsf{ACA}_0 + \bigcup_{\beta<\alpha} \mathsf{RFN}_{\mathbf{\Pi}^0_\infty}\Big(\mathsf{PA}(X) + \mathbf{R}^{\beta}_{\mathbf{\Pi}^0_\infty}\big(\mathsf{PA}(X)\big)\Big) \text{ by definition.} \\
&\equiv_{\mathbf{\Pi}^0_\infty} \mathsf{PA}(X) + \bigcup_{\beta<\alpha\ \theta\in\mathbf{\Pi}^0_\infty} \mathsf{RFN}_{\mathbf{\Pi}^0_\infty}\Big(\mathsf{PA}(X) + \mathbf{R}^{\beta}_{\mathbf{\Pi}^0_\infty}\big(\mathsf{PA}(X)\big)\Big)[\theta] \text{ by Lemma 4.4.10.} \\
&\equiv_{\mathbf{\Pi}^0_\infty} \mathsf{PA}(X) + \bigcup_{\beta<\alpha} \mathsf{RFN}_{\mathbf{\Pi}^0_\infty}\Big(\mathsf{PA}(X) + \mathbf{R}^{\beta}_{\mathbf{\Pi}^0_\infty}\big(\mathsf{PA}(X)\big)\Big) \text{ by Remark 4.4.15.} \\
&\equiv_{\mathbf{\Pi}^0_\infty} \mathbf{R}^{\alpha}_{\mathbf{\Pi}^0_\infty}(\mathsf{PA}(X)) \text{ by definition.}
\end{aligned}$$

This concludes the proof. ❑

The proof of the the main theorem of this section is now straightforward, given Theorem 4.4.9 and Lemma 4.4.16.

Theorem 4.4.17. $\mathbf{R}^{\alpha}_{\Pi^1_1}(\mathsf{ACA}_0)$ *is* $\Pi^1_1(\Sigma^0_2)$ *conservative over* $\mathbf{R}^{\varepsilon_\alpha}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0)$.

Proof. We reason as follows.

$$\begin{aligned}
\mathbf{R}^{\alpha}_{\Pi^1_1}(\mathsf{ACA}_0) &\equiv_{\mathbf{\Pi}^0_\infty} \mathbf{R}^{\alpha}_{\mathbf{\Pi}^0_\infty}(\mathsf{PA}(X)) \text{ by Lemma 4.4.16.} \\
&\equiv_{\mathbf{\Pi}^0_3} \mathbf{R}^{\varepsilon_\alpha}_{\mathbf{\Pi}^0_3}(\mathsf{EA}^+(X)) \text{ by Theorem 4.4.2.} \\
&\equiv_{\mathbf{\Pi}^0_3} \mathbf{R}^{\varepsilon_\alpha}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0) \text{ by Theorem 4.4.9.}
\end{aligned}$$

Note for each $\Pi^1_1(\Sigma^0_2)$ sentence φ we could find a $\mathbf{\Pi}^0_3$ sentence φ' such that RCA_0 proves the equivalence of φ and (the translation into the second order language of) φ'. Thus moreover we have

$$\mathbf{R}^{\alpha}_{\Pi^1_1}(\mathsf{ACA}_0) \equiv_{\Pi^1_1(\Sigma^0_2)} \mathbf{R}^{\varepsilon_\alpha}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0).$$

This completes the proof of the theorem. ❑

4.5 Reflection ranks and proof-theoretic ordinals

In this section we introduce the notion of *reflection rank*. We then use the results from the previous section to establish connections between reflection ranks and Π^1_1 proof-theoretic ordinals.

Reflection ranks

Recall that the reflection order $\prec_{\Pi^1_1}$ on r.e. extensions of ACA_0 is:

$$T_1 \prec_{\Pi^1_1} T_2 \iff^{\text{def}} T_2 \vdash \mathsf{RFN}_{\Pi^1_1}(T_1).$$

For a theory $T \supseteq \mathsf{ACA}_0$ we define the *reflection rank* $|T|_{\mathsf{ACA}_0} \in \mathbf{On} \cup \{\infty\}$ as the rank of T in the order $\prec_{\Pi^1_1}$.

Remark 4.5.1. We recall that as usual the rank function $\rho\colon A \to \mathbf{On} \cup \{\infty\}$ for a binary relation $(A, \lhd)$ is the only function such that $\rho(a) = \sup\{\rho(b) + 1 \mid b \lhd a\}$. Here the linear order $<$ on ordinals is extended to the class $\mathbf{On} \cup \{\infty\}$ by puting $\boldsymbol{\alpha} < \infty$, for all $\boldsymbol{\alpha} \in \mathbf{On}$. The operation $\boldsymbol{\alpha} \mapsto \boldsymbol{\alpha} + 1$ is extended to the class $\mathbf{On} \cup \{\infty\}$ by putting $\infty + 1 = \infty$. Note that $\rho(a) \in \mathbf{On}$ iff the cone $\{b \mid b \lhd a\}$ is well-founded with respect to $\lhd$.

Recall that Theorem 4.3.2 states that $|T|_{\mathsf{ACA}_0} \in \mathbf{On}$, for Π^1_1-sound T.

We will also consider the more general notion of reflection rank with respect to some other base theories. For second-order theories $U \supseteq \mathsf{RCA}_0$ we consider the reflection order $\prec_{\Pi^1_1(\Sigma^0_2)}$:

$$U_1 \prec_{\Pi^1_1(\Sigma^0_2)} U_2 \iff^{\text{def}} U_2 \vdash \mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(U_1).$$

Let us consider some base theory $T_0 \supseteq \mathsf{RCA}_0$. We define the set $\mathcal{E}\text{-}T_0$ of all theories U such that EA proves that $U \supseteq T_0$. For $U \in \mathcal{E}\text{-}T_0$ we denote by $|U|_{T_0}$ the rank of U in the order $(\mathcal{E}\text{-}T_0, \prec_{\Pi^1_1(\Sigma^0_2)})$. Note that $\Pi^1_1(\Sigma^0_2)$-sound extensions of T_0 have a well-founded rank in this ordering by Theorem 4.3.3.

Remark 4.5.2. For a theory T_0 given by a finite list of axioms the set $\mathcal{E}\text{-}T_0$ coincides with the set of all U such that $U \supseteq T_0$. Indeed, for any T_0 given by a finite list of axioms we have a Σ_1 formula in EA that expresses $U \supseteq T_0$ with U as a parameter (the Σ_1 formula states that there is a U-proof of the conjunction of all the axioms of T_0).

Remark 4.5.3. The definition of the rank $|T|_{\mathsf{ACA}_0}$ given in the beginning of the section coincides with the more general definition of rank, since in ACA_0 each Π^1_1 formula is equivalent to a $\Pi^1_1(\Sigma^0_2)$-formula and hence for any $T \supseteq \mathsf{ACA}_0$,

$$\mathsf{ACA}_0 \vdash \mathsf{RFN}_{\Pi^1_1}(T) \leftrightarrow \mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(T).$$

Straightforwardly from Theorem 4.3.3 we get the following.

Corollary 4.5.4. *If $U \supseteq \mathsf{RCA}_0$ is $\Pi^1_1(\Sigma^0_2)$-sound, then the rank $|U|_{\mathsf{RCA}_0} \in \mathbf{On}$. Hence for each $T_0 \supseteq \mathsf{RCA}_0$ and $\Pi^1_1(\Sigma^0_2)$-sound theory $U \in \mathcal{E}$-T_0 we have $|U|_{T_0} \in \mathbf{On}$.*

Remark 4.5.5. The converse of Corollary 4.5.4 is not true, there are $\Pi^1_1(\Sigma^0_2)$ unsound theories whose rank is an ordinal. In particular, for each consistent theory $T_0 \supseteq \mathsf{RCA}_0$, we have $|T_0 + \neg\mathsf{Con}(T_0)|_{T_0} = 0$. Indeed, assume $T_0 + \neg\mathsf{Con}(T_0) \vdash \mathsf{RFN}_{\Pi^1_1}(U)$, for some $U \in \mathcal{E}$-T_0. Then

$$\begin{aligned} T_0 + \neg\mathsf{Con}(T_0) &\vdash \mathsf{RFN}_{\Pi^1_1}(T_0) \\ &\vdash \mathsf{Con}(T_0) \\ &\vdash \bot. \end{aligned}$$

But by Gödel's Second Incompleteness Theorem $T_0 + \neg\mathsf{Con}(T_0)$ is consistent. This is to say that, though $T_0 + \neg\mathsf{Con}(T_0)$ is not Π^1_1 sound, $|T_0 + \neg\mathsf{Con}(T_0)|_{T_0} \in \mathbf{On}$.

Note that later we will introduce a notion of robust reflection rank that enjoys much better behavior and, in particular, satisfies the converse of Corollary 4.5.4.

Recall that for an ordinal notation α we denote by $|\alpha| \in \mathbf{On} \cup \{\infty\}$ the rank of the ordinal notation α in the order $\prec$.

The main proposition proved in this subsection is the following:

Proposition 4.5.6. *For each $\Pi^1_2(\Pi^0_2)$-sound theory T_0 and ordinal notation α:*

$$|\mathbf{R}^{\alpha}_{\Pi^1_1(\Sigma^0_2)}(T_0)|_{T_0} = |\alpha|.$$

In order to prove the proposition we first establish some lemmas. We state without proof the following two lemmas, the truth of which may easily be verified.

Lemma 4.5.7 (RCA_0)**.**
If T is $\Pi^1_2(\Pi^0_2)$ sound and α is a well-ordering, then $\mathbf{R}^{\alpha}_{\Pi^1_1(\Sigma^0_2)}(T)$ is $\Pi^1_1(\Sigma^0_2)$ sound.

Lemma 4.5.8 (RCA_0)**.** *If T is $\Pi^1_1(\Sigma^0_2)$ sound and φ is a true $\Sigma^1_1(\Pi^0_2)$ formula, then $T + \varphi$ is $\Pi^1_1(\Sigma^0_2)$ sound.*

Before proving proposition 4.5.6, we prove the following lemma.

Lemma 4.5.9. *If $|U|_{T_0} > |\alpha|$ then there is a true $\Sigma^1_1(\Pi^0_2)$ sentence φ such that*

$$U + \varphi \vdash \mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathbf{R}^{\alpha}_{\Pi^1_1(\Sigma^0_2)}(T_0)). \tag{4.10}$$

Proof. We prove the lemma by transfinite induction on $|\alpha|$. Since $|U|_{T_0} > |\alpha|$, there is a $V \in \mathcal{E}\text{-}T_0$ such that $U \vdash \mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(V)$ and $|V|_{T_0} \geqslant |\alpha|$. By the induction hypothesis there are true $\Sigma^1_1(\Pi^0_2)$ sentences φ_β, for all $\beta \prec \alpha$, such that

$$V + \varphi_\beta \vdash \mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathbf{R}^\beta_{\Pi^1_1(\Sigma^0_2)}(T_0)).$$

We now formalize the latter fact by a single $\Sigma^1_1(\Pi^0_2)$ sentence φ, which states that there is a sequence of Π^0_2 formulas $\langle \psi_\beta(Y) \mid \beta \prec \alpha \rangle$ without free variables other that Y and sequence of sets $\langle S_\beta \mid \beta \prec \alpha \rangle$ such that

- for all β, the formula $\psi_\beta(Y)$ holds on $Y = S_\beta$;
- for all β, we have $V + \exists Y\ \psi_\beta(Y) \vdash \mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathbf{R}^\beta_{\Pi^1_1(\Sigma^0_2)}(T_0))$.

It is easy to see that indeed we could form a $\Sigma^1_1(\Pi^0_2)$ sentence φ constituting the desired formalization.

Now let us show that φ is true. Without loss of generality, we may assume that each φ_β is of the form $\exists Y \theta_\beta(Y)$, where all $\theta_\beta(Y)$ are Π^0_2-formulas. We put each ψ_β to be θ_β and for each $\beta \prec \alpha$ we choose S_β so that $\theta_\beta(Y)$ holds on $Y = S_\beta$. Thus we see that φ is true.

We establish (4.10) by reasoning in $U + \varphi$ and showing that the theory $\mathbf{R}^\alpha_{\Pi^1_1(\Sigma^0_2)}(T_0)$ is $\Pi^1_1(\Sigma^0_2)$-sound. It is enough for us to establish the $\Pi^1_1(\Sigma^0_2)$-soundness of each finite subtheory of $\mathbf{R}^\alpha_{\Pi^1_1(\Sigma^0_2)}(T_0)$, i.e., each theory

$$T_0 + \mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathbf{R}^\beta_{\Pi^1_1(\Sigma^0_2)}(T_0)),$$

for $\beta \prec \alpha$. We know (from U) that V is $\Pi^1_1(\Sigma^0_2)$-sound. And also (from φ) we have a Π^0_2-formula $\psi_\beta(Y)$ such that

$$V + \exists Y\ \psi_\beta(Y) \vdash \mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathbf{R}^\beta_{\Pi^1_1(\Sigma^0_2)}(T_0))$$

and a set S_β such that $\psi_\beta(S_\beta)$ holds. From the $\Pi^1_1(\Sigma^0_2)$-soundness of V we infer the $\Pi^1_1(\Sigma^0_2)$-soundness of $V + \exists Y\ \psi_\beta(Y)$. Therefore $T_0 + \mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathbf{R}^\beta_{\Pi^1_1(\Sigma^0_2)}(T_0))$ is $\Pi^1_1(\Sigma^0_2)$-sound. ❑

We are now in a position to prove Proposition 4.5.6.

Proof. First let us notice that $|\mathbf{R}^\alpha_{\Pi^1_1(\Sigma^0_2)}(T_0)|_{T_0} \geqslant |\alpha|$. Indeed this inequality holds since there is a homomorphism $\beta \mapsto \mathbf{R}^\beta_{\Pi^1_1(\Sigma^0_2)}(T_0)$ of the low $\prec$-cone of α (the order $(\{\beta \mid \beta \leq \alpha\}, \prec)$) to the low $\prec_{\Pi^1_1(\Sigma^0_2)}$-cone of $\mathbf{R}^\alpha_{\Pi^1_1(\Sigma^0_2)}(T_0)$ in $\mathcal{E}\text{-}T_0$.

Now assume for a contradiction that $|\mathbf{R}^\alpha_{\Pi^1_1(\Sigma^0_2)}(T_0)|_{T_0} > |\alpha|$. In this case by Lemma 4.5.9 we have

$$\mathbf{R}^\alpha_{\Pi^1_1(\Sigma^0_2)}(T_0) + \varphi \vdash \mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathbf{R}^\alpha_{\Pi^1_1(\Sigma^0_2)}(T_0)),$$

for some true $\Sigma^1_1(\Pi^0_2)$ sentence φ. We derive

$$\begin{aligned}\mathbf{R}^{\alpha}_{\Pi^1_1(\Sigma^0_2)}(T_0) + \varphi &\vdash \mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathbf{R}^{\alpha}_{\Pi^1_1(\Sigma^0_2)}(T_0) + \varphi)\\ &\vdash \mathsf{Con}(\mathbf{R}^{\alpha}_{\Pi^1_1(\Sigma^0_2)}(T_0) + \varphi).\end{aligned}$$

So $\mathbf{R}^{\alpha}_{\Pi^1_1(\Sigma^0_2)}(T_0) + \varphi$ is inconsistent by Gödel's Second Incompleteness Theorem. Yet by Lemma 4.5.7, since T_0 is $\Pi^1_2(\Pi^0_2)$ sound, $\mathbf{R}^{\alpha}_{\Pi^1_1(\Sigma^0_2)}(T_0)$ is $\Pi^1_1(\Sigma^0_2)$ sound. Thus, by Lemma 4.5.8, $\mathbf{R}^{\alpha}_{\Pi^1_1(\Sigma^0_2)}(T_0) + \varphi$ is consistent. This is a contradiction. ❑

Proof-theoretic ordinals

For a theory $T \supseteq \mathsf{RCA}_0$ we write $|T|_{\mathsf{WO}}$ to denote the Π^1_1 *proof-theoretic ordinal* of T, which we define as the supremum of the ranks $|\alpha|$ of ordinal notations α such that $T \vdash \mathsf{WO}(\alpha)$. The formula $\mathsf{WO}(\alpha)$ is

$$\forall X((\exists \beta \prec \alpha)\ \beta \in X \rightarrow (\exists \beta \prec \alpha)(\beta \in X \wedge (\forall \gamma \prec \beta)\gamma \notin X)).$$

Remark 4.5.10. One may also define $|T|_{\mathsf{WO}}$ for second-order theories in terms of primitive recursive well-orders (alternatively recursive well-orders), i.e., $|T|_{\mathsf{WO}}$ then would be defined as the supremum of order types of primitive recursive (T-provably recursive) binary relations $\rhd$ such that $T \vdash \mathsf{WO}(\rhd)$. If T proves the well-orderedness of an ill-founded relation then this supremum by definition is ∞. We note that our definition coincides with the definitions above for $T \supseteq \mathsf{RCA}_0$. The connection between presentations of ordinals of various degrees of "niceness" is extensively discussed in M. Rathjen's survey [80], and the equivalence under consideration could be proved by a slight extension of the proof of [80, Proposition 2.19(i)].[3]

Theorem 4.5.11. $|\mathbf{R}^{\alpha}_{\Pi^1_1}(\mathsf{ACA}_0)|_{\mathsf{WO}} = |\varepsilon_\alpha|$.

In order to prove the theorem we first establish the following lemma:

Lemma 4.5.12. *For each* α

1. *the theory* ACA_0 *proves* $\mathsf{WO}(\alpha) \rightarrow \mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathbf{R}^{\alpha}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0))$*;*
2. *the theory* ACA^+_0 *proves* $\mathsf{WO}(\alpha) \rightarrow \mathsf{RFN}_{\Pi^1_1}(\mathbf{R}^{\alpha}_{\Pi^1_1}(\mathsf{ACA}_0))$.

We will derive Lemma 4.5.12 from the more general Lemma 4.5.13.

We will follow Simpson's formalization of countable coded models of the language of second-order arithmetic [92, Definition VII.2.1]. Under this definition a countable coded ω-model $\mathfrak{M}$ is a code for a countable family $W_0, W_1, \ldots$ of subsetes of $\mathbb{N}$, where $\{W_0, W_1, \ldots\}$

[3]The proof of [80, Proposition 2.19(i)] implicitly uses Σ_1-collection inside the theory T, although the claim is stated for all T containing PRA. But this issue doesn't affect the theories that we are interested in since $\mathsf{RCA}_0 \vdash \mathsf{B}\Sigma_1$

is the $\mathfrak{M}$-domain for sets of naturals. We note that the property "$\mathfrak{M}$ is a countable coded ω-model" is arithmetical. The expression $X \in \mathfrak{M}$ denotes the natural Σ^0_2 formula that expresses the fact that the set X is coded in a model $\mathfrak{M}$ (i.e. it is one of $X = W_i$, for some i). For each fixed second-order formula $\varphi(X_1, \dots, X_n, x_1, \dots, x_n)$ the expression $\mathfrak{M} \models \varphi(X_1, \dots, X_n, x_1, \dots, x_n)$ denotes the natural second-order formula that expresses that $\mathfrak{M}$ is a countable coded ω-model, sets $X_1, \dots, X_n$ are coded in $\mathfrak{M}$, and $\varphi(X_1, \dots, X_n, x_1, \dots, x_n)$ is true in $\mathfrak{M}$. We express the fact that that $\varphi(X_1, \dots, X_n, x_1, \dots, x_n)$ is true in $\mathfrak{M}$ by relativizing second-order quatifiers $\forall X$ and $\exists X$ to $\forall X \in \mathfrak{M}$ and $\exists X \in \mathfrak{M}$. Note that the latter quantifiers are in fact just first-order quantifiers. Hence $\mathfrak{M} \models \varphi(\vec{X}, \vec{x})$ is equivalent to a Π^0_m-formula, where m depends only on the depth of quantifier alternations in φ. For a fixed theory T given by a finite list of axioms, by $\mathfrak{M} \models T$ we mean the formula $\mathfrak{M} \models \varphi$, where φ is the conjunction of all the axioms of T.

For each theory $T_0 \supseteq \mathsf{RCA}_0$ given by a finite list of axioms we denote by T_0^+ the theory $T_0 + \mathsf{ACA}_0 +$ "every set is contained in an ω-model of T_0." We use this notation by analogy with ACA_0^+. We note that for $T_0 = \mathsf{ACA}_0$ the theory T_0^+ is just ACA_0^+ and for $T_0 = \mathsf{RCA}_0$ the theory T_0^+ is just ACA_0.

Lemma 4.5.13. *For each $T_0 \supseteq \mathsf{RCA}_0$ given by a finite list of axioms*

$$T_0^+ \vdash \forall \alpha \left(\mathsf{WO}(\alpha) \to \mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}\big(\mathbf{R}^{\alpha}_{\Pi^1_1(\Sigma^0_2)}(T_0)\big)\right).$$

Proof. We reason in T_0^+. We assume $\mathsf{WO}(\alpha)$ and claim $\mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathbf{R}^{\alpha}_{\Pi^1_1(\Sigma^0_2)}(T_0))$.

Note that it suffices to show that $\mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathbf{R}^{\alpha}_{\Pi^1_1(\Sigma^0_2)}(T_0))$ is true in all the ω-models of T_0. Indeed, since $\mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathbf{R}^{\alpha}_{\Pi^1_1(\Sigma^0_2)}(T_0))$ is a $\Pi^1_1(\Sigma^0_2)$ sentence, if it fails, this fact is witnessed by some set X and hence $\mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathbf{R}^{\alpha}_{\Pi^1_1(\Sigma^0_2)}(T_0))$ fails in all the ω-models of T_0 containing X.

Now let us consider an ω-model $\mathfrak{M}$ of T_0 and show $\mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathbf{R}^{\alpha}_{\Pi^1_1(\Sigma^0_2)}(T_0))$. We note that, for some fixed k, all the facts of the form $\mathfrak{M} \models \mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathbf{R}^{\beta}_{\Pi^1_1(\Sigma^0_2)}(T_0))$ are Π^0_k. In order to finish the proof it suffices to show $\mathfrak{M} \models \mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathbf{R}^{\beta}_{\Pi^1_1(\Sigma^0_2)}(T_0))$, for all $\beta \preceq \alpha$ by transfinite induction on $\beta \preceq \alpha$. By the induction hypothesis we know that $\mathfrak{M}$ is a model of $\mathbf{R}^{\beta}_{\Pi^1_1(\Sigma^0_2)}(T_0)$. Since $\mathfrak{M}$ is an ω-model we need to show that for all the (standard) proofs p of a $\Pi^1_1(\Sigma^0_2)$-sentence φ in $\mathbf{R}^{\beta}_{\Pi^1_1(\Sigma^0_2)}(T_0)$ the sentence φ is true in $\mathfrak{M}$. We consider some proof p of this form and apply the cut-elimination theorem for predicate calculus to make sure that all the intermediate formulas in the proof are of the complexity $\Pi^1_n(\Pi^0_m)$ for some externally fixed n and m (depending only on the complexity of the axioms of T_0). We proceed by showing by induction on formulas in the proof that all of them are true in the model $\mathfrak{M}$; we can do this since the satisfaction relation for $\Pi^1_n(\Pi^0_m)$-formulas in $\mathfrak{M}$ is arithmetical. ❑

Lemma 4.5.14.

$$\mathsf{RCA}_0 \vdash \forall \alpha \left(\mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}\big(\mathbf{R}^{\alpha}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0)\big) \to \mathsf{WO}(\alpha)\right)$$

Proof. We prove the lemma by reflexive induction on α in RCA_0. We reason in RCA_0 and assume the reflexive induction hypothesis

$$\forall \beta \prec \alpha \; \mathsf{Pr}_{\mathsf{RCA}_0}\Big(\mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}\big(\mathbf{R}^{\beta}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0)\big) \to \mathsf{WO}(\beta)\Big).$$

We need to show that:

$$\mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}\big(\mathbf{R}^{\alpha}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0)\big) \to \mathsf{WO}(\alpha) \tag{4.11}$$

So assume the antecedent of (4.11). From the reflexive induction hypothesis we see that for each individual $\beta \prec \alpha$ the theory $\mathbf{R}^{\alpha}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0)$ proves $\mathsf{WO}(\beta)$. Since $\mathsf{WO}(\beta)$ is a $\Pi^1_1(\Sigma^0_2)$-formula, we infer from the antecedent of (4.11) that $\forall \beta \prec \alpha \; \mathsf{WO}(\beta)$. Thus $\mathsf{WO}(\alpha)$. ❑

Now we are ready to prove Theorem 4.5.11

Proof. From Theorem 4.4.17 we know that

$$\mathbf{R}^{\alpha}_{\Pi^1_1}(\mathsf{ACA}_0) \equiv_{\Pi^1_1(\Sigma^0_2)} \mathbf{R}^{\varepsilon_\alpha}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0).$$

From Lemma 4.5.14 we see that $\mathbf{R}^{\varepsilon_\alpha}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0)$ proves $\mathsf{WO}(\beta)$ for each $\beta \prec \varepsilon_\alpha$ and thus $|\mathbf{R}^{\alpha}_{\Pi^1_1}(\mathsf{ACA}_0)|_{\mathsf{WO}} \geqslant |\varepsilon_\alpha|$.

In order to prove $|\mathbf{R}^{\alpha}_{\Pi^1_1}(\mathsf{ACA}_0)| \leqslant |\varepsilon_\alpha|$ let us assume that for some β the theory $\mathbf{R}^{\alpha}_{\Pi^1_1}(\mathsf{ACA}_0)$ proves $\mathsf{WO}(\beta)$ and then show that $|\beta| < |\varepsilon_\alpha|$. Indeed, by Lemma 4.5.12 the theory $\mathbf{R}^{\alpha}_{\Pi^1_1}(\mathsf{ACA}_0)$ proves $\mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathbf{R}^{\beta}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0))$. Hence

$$|\beta| < |\mathbf{R}^{\alpha}_{\Pi^1_1}(\mathsf{ACA}_0)|_{\mathsf{RCA}_0} = |\mathbf{R}^{\varepsilon_\alpha}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0)|_{\mathsf{RCA}_0}.$$

And Proposition 4.5.6 gives us

$$|\varepsilon_\alpha| = |\mathbf{R}^{\varepsilon_\alpha}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0)|_{\mathsf{RCA}_0} > |\beta|.$$

This completes the proof. ❑

Extensions of ACA^+_0

It is usually attributed to Kreisel that for extensions $T \sqsupseteq \mathsf{ACA}_0$ the Π^1_1 proof-theoretic ordinal $|T|_{\mathsf{WO}} = |T + \varphi|_{\mathsf{WO}}$, for any true Σ^1_1-sentence φ (see [76, Theorem 6.7.4,6.7.5]). We note that our notion of reflection rank $|T|_{\mathsf{ACA}_0}$ does not enjoy the same property.

Remark 4.5.15. Let us consider an ordinal notation system α for some large recursive ordinal, for example the Bachmann-Howard ordinal. Now we modify α to define pathological ordinal notation α'. The order $\prec_{\alpha'}$ is the restriction of $\prec_\alpha$ to numbers m such that $\forall x \leqslant m \; \neg \mathsf{Prf}_{\mathsf{ACA}_0}(x, 0 = 1)$. And α' corresponds to the same element of the domain of $\prec_\alpha$ as α (note that since ACA_0 is consistent this element is in the domain of $\prec_{\alpha'}$ as well). We

see externally that α' is isomorphic to α, since ACA_0 is consistent. Let us denote by Iso the true Σ^1_1-sentence that expresses the fact that α and α' are isomorphic. Clearly,

$$\mathsf{ACA}_0 + \mathsf{WO}(\alpha') + \mathsf{Iso} \sqsupseteq \mathsf{ACA}_0 + \mathsf{WO}(\alpha),$$

$$|\mathsf{ACA}_0 + \mathsf{WO}(\alpha') + \mathsf{Iso}|_{\mathsf{ACA}_0} \geqslant |\mathsf{ACA}_0 + \mathsf{WO}(\alpha)|_{\mathsf{ACA}_0}$$

and under our choice of α the rank $|\mathsf{ACA}_0 + \mathsf{WO}(\alpha)|_{\mathsf{ACA}_0}$ will be equal to the Bachmann-Howard ordinal. At the same time, the theory $\mathsf{ACA}_0 + \neg\mathsf{Con}(\mathsf{ACA}_0)$ proves that α' is isomorphic to some finite order and hence

$$\mathsf{ACA}_0 + \neg\mathsf{Con}(\mathsf{ACA}_0) \vdash \mathsf{WO}(\alpha').$$

Hence

$$|\mathsf{ACA}_0 + \mathsf{WO}(\alpha')|_{\mathsf{ACA}_0} \leqslant |\mathsf{ACA}_0 + \neg\mathsf{Con}(\mathsf{ACA}_0)|_{\mathsf{ACA}_0} = 0,$$

the latter equality follows from Remark 4.5.5. And thus

$$|\mathsf{ACA}_0 + \mathsf{WO}(\alpha')|_{\mathsf{WO}} < |\mathsf{ACA}_0 + \mathsf{WO}(\alpha') + \mathsf{Iso}|_{\mathsf{WO}}.$$

Accordingly, Iso is a true Σ^1_1 sentence that alters the reflection rank of the theory $\mathsf{ACA}_0 + \mathsf{WO}(\alpha')$.

We address this problem with two different results. First in Theorem 4.5.16 we show that for any extension $T \sqsupseteq \mathsf{ACA}^+_0$, $|T|_{\mathsf{ACA}_0} = |T|_{\mathsf{WO}}$. Second we introduce the notion of robust reflection rank $|\cdot|^\star_{\mathsf{ACA}_0}$ that enjoys a number of nice properties and at the same time coincides with reflection rank $|\cdot|_{\mathsf{ACA}_0}$, for many natural theories T (in particular, for any any T such that $T \equiv_{\Pi^1_1} \mathbf{R}^\alpha_{\Pi^1_1}(\mathsf{ACA}_0)$, for some ordinal notation α).

Theorem 4.5.16. *Suppose $T \sqsupseteq \mathsf{ACA}^+_0$ then*

$$|T|_{\mathsf{WO}} = |T|_{\mathsf{ACA}_0}.$$

We prove the following general theorem

Theorem 4.5.17. *Suppose that a $\Pi^1_2(\Pi^0_2)$-sound theory $T_0 \sqsupseteq \mathsf{RCA}_0$ is given by a finite list of axioms. Then for each $U \sqsupseteq T^+_0$ we have*

$$|U|_{\mathsf{WO}} = |U|_{T_0}.$$

Proof. Combining Lemma 4.5.13 and Proposition 4.5.6 we see that $|U|_{\mathsf{WO}} \leqslant |U|_{T_0}$. In order to show that $|U|_{\mathsf{WO}} \geqslant |U|_{T_0}$ we prove that for each $\boldsymbol{\alpha} < \min(\boldsymbol{\omega}^{\boldsymbol{CK}}_{\mathbf{1}}, |U|_{T_0})$ we have $\boldsymbol{\alpha} < |U|_{\mathsf{WO}}$. This is suffiecient, since both $|U|_{\mathsf{WO}}$ and $|U|_{T_0}$ are either ∞ or below $\boldsymbol{\omega}^{\boldsymbol{CK}}_{\mathbf{1}}$. We consider some $\boldsymbol{\alpha} < \min(\boldsymbol{\omega}^{\boldsymbol{CK}}_{\mathbf{1}}, |U|_{T_0})$ and fix an ordinal notation system α with $|\alpha| = \boldsymbol{\alpha}$. From Lemma 4.5.9 we see that there is a true $\Sigma^1_1(\Pi^0_2)$-sentence φ such that

$$U + \varphi \vdash \mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathbf{R}^\alpha_{\Pi^1_1(\Sigma^0_2)}(T_0)).$$

Since $T_0 \supseteq \mathsf{RCA}_0$, we have

$$U + \varphi \vdash \mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathbf{R}^{\alpha}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0)).$$

And hence by Lemma 4.5.14 we have $U + \varphi \vdash \mathsf{WO}(\alpha)$. Thus

$$\boldsymbol{\alpha} = |\alpha| < |U + \varphi|_{\mathsf{wo}} = |U|_{\mathsf{wo}}.$$

This completes the proof of the theorem. ❑

Robust reflection rank

The *robust reflection rank* $|U|^{\star}_{T_0}$ of a theory $U \in \mathcal{E}$-T_0 over a theory $T_0 \supseteq \mathsf{RCA}_0$ is defined as follows:

$$|U|^{\star}_{T_0} = \sup\{|U + \varphi|_{T_0} \colon \varphi \text{ is a true } \Sigma^1_1(\Pi^0_2)\text{-sentence}\}.$$

Proposition 4.5.18. *For theories $T_0 \supseteq \mathsf{RCA}_0$ and $U \in \mathcal{E}$-T_0 the robust reflection rank $|U|^{\star}_{T_0}$ is an ordinal iff U is $\Pi^1_1(\Sigma^0_2)$-sound.*

Proof. If U is $\Pi^1_1(\Sigma^0_2)$-sound then for any true $\Sigma^1_1(\Pi^0_2)$-sentence φ the theory $U + \varphi$ is $\Pi^1_1(\Sigma^0_2)$-sound. Thus, by Corollary 4.5.4 each rank $|U + \varphi|_{T_0} \in \mathbf{On}$ and so $|U|^{\star}_{T_0} \in \mathbf{On}$.

If U is not $\Pi^1_1(\Sigma^0_2)$-sound then there is a false $\Pi^1_1(\Sigma^0_2)$ sentence φ that U proves. Let ψ be a true $\Sigma^1_1(\Pi^0_2)$-sentence that is RCA_0-provably equivalent to $\neg\varphi$. Clearly, $U + \psi$ is inconsistent, so $U + \psi \prec_{\Pi^1_1(\Sigma^0_2)} U + \psi$ and hence $\infty = |U + \psi|_{T_0} = |U|^{\star}_{T_0}$. ❑

Proposition 4.5.19. *Suppose $T_0 \supseteq \mathsf{RCA}_0$ is $\Pi^1_2(\Pi^0_3)$-sound, $U \in \mathcal{E}$-T_0, and for some ordinal notation α we have $U \equiv_{\Pi^1_1(\Sigma^0_2)} \mathbf{R}^{\alpha}_{\Pi^1_1(\Sigma^0_2)}(T_0)$. Then*

$$|U|^{\star}_{T_0} = |U|_{T_0} = |\alpha|.$$

Proof. We use Proposition 4.5.6 and see that

$$|U|^{\star}_{T_0} \geqslant |U|_{T_0} = |\alpha|.$$

Let us assume for a contradiction that $|U|^{\star}_{T_0} > |\alpha|$. In this case from Lemma 4.5.9 there is a true $\Sigma^1_1(\Pi^0_2)$ sentence φ such that

$$U + \varphi \vdash \mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathbf{R}^{\alpha}_{\Pi^1_1(\Sigma^0_2)}(U)).$$

Of course, this implies that

$$U \vdash \varphi \to \mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathbf{R}^{\alpha}_{\Pi^1_1(\Sigma^0_2)}(U)).$$

Note that $\varphi \to \mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathbf{R}^\alpha_{\Pi^1_1(\Sigma^0_2)}(U))$ is a $\Pi^1_1(\Sigma^0_2)$ sentence. Thus, from the assumption that $U \equiv_{\Pi^1_1(\Sigma^0_2)} \mathbf{R}^\alpha_{\Pi^1_1(\Sigma^0_2)}(T_0)$, it follows that:

$$\begin{aligned}
\mathbf{R}^\alpha_{\Pi^1_1(\Sigma^0_2)}(T_0) &\vdash \varphi \to \mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathbf{R}^\alpha_{\Pi^1_1(\Sigma^0_2)}(U)) \\
\mathbf{R}^\alpha_{\Pi^1_1(\Sigma^0_2)}(T_0) + \varphi &\vdash \mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathbf{R}^\alpha_{\Pi^1_1(\Sigma^0_2)}(U)) \\
&\vdash \mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathbf{R}^\alpha_{\Pi^1_1(\Sigma^0_2)}(U) + \varphi) \text{ by Lemma 4.5.8.} \\
&\vdash \mathsf{Con}(\mathbf{R}^\alpha_{\Pi^1_1(\Sigma^0_2)}(U) + \varphi))
\end{aligned}$$

Thus, $\mathbf{R}^\alpha_{\Pi^1_1(\Sigma^0_2)}(T_0) + \varphi$ is inconsistent by Gödel's Second Incompleteness Theorem. On the other hand, since T_0 is $\Pi^1_2(\Pi^0_3)$ sound, $\mathbf{R}^\alpha_{\Pi^1_1(\Sigma^0_2)}(T_0)$ is $\Pi^1_1(\Sigma^0_2)$ sound by Lemma 4.5.7. Thus, $\mathbf{R}^\alpha_{\Pi^1_1(\Sigma^0_2)}(T_0) + \varphi$ is consistent by Lemma 4.5.8. This is a contradiction. ❑

Finally we connect the notions of robust reflection rank $|\cdot|^\star_{\mathsf{ACA}_0}$ and proof-theoretic ordinal $|\cdot|_{\mathsf{WO}}$:

Theorem 4.5.20. *For any theory $T \in \mathcal{E}$-ACA_0 with robust reflection rank $|T|^\star_{\mathsf{ACA}_0} = \boldsymbol{\alpha}$ we have $|T|_{\mathsf{WO}} = \boldsymbol{\varepsilon}_{\boldsymbol{\alpha}}$ (here by definition we put $\boldsymbol{\varepsilon}_\infty = \infty$).*

Proof. First let us show that $|T|_{\mathsf{WO}} \geqslant \boldsymbol{\varepsilon}_{\boldsymbol{\alpha}}$. We break into cases based on whether $\boldsymbol{\alpha} = \infty$ or $\boldsymbol{\alpha} \in \mathbf{On}$

Assume $\boldsymbol{\alpha} = \infty$. Then by Proposition 4.5.18 there is false Π^1_1 sentence φ that is provable in T. Now we could construct an ordinal notation α such that $\mathsf{WO}(\alpha)$ is ACA_0-provably equivalent to φ: we put φ in the tree normal form [92, Lemma V.1.4] and take α to be the Kleene-Brouwer order on the tree. Clearly, $T \vdash \mathsf{WO}(\alpha)$ and $|\alpha| = \infty$. Thus $|T|_{\mathsf{WO}} = \infty = \boldsymbol{\varepsilon}_{\boldsymbol{\alpha}}$.

Now assume that $\boldsymbol{\alpha} \in \mathbf{On}$. Let us consider some $\boldsymbol{\beta} < \boldsymbol{\varepsilon}_{\boldsymbol{\alpha}}$ and show that $|T|_{\mathsf{WO}} > \boldsymbol{\beta}$. From the definition of robust reflection rank it is easy to see that we could find some true $\Sigma^1_1(\Pi^0_2)$ sentence φ such that $\boldsymbol{\beta} < \boldsymbol{\varepsilon}_{\,|T+\varphi|_{\mathsf{ACA}_0}}$. Since $|T + \varphi|_{\mathsf{ACA}_0}$ is the rank of a Σ^0_1 binary relation, $|T + \varphi|_{\mathsf{ACA}_0} < \boldsymbol{\omega}^{\boldsymbol{CK}}_{\mathbf{1}}$. Thus we could choose an ordinal notation γ such that $|\gamma| < |T + \varphi|_{\mathsf{ACA}_0}$ but $\boldsymbol{\beta} < \boldsymbol{\varepsilon}_{|\gamma|+1}$. From Lemma 4.5.9 we infer that there is a true $\Sigma^1_1(\Pi^0_2)$-sentence φ' such that $T + \varphi + \varphi' \vdash \mathsf{RFN}_{\Pi^1_1}(\mathbf{R}^\gamma_{\Pi^1_1}(\mathsf{ACA}_0))$. We find a $\beta \prec \varepsilon_{\gamma+1}$ such that $|\beta| = \boldsymbol{\beta}$. By the same reasoning as in the proof of Theorem 4.5.11 we infer that $\mathbf{R}^{\gamma+1}_{\Pi^1_1}(\mathsf{ACA}_0) \vdash \mathsf{WO}(\beta)$. Thus $T + \varphi + \varphi' \vdash \mathsf{WO}(\beta)$. Hence $|T + \varphi + \varphi'|_{\mathsf{WO}} > \boldsymbol{\beta}$. From Kreisel's Theorem about Π^1_1 proof-theoretic ordinals of extensions of ACA_0 we infer that $|T|_{\mathsf{WO}} = |T + \varphi + \varphi'|_{\mathsf{WO}} > \boldsymbol{\beta}$.

Now let us show that $|T|_{\mathsf{WO}} \leqslant \boldsymbol{\varepsilon}_{\boldsymbol{\alpha}}$. Assume, for the sake of contradiction, that $|T|_{\mathsf{WO}} > \boldsymbol{\varepsilon}_{\boldsymbol{\alpha}}$. Then there is an ordinal notation β with $|\beta| = \boldsymbol{\varepsilon}_{\boldsymbol{\alpha}}$ such that $T \vdash \mathsf{WO}(\beta)$. Let us fix some ordinal notation α such that $|\alpha| = \boldsymbol{\alpha}$. Clearly, there is an isomorphism between β and ε_α. Let us denote by Iso the natural $\Sigma^1_1(\Pi^0_2)$-sentence expressing the latter fact. We see that $T + \mathsf{Iso} \vdash \mathsf{WO}(\varepsilon_\alpha)$. Thus by Lemma 4.5.12 we see that

$$T + \mathsf{Iso} \vdash \mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathbf{R}^{\varepsilon_\alpha}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0)).$$

From Theorem 4.4.17 we conclude that

$$T + \mathsf{Iso} \vdash \mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathbf{R}^{\alpha}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{ACA}_0)).$$

Since over ACA_0 every Π^1_1-formula is equivalent to a $\Pi^1_1(\Sigma^0_2)$-formula,

$$T + \mathsf{Iso} \vdash \mathsf{RFN}_{\Pi^1_1}(\mathbf{R}^{\alpha}_{\Pi^1_1}(\mathsf{ACA}_0)).$$

Therefore

$$|T|^{\star}_{\mathsf{ACA}_0} \geqslant |T + \mathsf{Iso}|_{\mathsf{ACA}_0} > |\mathbf{R}^{\alpha}_{\Pi^1_1}(\mathsf{ACA}_0)|_{\mathsf{ACA}_0} = |\alpha| = \boldsymbol{\alpha},$$

but $|T|^{\star}_{\mathsf{ACA}_0} = \boldsymbol{\alpha}$, a contradiction. ❑

4.6 Ordinal notation systems based on reflection principles

In this section we turn to ordinal notation systems based on reflection principles, like the one Beklemishev introduced in [8]. We will formally describe such a notation system momentarily, but, roughly, the elements of such notation systems are theories axiomatized by reflection principles and the ordering on them is given by consistency strength. Beklemishev endorsed the use of such notation systems as an approach to the well-known *canonicity problem* of ordinal notation systems. Since then, such notation systems have been intensively studied; see [29] for a survey of these notation systems and their properties.

We will consider ordinal notation systems based on the calculus RC^0 due to Beklemishev [11]. In earlier works, e.g. [8] on modal logic based ordinal analysis, ordinal notation systems arose from fragments of the polymodal provability logic GLP. However, this application of polymodal provability logic didn't required the full expressive power of GLP. Thus, starting from a work of Dashkov [20], strictly positive modal logics have been isolated that yield the same ordinal notation system as the logic GLP, but are much simpler from a technical point of view.

The set of formulas of RC^0 is given by the following inductive definition:

$$F ::= \top \mid F \wedge F \mid \Diamond_n F, \text{ where } n \text{ ranges over } \mathbb{N}.$$

An RC^0 sequent is an expression $A \vdash B$, where A and B are RC^0-formulas. The axioms and rules of inference of RC^0 are:

1. $A \vdash A$; $A \vdash \top$; if $A \vdash B$ and $B \vdash C$ then $A \vdash C$;
2. $A \wedge B \vdash A$; $A \wedge B \vdash B$; if $A \vdash B$ and $A \vdash C$ then $A \vdash B \wedge C$;
3. if $A \vdash B$ then $\Diamond_n A \vdash \Diamond_n B$, for all $n \in \mathbb{N}$;
4. $\Diamond_n \Diamond_n A \vdash \Diamond_n A$, for every $n \in \mathbb{N}$;

5. $\Diamond_n A \vdash \Diamond_m A$, for all $n > m$;

6. $\Diamond_n A \wedge \Diamond_m B \vdash \Diamond_n(A \wedge \Diamond_m B)$, for all $n > m$.

Let us describe the intended interpretation of RC^0-formulas in $\mathcal{L}_1$-sentences. The interpretation $\top^*$ of $\top$ is $0 = 0$. The interpretation $(A \wedge B)^*$ is $A^* \wedge B^*$. The interpretation $(\Diamond_n A)^*$ is $\mathsf{RFN}_{\Sigma_n}(A^*)$. A routine check by induction on the length of RC^0-derivations shows that if $A \vdash B$ then $\mathsf{EA} + A^* \vdash B^*$, for any RC^0-formulas A and B.

For a more extensive coverage of positive provability logic see [12].

We denote by $\mathcal{W}$ the set of all RC^0 formulas. The binary relation $<_n$, and the natural equivalence relation $\sim$ are given by

$$A <_n B \overset{\text{def}}{\iff} B \vdash \Diamond_n A, \qquad A \sim B \overset{\text{def}}{\iff} B \vdash A \text{ and } A \vdash B.$$

The Beklemishev ordinal notation system for ε_0 is the structure $(\mathcal{W}/\!\sim, <_0)$.

The following result is due to Beklemishev (see [10, 11]):

Theorem 4.6.1. $(\mathcal{W}/\!\sim, <_0)$ *is a well-ordering with the order type* ε_0.

The transitivity of $(\mathcal{W}/\!\sim, <_0)$ is trivial. The linearity of $(\mathcal{W}/\!\sim, <_0)$ is provable by a purely syntactical argument within the system RC^0. But Beklemishev's proof of the well-foundedness of $(\mathcal{W}/\!\sim, <_0)$ was based on the construction of an isomorphism with Cantor's ordinal notation system for ε_0, i.e., Cantor normal forms.

Here we will give a proof of the well-foundedness part of Theorem 4.6.1 by providing an alternative interpretation of the $\Diamond_n$'s by reflection principles in *second*-order arithmetic and then applying the results of §4.3 to derive well-foundedness.

Theorem 4.6.2. $(\mathcal{W}, <_0)$ *is a well-founded relation.*

Proof. We prove that the set $\mathcal{W}$ of RC^0-formulas is well-founded with respect to $<_0$.

We give an alternative interpretation of RC^0. According to this interpretation, the image $\top^*$ of $\top$ is $0 = 0$, $(A \wedge B)^*$ is $A^* \wedge B^*$, and $(\Diamond_n A)^*$ is $\mathsf{RFN}^1_{\Pi^1_{n+1}}(\mathsf{ACA}_0 + A^*)$.

We note that if $A \vdash B$ is a derivable RC^0-sequent then $\mathsf{ACA}_0 + A^* \vdash B^*$. This can be checked by a straightforward induction on RC^0-derivations. Also from the definition it is clear that for any A the theory $\mathsf{ACA}_0 + A^*$ is Π^1_1-sound (and in fact true A^* is true).

Now assume for a contradiction that there is an infinite descending chain $A_0 >_0 A_1 >_0 \dots$ of RC^0-formulas. Then $A_0^*, A_1^*, \dots$ is an infinite sequence of sentences such that $\mathsf{ACA}_0 + A_i^* \vdash \mathsf{RFN}^1_{\Pi^1_1}(\mathsf{ACA}_0 + A^*_{i+1})$. Henceforth we have a $\prec_{\Pi^1_1}$-descending chain of Π^1_1-sound extensions of ACA_0, contradicting Theorem 4.3.2. ❑

The key fact that we have used in this proof is that all the theories A_i^* are Π^1_1-sound. In fact all the theories under consideration are subtheories of ACA and hence the proof is naturally formalizable in $\mathsf{ACA}_0 + \mathsf{RFN}^1_{\Pi^1_1}(\mathsf{ACA})$.[4]

[4]The fact that $\mathsf{ACA} \equiv_{\Pi^1_\infty} \mathsf{RFN}^1_{\Pi^1_\infty}(\mathsf{ACA}_0)$ could be proved by a standard technique going back to Kreisel and Lévy [54]. A study of the exact correspondence between restrictions of the schemes of reflection and induction in the setting of second order arithmetic has been recently performed by Frittaion [34].

Now we show that the same kind of argument could be carried in ACA_0 itself.

Theorem 4.6.3. *For each $A \in \mathcal{W}$, the theory ACA_0 proves that $(\{B \in \mathcal{W} \mid B <_0 A\}, <_0)$ is well-founded.*

Proof. Note that in RC^0 any formula A follows from formulas $\Diamond_n \top$ such that, for all $\Diamond_m$ that occur in A, $m < n$; this fact could be proved by a straightforward induction on length of A. Clearly, for any such n, the set $\{B \in \mathcal{W} \mid B <_0 A\}$ is a subset of $\{B \in \mathcal{W} \mid B <_0 \Diamond_n \top\}$. Thus, without loss of generality, we may consider only the case of A being of the form $\Diamond_n \top$.

Now we reason in ACA_0. We assume for a contradiction that there is an infinite descending chain $\Diamond_n \top >_0 A_0 >_0 A_1 >_0 \dots$ of RC^0-formulas.

We construct a countably-coded ω-model $\mathfrak{M}$ of RCA_0 that contains this chain. Note that using arithmetical comprehension we could construct a (set encoding) partial satisfaction relation for $\mathfrak{M}$ that the sentence RCA_0 (conjunction of all axioms from some natural finite axiomatization of RCA_0) and all $\Pi^1_{n+1}(\Pi^0_3)$ formulas. We want to show that if RCA_0 proves some $\Pi^1_{n+1}(\Pi^0_3)$ sentence φ then φ is true in $\mathfrak{M}$. For this we consider any cut-free proof p of the sequent $\neg\mathsf{RCA}_0, \varphi$. And next by induction on subproofs of p show that all sequents in p are valid in $\mathfrak{M}$ (according to the partial satisfaction relation that we constructed above). Hence the principle $\mathsf{RFN}_{\Pi^1_{n+1}(\Pi^0_3)}(\mathsf{RCA}_0)$ holds in $\mathfrak{M}$.

We again define an alternative interpretation of RC^0. The interpretation $\top^*$ is $0 = 0$, the interpretations $(A \wedge B)^*$ are $A^* \wedge B^*$, and the interpretations $(\Diamond_k A_i)^*$ are $\mathsf{RFN}_{\Pi^1_{k+1}(\Pi^0_3)}(\mathsf{RCA}_0 + A_i^*)$. From the previous paragraph we see that $\mathfrak{M} \models (\Diamond_n \top)^*$. And since $\Diamond_n \top >_0 A_0$, we have $\mathfrak{M} \models (\Diamond_0 A_0)^*$, i.e., $\mathfrak{M} \models \mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0 + A_0^*)$. Thus in $\mathfrak{M}$ there is an infinite sequence of theories $\mathsf{RCA}_0 + A_0^*, \mathsf{RCA}_0 + A_1^*, \dots$ such that $\mathsf{RCA}_0 + A_i^* \vdash \mathsf{RFN}_{\Pi^1_1(\Pi^0_3)}(\mathsf{RCA}_0 + A_{i+1}^*)$ and $\mathsf{RFN}_{\Pi^1_1(\Sigma^0_2)}(\mathsf{RCA}_0 + A_0^*)$. Since $\mathfrak{M}$ is a model of RCA_0, by Theorem 4.3.3 we reach a contradiction. ❑

Chapter 5

A reduction principle for ω-model reflection

This chapter contains joint work with Fedor Pakhomov.

5.1 Introduction

Two sorts of principles are commonly called *reflection principles* in mathematical logic. First, according to *semantic* reflection principles, whatever is true in the universe holds in some set-sized model. The thought here is that structures within the universe reflect what is happening in the universe. Second, according to *syntactic* reflection principles, whatever is provable is true. The thought here is that we should endorse these principles by reflecting on the soundness of our axioms and inference rules.

In this chapter we establish a conservation theorem relating both types of reflection principles in second-order arithmetic. In particular, we establish a conservation theorem that reduces an ω model reflection principle to iterated syntactic reflection principles. There is a thorough proof-theoretic understanding of the latter in terms, e.g., of ordinal analysis. Accordingly, these reductions yield proof-theoretic analyses of ω model reflection principles.

We will be concerned in this paper with iterations of reflection along arbitrary (potentially non-recursive) well-orderings. We formally define these iterations using the language $\mathbf{L}_2$ that extends the standard language of second-order arithmetic with set-constants C_X for all sets X. The definitions of $\mathbf{L}_2$ formulas and the standard syntactic complexity classes for $\mathbf{L}_2$ can be given in ACA_0. Accordingly, throughout this paper we formalize our results in ACA_0. Moreover, we restrict our attention to axiomatic theories that extend ACA_0.

The main syntactic reflection principle we consider, $\mathbf{\Pi}^1_1$-$\mathsf{RFN}(T)$, informally says "all $\mathbf{\Pi}^1_1$ theorems of T are true." We will also be interested in the theories $\mathbf{\Pi}^1_1$-$\mathbf{R}^\alpha(T)$ that result from iterating this principle along well-orderings α. We will give precise definitions of these theories via Gödel's fixed point lemma in §5.2. Informally, one can think of them as defined

inductively, according to the following equation:

$$\mathbf{\Pi}^1_1\text{-}\mathbf{R}^\alpha(T) := T + \{\mathbf{\Pi}^1_1\text{-}\mathsf{RFN}\big(T + \mathbf{\Pi}^1_1\text{-}\mathbf{R}^\beta(T)\big) : \beta < \alpha\}$$

We will also consider ω model reflection principles. An ω model is an L_2 structure whose first-order part is $\mathbb{N}$ and whose second-order part is some subset of $\mathcal{P}(\mathbb{N})$. The semantic reflection principle we work with is "every set is contained in an ω model of T."

Our main theorem is the following:

Theorem 5.1.1 (ACA_0)**.** *For any $\mathbf{\Pi}^1_2$ axiomatized theory T, the following are equivalent:*

1. *Every set is contained in an ω model of T.*
2. $\forall\alpha\big(\mathsf{WO}(\alpha) \to \mathbf{\Pi}^1_1\text{-}\mathsf{RFN}(\mathbf{\Pi}^1_1\text{-}\mathbf{R}^\alpha(T))\big)$.

Theorem 5.1.1 provides a reduction of ω model reflection to iterated syntactic reflection. This reduction is desirable because of the distinct roles the two types of principles play in second-order arithmetic. On the one hand, ω model reflection principles are well-known in reverse mathematics, since many theories of interest can be axiomatized in terms of ω model reflection principles. On the other hand, iterated syntactic reflection principles are widely studied in ordinal analysis because of the systematic connections between iterated reflection and proof-theoretic ordinals. Thus, Theorem 5.1.1 opens the path to a systematic connection between ω model reflection and ordinal analysis.

Here is our plan for the rest of the chapter. In §5.2 we cover a number of preliminaries. We present the class language $\mathbf{L}_2$ for second-order arithmetic. We then define its syntactic complexity classes, their attendant reflection principles, and the iterations thereof. In §5.3 we define an infinitary proof system which is sound and complete with respect to ω models. The use of such a proof system is crucial for our main result. In §5.4 we prove the main theorems of our chapter. In particular, we prove Theorem 5.1.1, a reduction of ω model reflection to iterated syntactic reflection.

5.2 Preliminaries

Our base system is the system ACA_0. Since this theory is finitely axiomatizable we identify it with a sentence giving its finite axiomatization. Throughout this chapter we restrict our attention to theories extending ACA_0. So whenever we make a claim about "every theory T," we mean "every theory extending ACA_0."

Languages and Complexity Classes

In this chapter we will study reflection principles for formulas with set parameters. In the study of reflection and provability in first-order arithmetic it is common to study provability for formulas with number parameters; note the parameter in the expression

$$\forall x\big(\mathsf{Prv}_{\mathsf{PA}}(\ulcorner\varphi(\dot{x})\urcorner) \to \varphi(x)\big).$$

Here $\ulcorner\varphi(\dot{x})\urcorner$ denotes the Gödel number of the formula, $\varphi(\underline{x})$, where $\underline{x}$ is the numeral

$$\underbrace{S(\dots S}_{x \text{ times}}(0)\dots)).$$

This strategy is not available if we want to formalize claims about the provability of formulas with set parameters. Since there are no numerals for sets of natural numbers, we need to use a different approach to pass second-order variables inside provability predicates/reflection principles.

We write L_2 to denote the standard language of second-order arithmetic. We write $\mathbf{L}_2$ to denote the extension of L_2 with set-constants C_X for all sets X. From the external perspective $\mathbf{L}_2$ is a continuum-sized language. However, formulas of $\mathbf{L}_2$ can be encoded by sets and reasoned about within ACA_0. We will use the rest of this subsection to explain how this is accomplished.

L_2 formulas are finitary objects and are encoded by natural numbers. The code for an $\mathbf{L}_2$ formula $\varphi(C_{Y_1}, ..., C_{Y_n}, \vec{x})$ is a pair $\big(\varphi(X_1, ..., X_n, \vec{x}), \langle Y_1, ..., Y_n\rangle\big)$ where $\varphi(X_1, ..., X_n, \vec{x})$ is (a code for) an L_2-formula and $\langle Y_1, ..., Y_n\rangle$ is a sequence of sets. Note that whereas L_2 formulas are encoded by numbers, $\mathbf{L}_2$ formulas are thus encoded by sets.

Standard manipulations of (codes of) L_2 formulas (e.g., forming conjunctions, performing substitutions, etc.) is totally finitary and thus can be carried out in ACA_0 (indeed, in much weaker theories). Analogous manipulations of (codes of) $\mathbf{L}_2$ formulas is carried out on sets rather than on numbers. Nevertheless, ACA_0 can carry out these sorts of manipulations. The code of a formula formed, e.g., by conjunction is *arithmetic* in the codes of the conjuncts.

For any formula $\varphi(X_1, ..., X_m, \vec{x})$ with m set variables, there is a function cmp_φ which maps $X_1, ..., X_m$ to $\{\vec{x} : \varphi(X_1, ..., X_m, \vec{x})\}$. For any arithmetic φ, the graph of cmp_φ is defined by an arithmetic formula. The functions that manipulate (codes of) $\mathbf{L}_2$ formulas are expressible in terms of these definable comprehension functions. For instance, the code of a conjunction is the output of the comprehension function corresponding to an arithmetic operation applied to the codes of the conjuncts. Given a formula

$$\varphi(X_1, \dots, X_n, y_1, \dots, y_m) \in L_2$$

the expression

$$\ulcorner\varphi(\dot{X}_1, \dots, \dot{X}_n, \dot{y}_1, \dots, \dot{y}_m)\urcorner$$

is the term (built using the definable comprehension functions) denoting the code of the formula

$$\varphi(C_{X_1}, \dots, C_{X_n}, \underline{y_1}, \dots, \underline{y}_m).$$

As usual we write $\Pi^1_0 = \Sigma^1_0$ ($\mathbf{\Pi}^1_0 = \mathbf{\Sigma}^1_0$) to denote the class of L_2-formulas ($\mathbf{L}_2$-formulas) without second-order quantifiers. The class $\Pi^1_{n+1} \subseteq L_2$ ($\mathbf{\Pi}^1_{n+1} \subseteq \mathbf{L}_2$) consists of all formulas of the form $\forall\overrightarrow{Xx}\ \varphi$, where $\varphi \in \Sigma^1_n$ ($\varphi \in \mathbf{\Sigma}^1_n$) and $\overrightarrow{Xx}$ is a vector of variables that could contain both first and second order variables. The class $\Sigma^1_{n+1} \subseteq L_2$ ($\mathbf{\Sigma}^1_{n+1} \subseteq \mathbf{L}_2$) consists of

all formulas of the form $\exists \overrightarrow{Xx}\, \varphi$, where $\varphi \in \Pi^1_n$ ($\varphi \in \mathbf{\Pi}^1_n$) and $\overrightarrow{Xx}$ is some vector of variables that could contain both first and second-order variables.

In second-order arithmetic it is useful to work with countable sets of sets of naturals. To do this we represent a countable set $\mathcal{S} \subseteq \mathcal{P}(\mathbb{N})$ by a code of a countable sequence $\langle S_i \subseteq \mathbb{N} \mid i \in A\rangle$, $A \subseteq \mathbb{N}$ such that $\mathcal{S} = \{S_i \mid i \in A\}$. Formally, we use the predicate $X \dot{\in} Y$ that says:

$$\exists z \, \big(\langle z, 0\rangle \in Y \wedge \forall x (x \in X \leftrightarrow \langle z, x+1\rangle \in Y)\big).$$

Inside ACA_0 we work with countable $\mathbf{L}_2$-theories represented by sets T treated as codes for their set of axioms. The provability predicate $\mathsf{Prv}(T, \varphi)$ expresses that T is an $\mathbf{L}_2$-theory, φ is an $\mathbf{L}_2$-formula, and there is a proof P of φ in first-order logic such that all non-logical axioms in P are from T. Note that here the proof P by necessity is encoded by a set. However, Prv is equivalent to a Π^1_0 formula.

Reflection Principles

A standard construction allows us to define in ACA_0 partial truth definitions $\mathsf{Tr}_{\mathbf{\Pi}^1_n}(X)$, for the classes of formulas $\mathbf{\Pi}^1_n$. Here for any Π^1_n-formula $\varphi(\vec{X}, \vec{y})$ we have that

$$\mathsf{ACA}_0 \vdash \forall \vec{X}, \vec{y} \Big(\varphi(\vec{X}, \vec{y}) \leftrightarrow \mathsf{Tr}_{\mathbf{\Pi}^1_n}\big(\varphi(\vec{X}, \vec{y})\big)\Big)$$

Note that the formulas as $\mathsf{Tr}_{\mathbf{\Pi}^1_n}(X)$ are Π^1_n-formulas. We have truth definitions $\mathsf{Tr}_{\mathbf{\Sigma}^1_n}(X)$, for $n \geqslant 1$ with analogous properties as well.

For a theory T we put $\mathbf{\Pi}^1_1$-$\mathsf{RFN}(T)$, $n \geqslant 1$, to be the $\mathbf{L}_2$-sentence

$$\forall \varphi \in \mathbf{\Pi}^1_1 \, (\mathsf{Prv}(T, \varphi) \rightarrow \mathsf{Tr}_{\mathbf{\Pi}^1_1}(\varphi)).$$

Recall that an ω-model $\mathfrak{M}$ of second-order arithmetic is a structure whose interpretation of the natural numbers is standard, and the sort of sets of naturals is interpreted by some subset $\mathcal{S}_{\mathfrak{M}}$ of $\mathcal{P}(\mathbb{N})$. We reserve Fraktur letters $\mathfrak{M}$, $\mathfrak{N}$, ... for ω models. If $\mathcal{S}_{\mathfrak{M}}$ is countable, then the ω-model $\mathfrak{M}$ is called countable. Formally, an ω-model $\mathfrak{M}$ is a code for a countable family $\mathcal{S}_{\mathfrak{M}}$ of sets, and a satisfaction relation $\models_{\mathfrak{M}}$ that is defined on $L_{\mathfrak{M}}$, where $L_{\mathfrak{M}}$ is the set of all $\mathbf{L}_2$-sentences that contain constants C_X only for $X \in \mathcal{S}_{\mathfrak{M}}$. Full satisfaction predicates are available only in ACA^+_0, but partial satisfaction predicates are available in ACA_0. The partial satisfaction relation $\models_{\mathfrak{M}}$ should satisfy the usual Tarski clauses. When we work in ACA_0 the expression $\mathfrak{M} \models \varphi$ for $\varphi \in \mathbf{L}_2$ means that $\varphi \in L_{\mathfrak{M}}$ and $\models_{\mathfrak{M}} \varphi$. And the expression $\mathfrak{M} \models T$ for an $\mathbf{L}_2$-theory T means that $T \subseteq L_{\mathfrak{M}}$ and for any axiom φ of T we have $\models_{\mathfrak{M}} \varphi$. The expression $X \in \mathfrak{M}$ is a shorthand for $X \dot{\in} \mathcal{S}_{\mathfrak{M}}$. We will also be interested in the reflection principle "every set is contained in an ω model of T." This is equivalent to the claim that every true $\mathbf{\Sigma}^1_1$ is satisfied by some ω model of T, a fact which we will use in the proof of Theorem 5.1.1.

Linear Orders and Iterated Reflection

We consider linear orders α defined on subsets of natural numbers. We encode such orders as pairs $\langle D_\alpha, \prec_\alpha \rangle$, where $D_\alpha \subseteq \mathbb{N}$ and $\prec_\alpha \subseteq D_\alpha^2$ is a transitive irreflexive binary relation. For $x \in D_\alpha$ we denote as $\mathsf{cone}(\alpha, x)$ the set $\{y \in D_\alpha \mid y \prec_\alpha x\}$. Clearly, cone could be expressed by a Π^1_0 comprehension term.

We will now turn to defining iterations of the reflection principle $\mathbf{\Pi}^1_1\text{-}\mathsf{RFN}(T)$ along linear orders α. In order to define its iterates $\mathbf{\Pi}^1_1\text{-}\mathbf{R}^{\cdot}(\cdot)$ we will define formulas $\mathbf{\Pi}^1_1\text{-}\mathsf{RFN}^{\mathsf{it}}(T, \alpha)$, where T and α are free set variables and x is a free number variable. We define $\mathbf{\Pi}^1_1\text{-}\mathsf{RFN}^{\mathsf{it}}(T, \alpha)$ as a fixed point that satisfies:

$$\mathsf{ACA}_0 \vdash \mathbf{\Pi}^1_1\text{-}\mathsf{RFN}^{\mathsf{it}}(T, \alpha) \leftrightarrow \mathbf{\Pi}^1_1\text{-}\mathsf{RFN}\big(T + \{\mathbf{\Pi}^1_1\text{-}\mathsf{RFN}^{\mathsf{it}}(\dot{T}, \mathsf{cone}(\dot{\alpha}, \dot{x})) \mid x \in D_\alpha\}\big).$$

We put

$$\mathbf{R}^\alpha(T) = T + \{\mathbf{\Pi}^1_1\text{-}\mathsf{RFN}^{\mathsf{it}}(T, \mathsf{cone}(\alpha, \dot{x})) \mid x \in D_\alpha\}.$$

And thus

$$\mathsf{ACA}_0 \vdash \mathbf{\Pi}^1_1\text{-}\mathsf{RFN}^{\mathsf{it}}(T, \alpha) \leftrightarrow \mathbf{\Pi}^1_1\text{-}\mathsf{RFN}\big(\mathbf{R}^\alpha(T)\big).$$

Clearly $\mathbf{\Pi}^1_1\text{-}\mathbf{R}^\alpha(T)$ is a Π^1_0-comprehension term depending on α and T.

5.3 ω-proofs

In this section we will show that with our choice of ACA_0 as base system, the principle of ω-model reflection is fairly robust with respect to the choice of particular formalization. Namely, we will show the equivalence of the variants of reflection based on ω-models and cut-free ω-proofs. Note that David Fernández-Duque [30] proved that for certain other similar reflection principles these equivalences aren't provable in ACA_0; namely he considered reflection principles based on certain formalizations of provability in ω-logic that were not based on the notion of ω-proof.

Defining ω proofs for $\mathbf{L}_2$

First let us formulate the variant of ω-logic for $\mathbf{L}_2$. This logic will be a variant of the Tait calculus. Formulas are built up from literals using the connectives $\wedge$, $\vee$ and quantifiers $\forall x$, $\exists x$, $\forall X$, $\exists X$. Literals are atomic $\mathbf{L}_2$-formulas φ and their negations $\sim\varphi$. As usual for any formula φ, its negation $\neg\varphi$ is defined inductively by the De Morgan rules. That is, $\neg\varphi$ is the result of switching any connective and quantifier with the dual, switching positive literals φ with $\sim\varphi$ and switching negative literals $\sim\varphi$ with φ. Sequents are countable sets of formulas without free natural number variables (we allow free set variables). The axioms and rules of the logic are:

$$\frac{}{\Gamma, t = v}\text{, if } \mathsf{val}(t) = \mathsf{val}(v)\ (\mathsf{Ax}_1); \quad \frac{}{\Gamma, \sim t = v;}\text{, if } \mathsf{val}(t) \neq \mathsf{val}(v)\ (\mathsf{Ax}_2);$$

$$\frac{}{\Gamma, t \in C_A}\text{, if } \mathsf{val}(t) \in A\ (\mathsf{Ax}_3); \quad \frac{}{\Gamma, \sim t \in C_A}\text{, if } \mathsf{val}(t) \notin A\ (\mathsf{Ax}_4);$$

$$\frac{}{\Gamma, t \in X, \sim t \in X}\ (\mathsf{Ax}_5); \quad \frac{\Gamma, \varphi \qquad \Gamma, \psi}{\Gamma, \varphi \wedge \psi}\ (\wedge\text{-}\mathsf{Int}); \quad \frac{\Gamma, \varphi, \psi}{\Gamma, \varphi \vee \psi}\ (\vee\text{-}\mathsf{Int});$$

$$\frac{\Gamma, \varphi(\underline{n}), \qquad \text{for all } n \in \mathbb{N}}{\Gamma, \forall x\ \varphi(x)}\ (\forall_1\text{-}\mathsf{Int}); \quad \frac{\Gamma, \varphi(t)}{\Gamma, \exists x\ \varphi(x)}\ (\exists_1\text{-}\mathsf{Int});$$

$$\frac{\Gamma, \varphi(Y)}{\Gamma, \forall X\ \varphi(X)}\text{, if } Y \notin \mathsf{FV}(\Gamma)(\forall_2\text{-}\mathsf{Int}); \quad \frac{\Gamma, \varphi(C_A)}{\Gamma, \exists X\ \varphi(X)}\ (\exists_2\text{-}\mathsf{Int}_1); \quad \frac{\Gamma, \varphi(Y)}{\Gamma, \exists X\ \varphi(X)}\ (\exists_2\text{-}\mathsf{Int}_2);$$

$$\frac{\Gamma, \varphi \qquad \Gamma, \neg\varphi}{\Gamma}\ (\mathsf{Cut}). \quad \frac{\Gamma}{\Gamma}\ (\mathsf{Rep});$$

A *pre-proof* is a any true that accords with these axioms and rules in the sense that its leaves are axioms and each child node follows from applying one of the rules. Note that a pre-proof may be ill-founded. By a *proof* we mean a well-founded pre-proof. A sequent Γ is ω-provable if there is a well-founded proof-tree with Γ as its conclusion. We write $\vdash_\omega \Gamma$ if the sequent Γ has an ω-proof. And we write $\vdash_0 \Gamma$ if the sequent Γ has a cut-free ω-proof.

Details of encoding ω-proofs

We now describe in some detail how we encode infinitary proof trees in ACA_0. We encode sequents as codes for countable sets of $\mathbf{L}_2$-formulas. Due to the way our encoding works, the same sequent could have multiple representations. Note that equality on codes of sequents coincides with extensional equality:

$$X \dot{=} Y \iff^{\text{def}} \forall Z(Z \dot{\in} X \leftrightarrow Z \dot{\in} Y).$$

And it is easy to see that $X \dot{=} Y$ is equivalent to a Π^1_0 formula.

It is useful to define not only the notion of *proof* but also the notion of *pre-proof*, where a pre-proof is a possibly ill-founded derivation tree. More formally, a pre-proof P is (a code for) a triple Sh_P, Sq_P, Rl_P. Here Sh_P is a "proof-shape" tree $\langle I_P, r_P, \prec_P \rangle$, where $I_P \subseteq \mathbb{N}$ is the domain of the tree, $r_P \in I_P$ is the root of the tree, and $x \prec_P y$ is the binary relation on I_P with the intended meaning that x is a child of y. We require that for any $i \in I_P$ there exists unique $\prec_P$-path from it to the root

$$i = i_0 \prec_P i_1 \prec_P \ldots \prec_P i_n = r_P.$$

We require Sq_P to be an assignment of sequents $\langle \Delta_i \mid i \in I_P \rangle$ to the nodes of the tree Sh_P. Finally, Rl_P is an assignment of rules $\langle R_i \mid i \in I_P \rangle$ to the nodes of the tree Sh_P. Each R_i contains all the information about the applied rule. First it contains the rule type (Ax_1, Ax_2, Ax_3, Ax_4, Ax_5, $\wedge$-Int, $\vee$-Int, $\forall_1$-Int, $\exists_1$-Int, $\forall_2$-Int, $\exists_2$-Int_1, $\exists_2$-Int_2). And it contains the information specific to each particular rule type. Let us specify what this information is in the case when R_i is of the type $\forall_1$-Int, the cases of all the other rule types are analogous.

The rule R_i should be supplied with the sequent Γ_i, variable x_i, formula $\varphi_i(x_i)$ and sequents of indices of the premises $\langle p_{i,n} \mid n \in \mathbb{N}\rangle$. It is required that $\Delta_i \dot{=} (\Gamma_i, \forall x_i\ \varphi_i(x_i))$, that all $\Delta_{p_{i,n}} \dot{=} (\Gamma_i, \varphi_i(\underline{n}))$, that $\{p_{i,n} \mid n \in \mathbb{N}\} = \{j \in I_P \mid j \prec_P i\}$, and that $p_{i,n}$ are pairwise distinct. For a pre-proof P the sequent Γ_{r_P} is called the conclusion of P. A pre-proof P is called a proof if $\prec_P$ is a well-founded relation.

We write $\vdash_\omega \Gamma$ if the sequent Γ has an ω-proof. And we write $\vdash_0 \Gamma$ if the sequent Γ has a cut-free ωproof. We note that ACA_0 cannot the prove full cut-elimination theorem for ω-logic (cut-elemination for ω-logic requires the system ACA_0^+; however, ACA_0 *can* show that it is possible to eliminate all the cuts of the highest rank, see [38, Theorem 6.4.1]). Due to this issue we formulate several variants of ω-completeness theorems.

Recall that we write $\mathfrak{M} \models T$ if all axioms of theory T hold in the model $\mathfrak{M}$. At the same time for closed sequents Γ (i.e. sequents without free variables) we will write $\mathfrak{M} \models \Gamma$ if some formula $\varphi \in \Gamma$ holds in $\mathfrak{M}$. This is an abuse of notation since both sequents and theories are represented by codes of (countable) sets of $\mathbf{L}_2$-formulas. However, it will be always clear from context whether a particular object is a theory or a sequent (in particular we denote theories by capital Latin letters T, U and sequents by capital Greek letters Γ, Δ). For a theory T we denote by $\neg T$ the sequent $\{\neg\varphi \mid \varphi \text{ is an axiom of } T\}$.

Completeness theorems for cut-free ω proofs

We now describe in detail a completeness theorem for ω proofs with respect to ω models. Our completeness theorem is proved using Schütte's method of deduction chains. Thus, before proving the theorem we will work up to the definition of a *deduction chain* for a sequent Γ and a countable fragment H of $\mathbf{L}_2$. First, the definition of a *countable fragment* of $\mathbf{L}_2$:

Definition 5.3.1. A *countable fragment* H of $\mathbf{L}_2$ is a countable set of $\mathbf{L}_2$-formulas such that for any L_2 formula $\varphi(X_1, ..., X_n)$ and set constants $C_1, ..., C_n$ occurring in H the formula $\varphi(C_1, ..., C_n)$ is in H.

When we are working with a sequent Γ and a countable fragment H, we will assume that:

1. H comes with a fixed enumeration $Y_0, Y_1, ...$ of the free set variables in H that do not occur free in Γ.
2. H comes with a fixed enumeration $A_0, A_1, ...$ where each A_i is either:
   a) an H formula φ_i that does not start with $\exists$
   b) a pair $\langle \exists x \varphi_i(x), t\rangle$ where $\exists x \varphi_i(x)$ is an H formula and t is a closed term or
   c) a pair $\langle \exists X \varphi_i(X), U\rangle$ where $\exists X \varphi_i(X)$ is an H formula and U is either a second order variable or second order constant.

   We require that the sequence $A_0, A_1, ...$ covers all formulas and pairs of the form we describe; moreover, we require that each such formula and pair occur infinitely many times in the enumeration.

A sequent Δ is *axiomatic* if it contains an instance of one of the axioms (1)–(5).

Definition 5.3.2. A *deduction chain* for a sequent Γ and a countable fragment of H of $\mathbf{L}_2$ is a finite sequence $\Delta_0, \Delta_1, ..., \Delta_k$ of sequents (i.e., countable sets) of constant $\mathbf{L}_2$ formulas satisfying the following conditions:

1. Δ_0 is the sequent Γ.
2. For all numbers i less than k, Δ_i is not axiomatic.
3. If A_i is $\varphi \wedge \psi$ and $\varphi \wedge \psi \in \Delta_i$, then Δ_{i+1} is either Δ_i, φ or Δ_i, ψ.
4. If A_i is $\varphi \vee \psi$ and $\varphi \vee \psi \in \Delta_i$, then Δ_{i+1} is Δ_i, φ, ψ.
5. If A_i is $\forall x \varphi(x)$ and $\forall x \varphi(x) \in \Delta_i$, then, for some $n \in \mathbb{N}$, Δ_{i+1} is $\Delta_i, \varphi(\bar{n})$.
6. If A_i is $\forall X \varphi(X)$ and $\forall X \varphi(X) \in \Delta_i$, then Δ_{i+1} is $\Delta_i, \varphi(Y_i)$.
7. If A_i is $\langle \exists x \varphi(x), t \rangle$ and $\exists x \varphi(x) \in \Delta_i$, then Δ_{i+1} is $\Delta_i, \varphi(t)$.
8. If A_i is $\langle \exists X \varphi(X), U \rangle$ and $\exists X \varphi(X) \in \Delta_i$, then Δ_{i+1} is $\Delta_i, \varphi(U)$.
9. Otherwise, $\Delta_{i+1} = \Delta_i$.

This concludes the definition of *deduction chains*.

Definition 5.3.3. Given a sequent Γ and countable fragment H, we write $\mathbb{DT}[\Gamma, H]$ to denote the ω branching tree of all deduction chains for Γ and H. We call $\mathbb{DT}[\Gamma, H]$ the *canonical tree* of Γ, H.

Remark 5.3.4. Note that the tree $\mathbb{DT}[\Gamma, H]$ constitutes a cut-free pre-proof in our proof system. So if $\mathbb{DT}[\Gamma, H]$ is well-founded, then $\mathbb{DT}[\Gamma, H]$ constitutes a cut-free ω proof of Γ.

The following standard lemma follows from the definition of deduction chains.

Lemma 5.3.5. *Suppose $\mathbb{DT}[\Gamma, H]$ is ill-founded with path $\mathbb{P}$. Then:*

1. *$\mathbb{P}$ does not contain any literals that are true in $\mathbb{N}$.*
2. *$\mathbb{P}$ does not contain formulas $s \in K_i$ and $t \notin K_i$ for constant terms s and t such that $s^{\mathbb{N}} = t^{\mathbb{N}}$.*
3. *If $\mathbb{P}$ contains $E_0 \vee E_1$, then $\mathbb{P}$ contains E_0 and E_1.*
4. *If $\mathbb{P}$ contains $E_0 \wedge E_1$, then $\mathbb{P}$ contains E_0 or E_1.*
5. *If $\mathbb{P}$ contains $\exists x F(x)$, then $\mathbb{P}$ contains $F(\bar{n})$ for all n.*
6. *If $\mathbb{P}$ contains $\forall x F(x)$, then $\mathbb{P}$ contains $F(\bar{n})$ for some n.*

7. *If* $\mathbb{P}$ *contains* $\exists X F(X)$*, then* $\mathbb{P}$ *contains* $F(U)$ *for all set variables and constants* U*.*

8. *If* $\mathbb{P}$ *contains* $\forall X F(X)$*, then* $\mathbb{P}$ *contains* $F(U)$ *for some set variable/constant* U*.*

To see why clauses 1 and 2 of Lemma 5.3.5 are true, note that if $\mathbb{P}$ contained a true atomic sentence φ, then φ would belong to an axiomatic sequent, but by definition deduction chains do not contain axiomatic sequents.

Now we are ready to prove our completeness theorems for ω models.

Theorem 5.3.6. *(*ACA_0^+*) For any closed sequent* Γ *the following are equivalent:*

1. $\vdash_0 \Gamma$*;*

2. *There exists a family* S *of sets such that for any* ω*-model* $\mathfrak{M} \supseteq S$ *we have* $\mathfrak{M} \models \Gamma$*.*

Proof. (1) implies (2) follows from the soundness of the proof system with respect to ω-models.

For (2) implies (1) we prove the contrapositive. Assume that Γ does not have a cut-free ω proof. Let S be a family of sets and let H be a countable fragment in which all sets in S are named. Note that $\mathbb{DT}[\Gamma, H]$ is ill-founded; otherwise, it would constitute a cut-free ω proof of Γ. We will use an infinite path through $\mathbb{DT}[\Gamma, H]$ to define an ω model $\mathfrak{M}$ containing the sets named in H (and so *a fortiori* the sets in S) such that Γ fails in $\mathfrak{M}$.

Let P be a path through $\mathbb{DT}[\Gamma, H]$, and let $\mathbb{P}$ be the set of all formulas that occur in P. For any set term (variable or constant) K, we now assign a subset $\mathsf{val}(K)$ of $\mathbb{N}$ to K as follows:

$$\mathsf{val}(K) := \{t^{\mathbb{N}} : \ t \text{ is a constant } \mathbf{L}_2 \text{ term and } (t \notin K) \text{ belongs to } \mathbb{P}\}.$$

It is easy to verify, given the axioms of our proof system, that for any $C_A \in H$, $\mathsf{val}(C_A)$ is the set A.

Let M be the weak ω model given by relativizing the second-order quantifiers to the disjoint union of the values $\mathsf{val}(K_n)$. Since we are reasoning in ACA_0^+ we may enrich M with a full satisfaction class, yielding an ω model $\mathfrak{M}$. An induction on the complexity of formulas (making use of Lemma 5.3.5) shows that for any formula φ, $\varphi \in \mathbb{P}$ only if $\mathfrak{M} \not\models \varphi$. Thus, the assumption that $\mathbb{DT}[\Gamma, H]$ is ill-founded implies that there is an ω model $\mathfrak{M}$ containing each set named by a constant in H in which every sentence in Γ is false. ❑

5.4 Reduction for ω-model reflection

In this section we prove the main result of this chapter. First we prove a lemma, which can be viewed as an analogue of Feferman's completeness theorem for iterated Π^1_1 reflection. Then we prove Theorem 5.4.6 (Theorem 5.1.1 in the introduction), which provides a reduction of ω model reflection to iterated syntactic reflection.

An analogue of Feferman's theorem

The ω rule provides one route to proving all arithmetical truths; indeed, the recursive ω rule suffices as shown by Shoenfield in [91]. Feferman provided another route in [26]. Recall that for a theory T in the language of first-order arithmetic, the uniform reflection schema $\mathsf{RFN}(T)$ for T is the set of all sentences of the form:

$$\forall \vec{x}\Big(\mathsf{Pr}_T\big(\varphi(\vec{x})\big) \to \varphi(\vec{x})\Big)$$

where $\varphi(\vec{x})$ is a formula in the language of first-order arithmetic. Given an effective ordinal notation system $\prec$ we may then use the fixed point lemma to define the iterates of uniform reflection as follows:

$$\begin{aligned}\mathsf{RFN}^0(T) &:= T\\ \mathsf{RFN}^\alpha(T) &:= T + \bigcup_{\beta<\alpha} \mathsf{RFN}\big(\mathsf{RFN}^\beta(T)\big) \text{ for } \alpha \succ 0.\end{aligned}$$

Theorem 5.4.1 (Feferman)**.** *For any true arithmetical sentence φ, there is a representation α of a recursive ordinal such that $\mathsf{PA} + \mathsf{RFN}^\alpha(\mathsf{PA}) \vdash \varphi$.*

Feferman's proof makes crucial use of Shoenfield's completeness theorem for the recursive ω rule. In particular, Feferman shows that applications of the recursive ω rule can be simulated by iterating uniform reflection along a carefully selected ordinal notation. In [88], Schmerl cites this result (among others) as evidence that the uniform reflection principle is a formalized analogue of the ω rule.

In this subsection we will show that if a sequent of $\mathbf{\Pi}^1_1$ formulas can be proved from a $\mathbf{\Pi}^1_2$ axiomatized theory T by applying the ω rule, then it can also be proved by iterating $\mathbf{\Pi}^1_1$ reflection. Thus, our main lemma is an analogue of Feferman's completeness theorem.

Lemma 5.4.2. *Suppose that T is a $\mathbf{\Pi}^1_2$ axiomatized theory, Γ is a sequent of $\mathbf{\Pi}^1_1$ formulas, and P is a cut-free ω-proof of $\neg T, \Gamma$ with Kleene-Brouwer rank δ. Then $\mathbf{\Pi}^1_1\text{-}\mathbf{R}^\delta(T) \vdash \bigvee \Gamma$.*

Proof. Let Λ be the statement of the lemma. We will prove Λ by Löb's Theorem. That is, we will work in ACA_0 and prove the statement $\mathsf{Pr}_{\mathsf{ACA}_0}(\Lambda) \to \Lambda$. It will then follow by Löb's Theorem that ACA_0 proves Λ.

So work in ACA_0 and suppose that the statement of the lemma is provable in ACA_0. Let T and Γ be as in the statement of the theorem. Let δ be the Kleene-Brouwer rank of the canonical tree P for $\neg T, \Gamma$. We split into cases based on the final rule applied in P.

In each case $\neg T, \Gamma$ is being inferred from a sequence of sequents Δ_i which are the conclusions of canonical trees with Kleene-Brouwer ranks $\delta_i < \delta$. Our initial assumption that the statement of the lemma is provable in ACA_0 yields that

$$\mathsf{ACA}_0 \vdash \text{ “for all } i,\ \mathbf{\Pi}^1_1\text{-}\mathbf{R}^{\delta_i}(T) \text{ proves } \bigvee \Delta_i\text{.”}$$

Which straightforwardly implies

$$\mathsf{ACA}_0 \vdash \text{ “for all } i\text{, if } \boldsymbol{\Pi}^1_1\text{-}\mathbf{R}^\delta(T) \text{ then } \mathsf{True}_{\boldsymbol{\Pi}^1_1}(\bigvee \Delta_i)\text{.”}$$

Which in turn implies

$$\boldsymbol{\Pi}^1_1\text{-}\mathbf{R}^\delta(T) \vdash \forall i \mathsf{True}_{\boldsymbol{\Pi}^1_1}(\bigvee \Delta_i).$$

It suffices to check that this guarantees that $\boldsymbol{\Pi}^1_1\text{-}\mathbf{R}^\delta(T) \vdash \bigvee \Gamma$.

Since our canonical tree is cut-free, for each i, $\bigvee \Delta_i$ consists of $\boldsymbol{\Sigma}^1_2$ formulas (subformulas of negations of axioms of T) and $\boldsymbol{\Pi}^1_1$ formulas (subformulas of members of Γ). $\boldsymbol{\Pi}^1_1\text{-}\mathbf{R}^\delta(T)$ automatically rejects the negations of T's axioms and so accepts the $\boldsymbol{\Pi}^1_1$ parts of these sequents (consisting only of subformulas of members of Γ). Then after checking, case-by-case, the soundness of each proof rule, $\boldsymbol{\Pi}^1_1\text{-}\mathbf{R}^\delta(T)$ infers $\bigvee \Gamma$ from $\forall i \mathsf{True}_{\boldsymbol{\Pi}^1_1}(\bigvee \Delta_i)$. ❑

A special case of the main theorem

Before proving the main theorem we quickly prove a special case of the main theorem, namely, for $T = \mathsf{ACA}_0$. We will appeal to this special case in our proof of the main theorem. The special case follows easily from combining a theorem of Marcone and Montalbán with a theorem of the authors. First, the theorem of Marcone–Montalbán from [66]:

Theorem 5.4.3. *(Marcone–Montalbán)* RCA_0 *proves that the following are equivalent:*

1. *Every set is contained in an* ω *model of* ACA_0.
2. $\forall \alpha \Big(\mathsf{WO}(\alpha) \to \mathsf{WO}(\varepsilon_\alpha) \Big)$.

The following theorem is merely a relativization of a theorem of the authors in a previous paper; see Theorem 1.5 in [72]:

Theorem 5.4.4. *Provably in* ACA_0*, for any* α*, if* $\boldsymbol{\Pi}^1_1\text{-}\mathsf{RFN}\big(\boldsymbol{\Pi}^1_1\text{-}\mathbf{R}^\alpha(\mathsf{ACA}_0)\big)$ *then* $\mathsf{WO}(\varepsilon_\alpha)$.

Now the lemma we need follows easily.

Lemma 5.4.5. *(*ACA_0*) Suppose* $\forall \alpha \big(\mathsf{WO}(\alpha) \to \boldsymbol{\Pi}^1_1\text{-}\mathsf{RFN}(\boldsymbol{\Pi}^1_1\text{-}\mathbf{R}^\alpha(\mathsf{ACA}_0)) \big)$. *Then every set is contained in an* ω *model of* ACA_0.

Proof. Reason in ACA_0. Suppose $\forall \alpha \big(\mathsf{WO}(\alpha) \to \boldsymbol{\Pi}^1_1\text{-}\mathsf{RFN}(\boldsymbol{\Pi}^1_1\text{-}\mathbf{R}^\alpha(\mathsf{ACA}_0)) \big)$. By Theorem 5.4.3 it suffices to show that for any α, if α is well-ordered then so is ε_α. We reason as follows:

$$\begin{aligned} \mathsf{WO}(\alpha) &\Rightarrow \boldsymbol{\Pi}^1_1\text{-}\mathsf{RFN}\big(\boldsymbol{\Pi}^1_1\text{-}\mathbf{R}^\alpha(\mathsf{ACA}_0)\big) \text{ by assumption} \\ &\Rightarrow \mathsf{WO}(\varepsilon_\alpha) \text{ by Theorem 5.4.4} \end{aligned}$$

This completes the proof of the lemma. ❑

The main theorem

Now for the proof of the main theorem:

Theorem 5.4.6. *For any $\mathbf{\Pi}^1_2$-axiomatizable theory T, the following are equivalent:*

1. *Every set is contained in an ω model of T.*
2. $\forall\alpha\big(\mathsf{WO}(\alpha) \to \mathbf{\Pi}^1_1\text{-}\mathsf{RFN}(\mathbf{\Pi}^1_1\text{-}\mathbf{R}^\alpha(T))\big)$.

Proof. The 1→2 direction is relatively straightforward.

Assume, for contradiction, that 1 is true but 2 is false. Since 2 is false, there is a well-ordering α such that $\mathbf{\Pi}^1_1\text{-}\mathsf{RFN}(\mathbf{\Pi}^1_1\text{-}\mathbf{R}^\alpha(T))$ is false. So for some false $\mathbf{\Pi}^1_1$ sentence φ,

$$\mathbf{\Pi}^1_1\text{-}\mathbf{R}^\alpha(T) \vdash \varphi. \tag{5.1}$$

Note that $\neg\varphi$ is a true $\mathbf{\Sigma}^1_1$ statement. By 1, we infer that there is an ω model $\mathfrak{M}$ of T such that:

$$\mathfrak{M} \vDash \neg\varphi. \tag{5.2}$$

On the other hand, by induction, we can show that $\mathfrak{M}$ satisfies $\mathbf{\Pi}^1_1\text{-}\mathbf{R}^\alpha(T)$. Assume that for every $\beta < \alpha$, $\mathfrak{M} \vDash \mathbf{\Pi}^1_1\text{-}\mathbf{R}^\beta(T)$. If $\mathfrak{M} \nvDash \mathbf{\Pi}^1_1\text{-}\mathbf{R}^\alpha(T)$ then

$$\mathfrak{M} \vDash \exists\beta < \alpha \text{ ``}\mathbf{\Pi}^1_1\text{-}\mathbf{R}^\beta(T) \text{ proves a false } \mathbf{\Pi}^1_1 \text{ statement } \psi.\text{''}$$

Since $\mathfrak{M}$ is an ω-model, it is correct about what is provable. That is, this claim must be witnessed in $\mathfrak{M}$ by a standard proof. However, for any $\beta < \alpha$ and $\mathbf{\Pi}^1_1$ statement ψ, if $\mathbf{\Pi}^1_1\text{-}\mathbf{R}^\beta(T)$ proves ψ then since $\mathfrak{M}$ is a model of $\mathbf{\Pi}^1_1\text{-}\mathbf{R}^\beta(T)$, $\mathfrak{M}$ is a model of ψ, and thus, that ψ is a true $\mathbf{\Pi}^1_1$ statement.

Thus, we conclude that

$$\mathfrak{M} \vDash \mathbf{\Pi}^1_1\text{-}\mathbf{R}^\alpha(T) \tag{5.3}$$

But 5.1, 5.2, and 5.3 are jointly inconsistent.

The 2→1 direction is less straightforward, but we have already laid the groundwork. We assume 2. We want to prove 1, i.e., that every true $\mathbf{\Sigma}^1_1$ sentence is satisfied by an ω model of T. So let φ be a true $\mathbf{\Sigma}^1_1$ sentence. We break into cases based on whether there is a cut-free ω proof of $\neg T, \neg\varphi$.

Case I: There is no such proof, i.e., $\nvdash_0 \neg T, \neg\varphi$. By Lemma 5.4.5 assumption 2 implies that every set is contained in an ω model of ACA_0. That is, we are allowed to use theorems that are available in ACA^+_0. By Theorem 5.3.6, for every family S of sets there is an ω model $\mathfrak{M} \supseteq S$ satisfying $T + \varphi$. This yields 1.

Case II: There is such a proof, i.e., $\vdash_0 \neg T, \neg\varphi$. Let δ be the Kleene-Brouwer rank of the canonical proof tree of $\neg T, \neg\varphi$. By 2 we can iterate reflection along δ, yielding $\mathbf{\Pi}^1_1\text{-}\mathsf{RFN}(\mathbf{\Pi}^1_1\text{-}\mathbf{R}^\delta(T))$. On the other hand, by Lemma 5.4.2, $\mathbf{\Pi}^1_1\text{-}\mathbf{R}^\delta(T) \vdash \neg\varphi$. Combining these two observations, we conclude that $\neg\varphi$ is true, contradicting our choice of φ. ❑

Chapter 6

Incompleteness and jump hierarchies

This chapter contains joint work with Patrick Lutz.

6.1 Introduction

In this paper we explore a connection between Gödel's second incompleteness theorem and recursion-theoretic jump hierarchies. Our primary technical contribution is a method for proving the well-foundedness of jump hierarchies; this method crucially involves the second incompleteness theorem. We use this technique to provide a proof of the following theorem:

Theorem 6.1.1. *There is no sequence $(A_n)_{n<\omega}$ of reals such that, for each n, the hyperjump of A_{n+1} is hyperarithmetical in A_n.*

This theorem is an immediate consequence of a result of Spector's, namely that if $\mathcal{O}^A \leqslant_H B$ then $\omega_1^A < \omega_1^B$ (so the existence of such a sequence $(A_n)_{n<\omega}$ would imply the existence of a descending sequence $\omega_1^{A_0} > \omega_1^{A_1} > \dots$ in the ordinals). We provide an alternative proof that makes no mention of admissible ordinals, and which has the additional benefit of showing the theorem is provable in ACA_0.

Here is a brief sketch of how our alternative proof works: Consider the theory $\mathsf{ACA}_0 + \mathsf{DS}$ where DS is a sentence asserting the existence of a sequence of reals as described in Theorem 6.1.1. We work *inside* the theory and let $A_0, A_1, \dots$ be such a sequence. ACA_0 proves that if the hyperjump of a real exists then there is a β-model (a model that is correct for Σ^1_1 sentences) containing it. In this case $\mathcal{O}^{A_1}$ exists so there is a β-model containing A_1. Moreover, since all A_n's for $n \geqslant 1$ are hyperarithmetical in A_1, the β-model will contain all of them. All β-models are models of ACA_0 (in fact, ATR_0) so it appears this model is a model of the theory $\mathsf{ACA}_0 + \mathsf{DS}$, meaning that the theory proves its own consistency. By Gödel's second incompleteness theorem, this implies that ACA_0 proves $\neg\mathsf{DS}$.

There is one problem, however. Just because the model contains all the elements of the sequence $(A_n)_{n\geqslant 1}$ does not mean it contains the sequence itself (here we are thinking of the sequence as a single real whose slices are the A_n's). Indeed, the sequence itself could be

much more complicated than any single real in the sequence. In our proof, we overcome this flaw by showing that if there is a descending sequence then there is a descending sequence that is relatively simple—in fact there is one that is hyperarithmetic relative to A_1. This means the β-model above really does contain a descending sequence.

In [31], H. Friedman uses similar ideas to prove the following theorem originally due to Steel:

Theorem 6.1.2 (Steel)**.** *Let $P \subset \mathbb{R}^2$ be arithmetic. Then there is no sequence $(A_n)_{n<\omega}$ such that for every n,*

(i) $A_n \geqslant_T A'_{n+1}$ and

(ii) A_{n+1} is the unique B such that $P(A_n, B)$.

In these proofs we move from the second incompleteness theorem to the well-foundedness (or near well-foundedness) of recursion-theoretic jump hierarchies. In fact, the implication goes in both directions: the well-foundedness of appropriate jump hierarchies entails semantic versions of the second incompleteness theorem. For example, theorem 6.1.1 yields a simple and direct proof of the following semantic version of the second incompleteness theorem originally due to Mummert and Simpson (recall that $\mathcal{L}_2$ is the standard two-sorted language of second order arithmetic):

Theorem 6.1.3 (Mummert–Simpson)**.** *Let T be a recursively axiomatized $\mathcal{L}_2$ theory. For each $n \geqslant 1$, if there is a β_n-model of T then there is a β_n-model of T which contains no countable coded β_n-models of T.*

In fact, our proof sharpens the Mummert-Simpson result somewhat by dropping the requirement that T be recursively axiomatized.

A different semantic version of the second incompleteness theorem also follows from theorem 6.1.2, as observed by Steel in [98]. Namely, the following:

Theorem 6.1.4 (Steel)**.** *Let T be an arithmetically axiomatized $\mathcal{L}_2$ theory extending ACA_0. If T has an ω-model then T has an ω-model which contains no countable coded ω-models of T.*

These results all point to a general connection between incompleteness and well-foundedness. Elucidating this connection is the central goal of this paper. Though many of the theorems we prove could also be proved from the application of known methods, we believe that the new techniques are more conducive to achieving our central goal. Additionally, our techniques are able to prove somewhat sharper results than the original methods.

We also investigate directly the well-founded hierarchy at the center of theorem 6.1.1. It follows from that theorem that the relation $A \prec B$ defined by $\mathcal{O}^A \leqslant_H B$ is a well founded partial order. We call the $\prec$ rank of a real its *Spector rank*. There is a recursion-theoretically natural characterization of the Spector ranks of reals:

Theorem 6.1.5. *For any real A, the Spector rank of A is α just in case ω_1^A is the $(1+\alpha)^{th}$ admissible ordinal.*

It follows, assuming suitable large cardinal hypotheses, that, on a cone, the Spector rank of X is ω_1^X.

Here is our plan for the rest of the paper. In §6.2 we describe related research. In §6.3 we prove the main theorem. In §6.4 we provide an alternative proof of the Mummert-Simpson theorem. In §6.5 we turn to the calculation of Spector ranks.

6.2 Second incompleteness & well-foundedness

The second incompleteness theorem implies the well-foundedness of various structures (in particular, sequences of models). In turn, the well-foundedness of structures sometimes yields a semantic version of the second incompleteness theorem (in the form of a minimum model theorem). It is worth emphasizing that the former argument does not rely on the theory of transfinite ordinals and the latter argument does not rely on self-reference or fixed point constructions. This point allows us to sharpen certain results. Because we avoid the use of ordinals, we can verify that Theorem 6.1.1 is provable in ACA_0; because we avoid self-reference, we can drop the restriction in the statement of Theorem 6.1.3 that T be recursively axiomatized.

We will now describe both types of arguments, describe their historical antecedents, and point to related research.

Well-foundedness via incompleteness

To derive well-foundedness from incompleteness we work in the theory $T+$ "there is a descending sequence," where T is sound and sufficiently strong. We build a model of T containing a tail of the sequence, yielding a consistency proof of $T+$ "there is a descending sequence" *within the theory* $T+$ "there is a descending sequence." By the second incompleteness theorem, this means that T proves that there are no descending sequences.

The main difficulties lie in building a model that is correct enough that if a descending sequence is in the model, the model knows it is descending and in finding a T that is strong enough to prove the model exists but weak enough that the model built satisfies it.

As far as we know, the first arguments of this type are due to H. Friedman. We were inspired, in particular, by H. Friedman's [31] proof of a theorem originally due to Steel [98].

Theorem 6.1.2 (Steel)**.** *Let $P \subset \mathbb{R}^2$ be arithmetic. Then there is no sequence $(A_n)_{n<\omega}$ such that for every n,*

(i) $A_n \geqslant_T A'_{n+1}$ *and*

(ii) A_{n+1} *is the unique B such that $P(A_n, B)$.*

Steel's proof is purely recursion-theoretic, whereas Friedman's proof appeals to the second incompleteness theorem. In particular, Friedman supposes that there is an arithmetic counter-example P to Steel's Theorem. He then works in the theory $T := \mathsf{CCA} +$ "P produces a descending sequence" and uses P to build ω-models of arbitrarily large fragments of T. This yields a proof of $\mathsf{Con}(T)$ in T, whence T is inconsistent by Gödel's second incompleteness theorem.

Recently, Pakhomov and the second named author developed proof-theoretic applications of this technique in [72]. They show that there is no sequence $(T_n)_{n<\omega}$ of Π^1_1 sound extensions of ACA_0 such that, for each n, T_n proves the Π^1_1 soundness of T_{n+1}. This result is proved by appeal to the second incompleteness theorem, though it could be proved by showing that a descending sequence $(T_n)_{n<\omega}$ of theories would induce a descending sequence in the ordinals (namely, the associated sequence of proof-theoretic ordinals). They also show that, "on a cone," the rank of a theory in this well-founded ordering coincides with its proof-theoretic ordinal. These results are strikingly similar to the main theorems of this paper.

Incompleteness via well-foundedness

Here is an informal argument for incompleteness via well-foundedness. Suppose that second incompleteness fails, i.e. that a consistent T proves its own consistency. If T also proves the completeness theorem, then every model $\mathfrak{M}$ of T has (what it is by the lights of $\mathfrak{M}$) a model within it. This produces a nested sequence of models. If these models can be indexed by ordinals, then this produces a descending sequence of ordinals. So the well-foundedness of the ordinals produces some form of the second incompleteness theorem. If we know that the models form a well-founded structure, we can argue directly, without the detour through the ordinals.

To sharpen this argument one must know that the objects that are "models of T" in the sense of $\mathfrak{M}$ are genuinely models of T. So one must restrict one's attention to structures that are sufficiently correct. In addition, one must clarify the relation by which the models are being compared and prove that it is well-founded.

An early argument of this sort is attributed to Kuratowski (see [47, 56]). Set theory cannot prove the following strong form of the consistency of set theory: that there is an α such that V_α is a model of set theory. For if it does then there is α such that V_α is a model of set theory. Since V_α is a model of set theory, there is also a $\beta < \alpha$ such that V_β is a model of set theory. Iterating this argument produces an infinite descending sequence of ordinals. Contradiction.

Steel has also developed an argument of this sort. Using his Theorem 6.1.2, he demonstrates that if an arithmetically axiomatized theory of second order arithmetic extends ACA_0 and has an ω-model then it has an ω-model which contains no countable coded ω-models of the theory.

Kripke structures

We conclude this discussion of related work with the following observation. Formalized in the language of modal logic, the statement of Gödel's second incompleteness theorem characterizes well-founded Kripke frames. Indeed, the formalization corresponds to the least element principle:

$$\Diamond\varphi \to \Diamond(\varphi \wedge \neg\Diamond\varphi).$$

Its contrapositive (writing ψ for $\neg\varphi$) is a modal formalization of Löb's theorem which corresponds to induction[1]:

$$\Box(\Box\psi \to \psi) \to \Box\psi$$

Beklemishev has suggested that this observation is connected with ordinal analysis. In [8], he uses a modal logic of provability known as GLP to develop both an ordinal notation system for ε_0 and a novel consistency proof of PA.

6.3 The main theorem

In this section we provide our alternative proof of Theorem 6.1.1.

Outline of proof

In broad strokes, here is our strategy. We will consider a statement DS which states that there *is* a descending sequence in the hyperjump hierarchy. We then work in the theory $\mathsf{ACA}_0 + \mathsf{DS}$ and derive the statement $\mathsf{Con}(\mathsf{ACA}_0 + \mathsf{DS})$. By Gödel's second incompleteness theorem, this implies that there is a proof of $\neg\mathsf{DS}$ in ACA_0.

To derive $\mathsf{Con}(\mathsf{ACA}_0 + \mathsf{DS})$ in $\mathsf{ACA}_0 + \mathsf{DS}$, we use the hyperjump of a real to construct a coded β-model of ACA_0 containing that real. In particular, if we are given a descending sequence then we can use the existence of the hyperjump of the second real in the sequence to find a β-model containing all the elements of the tail of the sequence. The point is that the tail of a descending sequence is again a descending sequence and β-models are correct enough to verify this.

The only problem is that while the β-model we found contains all the elements of the tail it may not contain the tail itself (i.e. it may not contain the recursive join of all the elements of the tail). Our strategy to fix this is to essentially to show that there is a family of descending sequences which is arithmetically definable relative to some parameter whose hyperjump exists. A β-model containing this parameter must then contain an element of this family (because β-models contain witnesses to all Σ^1_1 statements).

For the parameter, we will use a countable coded β-model which contains a tail of the original descending sequence. The arithmetic formula will then essentially say that the β-model believes each step along the sequence is descending. The point is that we have replaced

[1]Note that since we replaced ϕ with $\neg\phi$ before taking the contrapositive, the two modal statements are equivalent only as *schemas*.

a Π^1_1 formula saying the sequence is descending by an arithmetic formula talking about the truth predicate of some coded model and that β-models are correct enough that this does not cause any errors.

The β-model will just come from the existence of the hyperjump of some element of the original sequence, and we can guarantee the hyperjump of the model exists by taking one more step down the original descending sequence.

Useful facts

In this section, we record the facts about β-models that we will use in the proof of the main theorem. Unless otherwise noted, proofs of all propositions in this section can be found in [92].

Definition 6.3.1. A β-model is an ω-model M of second order arithmetic such that for any Σ^1_1 sentence φ with parameters in M, $M \vDash \varphi$ if and only if φ is true.

Proposition 6.3.2 ([92], Lemma VII.2.4, Theorem VII.2.7)**.** *Provably in* ACA_0*, all countable coded* β*-models satisfy* ATR_0 *(and hence also* ACA_0*).*

Proposition 6.3.3 ([92], Lemma VII.2.9)**.** *Provably in* ACA_0*, for any* X*,* $\mathcal{O}^X$ *exists if and only if there is a countable coded* β*-model containing* X*.*

Proposition 6.3.4. *All of the following can be written as Boolean combinations of* Σ^1_1 *formulas and hence are absolute between* β*-models*

1. *A is the hyperjump of B.*
2. *$A \leqslant_H B$*
3. *M is a countable coded β-model.*

Proof of the main theorem

Theorem 6.1.1. *There is no sequence* $(A_n)_{n<\omega}$ *of reals such that, for each* n*, the hyperjump of* A_{n+1} *is hyperarithmetical in* A_n*.*

Proof. It suffices to prove the inconsistency of the theory $\mathsf{ACA}_0 + \mathsf{DS}$, where

$$\mathsf{DS} := \exists X \forall n (\mathcal{O}^{X_{n+1}} \text{ exists and } \mathcal{O}^{X_{n+1}} \leqslant_H X_n).$$

To do this, we reason in $\mathsf{ACA}_0 + \mathsf{DS}$ and derive $\mathsf{Con}(\mathsf{ACA}_0 + \mathsf{DS})$. The inconsistency of $\mathsf{ACA}_0 + \mathsf{DS}$ then follows from Gödel's second incompleteness theorem.

Reasoning in $\mathsf{ACA}_0 + \mathsf{DS}$:

Let A witness DS. That is, for all n, $\mathcal{O}^{A_{n+1}}$ exists and $\mathcal{O}^{A_{n+1}} \leqslant_H A_n$. Our goal is now to show there is a model of $\mathsf{ACA}_0 + \mathsf{DS}$.

Claim. *There is a countable coded β-model $\mathfrak{M}$ coded by M such that $\mathcal{O}^M$ exists and $\mathfrak{M}$ contains A_n for all sufficiently large n.*

The proof of Proposition 6.3.3 in [92] actually shows that for any X, if $\mathcal{O}^X$ exists then X is contained in a countable coded β-model which is coded by a real that is recursive in $\mathcal{O}^X$. So A_2 is contained in some countable coded β-model $\mathfrak{M}$, coded by M, such that $M \leqslant_T \mathcal{O}^{A_2} \leqslant_H A_1$. Hence $\mathcal{O}^M \leqslant_T \mathcal{O}^{A_1}$. Since $\mathcal{O}^{A_1}$ exists, so does $\mathcal{O}^M$. And since $\mathfrak{M}$ is closed under hyperarithmetic reducibility, $\mathfrak{M}$ contains A_n for all $n \geqslant 2$.

Claim. *There is an arithmetic formula φ such that*

(i) $\exists X\, \varphi(M, X)$

(ii) For any X, if $\varphi(M, X)$ holds then X is a witness of DS

where M is as in the previous claim.

Basically $\varphi(M, X)$ says that X is a sequence of reals whose elements are in $\mathfrak{M}$ and for each n, $\mathfrak{M}$ believes that $\mathcal{O}^{X_{n+1}}$ exists and is hyperarithmetical in X_n. More precisely $\varphi(M, X)$ is the sentence

$$\forall n\, (X_{n+1}, X_n \in \mathfrak{M} \wedge \mathfrak{M} \vDash \text{“}\exists Y\, [Y = \mathcal{O}^{X_{n+1}} \wedge Y \leqslant_H X_n]\text{”}).$$

To see why $\varphi(\mathfrak{M}, X)$ has a solution, recall that $\mathfrak{M}$ contains A_n for all n sufficiently large. Let X be the sequence A but with the first few elements removed so that $\mathfrak{M}$ contains all elements in X. For each n, the fact that A is a witness of DS guarantees that there is some Y such that $\mathcal{O}^{X_{n+1}} = Y$ and $Y \leqslant_H X_n$. Since $\mathfrak{M}$ contains X_n and since β-models are closed under hyperarithmetic reducibility, $\mathfrak{M}$ contains Y. And by proposition 6.3.4, β-models are sufficiently correct that $\mathfrak{M} \vDash \text{“}Y = \mathcal{O}^{X_{n+1}} \wedge Y \leqslant_H X_n$.”

Suppose X is a sequence such that $\varphi(M, X)$ holds. Then for each n there is a Y such that $\mathfrak{M} \vDash \text{“}Y = \mathcal{O}^{X_{n+1}} \wedge Y \leqslant_H X_n$.” By proposition 6.3.4, both clauses of the conjunction are absolute between β-models. Hence $\mathcal{O}^{X_{n+1}}$ exists and is hyperarithmetical in X_n. So X is a witness of DS.

Claim. *There is a model of* $\mathsf{ACA}_0 + \mathsf{DS}$.

By proposition 6.3.3, there is a β-model $\mathfrak{N}$ that contains M. Since $\mathfrak{N}$ is a β-model, by proposition 6.3.2, it is a model of ACA_0.

Since the Σ^1_1 formula $\exists X\, \varphi(M, X)$ holds and $\mathfrak{N}$ is correct for Σ^1_1 formulas with parameters from $\mathfrak{N}$, there is some X in $\mathfrak{N}$ such that $\mathfrak{N} \vDash \varphi(X, M)$. And since $\mathfrak{N}$ is a β-model, it is correct about this fact—that is, $\varphi(X, M)$ really does hold. Since $\varphi(M, X)$ holds, X is a witness to DS. The point now is just that $\mathfrak{N}$ is correct enough to see that X is a witness to DS. In detail: for each n, $\mathcal{O}^{X_{n+1}}$ exists and is hyperarithmetical in X_n. Since X_n is in $\mathfrak{N}$, this means $\mathcal{O}^{X_{n+1}}$ is in $\mathfrak{N}$. And by proposition 6.3.4, $\mathfrak{N}$ agrees that it is the hyperjump of X_{n+1} and that it is hyperarithmetical in X_n. Therefore $\mathfrak{N}$ agrees that X is a witness to DS. ❑

Remark 6.3.5. The previous proof actually demonstrates that ACA_0 proves Theorem 6.1.1. The original Spector proof relies on the theory of admissible ordinals, so it is unlikely to be formalizable in systems weaker than ATR_0.

6.4 Semantic incompleteness theorems

Steel derives the following theorem as a corollary of his Theorem 6.1.2.

Theorem 6.1.4 (Steel)**.** *Let T be an arithmetically axiomatized $\mathcal{L}_2$ theory extending ACA_0. If T has an ω-model then T has an ω-model which contains no countable coded ω-models of T.*

Because ω-models are correct for arithmetic statements, we can restate this as

Corollary 6.4.1. *Let T be an arithmetically axiomatized $\mathcal{L}_2$ theory extending ACA_0. If there is an ω-model of T then there is an ω-model of T + "there is no ω-model of T".*

Similarly, we can use Theorem 6.1.1 to prove the following theorem originally proved in a slightly weaker form by Mummert and Simpson in [69]. Note that this time we do not need any assumptions about the theory T.

Theorem 6.4.2. *Let T be an $\mathcal{L}_2$ theory. If there is a β-model of T then there is a β-model of T that contains no countable coded β-models of T.*

Proof. Suppose not. Then every β-model of T contains a countable coded β-model of T. Let $\mathfrak{M}$ be a β-model of T. So $\mathfrak{M}$ contains some countable coded β-model $\mathfrak{N}_0$ coded by a real N. Similarly $\mathfrak{N}_0$ contains a countable coded β-model of T, $\mathfrak{N}_1$, coded by a real N_1. In this manner we can define a sequence of countable β-models of T, $\mathfrak{N}_0, \mathfrak{N}_1, \mathfrak{N}_2, \ldots$ along with their codes $N_0, N_1, N_2, \ldots$

But for each n, $N_{n+1} \in \mathfrak{N}_n$ and since $\mathfrak{N}_n$ is a β-model it is correct about all Π^1_1 facts about N_{n+1}. In other words, $\mathcal{O}^{N_{n+1}}$ is arithmetic in N_n. So $N_0, N_1, \ldots$ provides an example of the type of descending sequence in the hyperdegrees shown not to exist in theorem 6.1.1. ❑

In fact, this same proof actually yields a seemingly stronger result. A β_n-model is defined to be an ω-model of second order arithmetic which is correct for all Σ^1_n statements with parameters from the model. The same proof as above proves the theorem mentioned in the introduction (where once again our new proof shows that the assumption that T is recursively axiomatized can be dropped):

Theorem 6.1.3 (Mummert–Simpson)**.** *Let T be a recursively axiomatized $\mathcal{L}_2$ theory. For each $n \geqslant 1$, if there is a β_n-model of T then there is a β_n-model of T which contains no countable coded β_n-models of T.*

Since the statement that a real is the code for a β_n-model is Π^1_n, β_n-models are correct about such statements. Thus we can restate the above theorem to get the following sharpening of a theorem of Mummert and Simpson (once again dropping the requirement that T be recursively axiomatized):

Theorem 6.4.3. *Let T be an $\mathcal{L}_2$ theory. For each $n \geqslant 1$, if there is a β_n-model of T, then there is a β_n-model of*

$$T + \text{“there is no countable coded } \beta_n\text{-model of } T\text{.”}$$

From this we immediately infer the following corollary, a strengthened version of Mummert and Simpson's Corollary 2.4 from [69]:

Corollary 6.4.4. *Let T be an $\mathcal{L}_2$ theory. For each $n \geqslant 1$, if T has a β_n-model then T has a β_n model that is not a β_{n+1} model.*

Proof. Let T be an $\mathcal{L}_2$ theory with a β_n model. By Theorem 6.4.3, there is a β_n model $\mathfrak{M}$ of $T+$“there is no countable coded β_n-model of T.” The latter is a false Π^1_{n+1} sentence, whence $\mathfrak{M}$ is not a β_{n+1} model. ❑

6.5 Spector ranks

Define a relation $\prec$ on pairs of reals by $A \prec B$ iff $\mathcal{O}^A \leqslant_H B$. By theorem 6.1.1, this relation is well-founded and therefore reals can be assigned ordinal ranks according to it. Let's refer to the $\prec$-rank of a real as its *Spector rank*. In this section we will calculate the Spector ranks of reals, showing that we get the same ranks as those induced by the ω_1's of reals.

We will need to use the following theorem due to Spector:

Theorem 6.5.1 (Spector)**.** *For any reals A and B:*

1. *If $\mathcal{O}^B \leqslant_H A$ then $\omega_1^B < \omega_1^A$.*
2. *If $B \leqslant_H A$ and $\omega_1^B < \omega_1^A$ then $\mathcal{O}^B \leqslant_H A$.*

We will also need to use the following theorem of Sacks:

Theorem 6.5.2 (Sacks)**.** *If λ is an admissible ordinal greater than ω and X is a real such that X computes a presentation of λ (i.e. $\lambda < \omega_1^X$) then there is a real Y that is hyperarithmetical in X such that $\omega_1^Y = \lambda$.*

This theorem is typically stated without the requirement that Y is hyperarithmetical in X, though this is implicit in all or nearly all extant proofs of the theorem.

For instance, in [99] Steel uses the method of forcing with tagged trees to prove Sacks' theorem. In that case, the real Y is obtained as the reduct of a generic filter over L_λ. Since any presentation of λ can hypercompute such a generic (if you can compute a presentation

of λ then it just takes $\omega \cdot (\lambda + 1)$ jumps to compute the theory of L_λ), X can hypercompute a Y witnessing Sacks' theorem.

Using these two theorems, the calculation of Spector ranks follows relatively easily.

ranks*

Remark 6.5.3. The only reason we need to say $(1 + \alpha)^{\text{th}}$ admissible rather than the α^{th} admissible is that the way admissible is usually defined, ω is an admissible ordinal but unlike all other countable admissible ordinals, it is not the ω_1 of any real.

Proof. We will argue by induction on α that for any A if $rank(A) > \alpha$ then ω_1^A is greater than the $(1 + \alpha)^{\text{th}}$ admissible ordinal and conversely that if ω_1^A is greater than the $(1 + \alpha)^{\text{th}}$ admissible then $rank(A) > \alpha$.

First suppose $rank(A) > \alpha$. So there is some B of rank α such that $\mathcal{O}^B \leqslant_H A$. By Spector's result, theorem 6.5.1, this implies $\omega_1^B < \omega_1^A$. And by the induction assumption, ω_1^B is at least the $(1 + \alpha)^{\text{th}}$ admissible so ω_1^A is greater than the $(1 + \alpha)^{\text{th}}$ admissible.

Now suppose that ω_1^A is greater than the $(1 + \alpha)^{\text{th}}$ admissible. Let λ denote the $(1 + \alpha)^{\text{th}}$ admissible. By Sacks' theorem, there is some B hyperarithmetical in A such that $\omega_1^B = \lambda$. Since $\omega_1^B < \omega_1^A$, Spector's theorem implies that $\mathcal{O}^B \leqslant_H A$ and hence $rank(B) < rank(A)$. By the induction assumption, $rank(B)$ is at least α, so $rank(A) > \alpha$. ❑

Theorem 6.5.4 (Silver)**.** *If α is admissible relative to $0^\sharp$ then α is a cardinal in L.*

Hence if X is a real in the cone above $0^\sharp$ then ω_1^X is a cardinal in L. Suppose that ω_1^X is the α^{th} admissible. Since ω_1^X is a cardinal in L, it follows that actually $\alpha = \omega_1^X = \omega_\alpha^{CK}$. So if $0^\sharp$ exists then on a cone, the Spector rank of a real X is equal to ω_1^X.

Theorem 6.5.5. *If $0^\sharp$ exists, then for all A on a cone, the Spector rank of A is ω_1^A.*

Alternatively, one can infer the previous theorem from the following proposition due to Martin.

Proposition 6.5.6 (Martin)**.** *Assuming appropriate determinacy hypotheses, if F is a degree invariant function from reals to (presentations of) ordinals such that $F(A) \leqslant \omega_1^A$, then either F is constant on a cone or $F(A) = \omega_1^A$ on a cone.*

One could also consider the analogous relation given by replacing hyperarithmetic reducibility and the hyperjump with Turing reducibility and the Turing jump. Namely, define $\prec_T$ by $A \prec_T B$ iff $A' \leqslant_T B$. By results of Harrison (see [43]), this relation is not well-founded. However, it *is* well-founded if we restrict ourselves to the hyperarithmetic reals, as shown by Putnam and Enderton in [25]. In that paper, Putnam and Enderton also show that the rank of a hyperarithmetic real A in this relation is "within 2" of the least α such that A cannot compute $0^{(\alpha)}$. More precisely, if the rank of A is α then A cannot compute $0^{(\alpha+1)}$ and if A cannot compute $0^{(\alpha)}$ then the rank of A is at most $\alpha + 2$.

Part III

Foundational Reflections

Chapter 7

Epistemic aspects of consistency proofs

7.1 Introduction

Are the fundamental principles of mathematical reasoning consistent? By a *consistency proof* I mean a proof that some such principles are consistent. Though some consistency proofs are widely regarded as mathematically significant, it is controversial whether they are epistemically significant. This controversy has primarily concerned the prospects of using consistency proofs to address skepticism about consistency. I will focus on a different epistemic aspect of consistency proofs, namely their explanatory value. In particular, I will argue that a consistency proof can improve one's understanding of *why* a theory is consistent, regardless of how it affects one's credence *that* it is consistent. On this basis, I will argue against certain restrictive norms on mathematical practice.

Modern consistency proofs are descended from an early twentieth century research program known as *Hilbert's Program*. The program had two parts: (i) axiomatize mathematics and (ii) prove the consistency of the axioms. The consistency proof was intended to dispel skeptical worries about the consistency of classical mathematics.[1] To have dialectical force against skeptics, (ii) would have to have been carried out using only methods acceptable to skeptics. Thus, Hilbert proposed that the consistency proof use only the restricted methods of *finitistic mathematics*, which he considered unimpeachable. In 1931, Hilbert's program reached a major obstacle in the form of Gödel's [40] second incompleteness theorem.

Theorem 7.1.1 (Gödel)**.** *No reasonable axiomatic theory proves the consistency of its own axioms.*[2]

[1]This summary is inadequate for conveying the range of issues that Hilbert wished to address. Many contemporary logicians and philosophers have focused on other aspects of Hilbert's program, such as the elimination of ideal objects from proofs of real statements.

[2]By a reasonable theory I mean a consistent theory that is recursively axiomatized and that interprets a modicum of arithmetic. I am eliding other technical complications, e.g., that we must specify in advance

It follows from Gödel's theorem that if the principles of finitistic mathematics are codifiable in a reasonable axiomatic theory, then they do not prove their own consistency, much less the consistency of stronger theories.[3] Thus, it is generally agreed that Hilbert's program failed.[4] Some logicians were undeterred, however. Not long after Gödel proved the incompleteness theorems, Gentzen [35] produced a consistency proof of arithmetic.[5]

Theorem 7.1.2 (Gentzen)**.** *If the ordinal number ε_0 is well-founded, then the axioms of arithmetic are consistent.*[6]

Gentzen's proof was remarkably finitistic, despite the limitations of Gödel's theorem. The only principle invoked in the proof that is not obviously finitistically acceptable is the well-foundedness of the ordinal number ε_0. Gentzen's consistency proof marked the beginning of a research program known as *ordinal analysis*,[7] whereby similar consistency proofs have been discovered for a wide range of axiomatic theories. Ordinal analysis has not yet reached the level of classical analysis. Calculating the *proof theoretic ordinal* of analysis, i.e., determining which ordinal number must be invoked in a consistency proof of analysis, is among the most well-known and difficult open problems in mathematical logic.[8]

Gentzen [37] described his consistency proof as a "real *vindication* of the disputable parts of elementary number theory," but opinions have not converged on Gentzen's conclusion. Tarski famously remarked that Gentzen's proof increased his credence in arithmetic's consistency "only by an epsilon." More recently, Voedvodsky, a prominent mathematician and skeptic of consistency, called Gentzen's argument "suspicious" and "not very convincing."[9]

By proving a statement, one certifies the truth of that statement. However, in addition to certifying *that* a statement is true, some proofs also explain *why* a statement is true.[10] I do not expect that Gentzen's proof will increase anyone's credence that arithmetic is consistent.

---

some specific consistency predicate and presentation of the theory's axioms.

[3]There is not universal agreement as to what the axioms of finitistic mathematics are, an issue to which I will return in §3.1.

[4]Some philosophers, most notably Detlefsen, have argued that Hilbert's program can survive Gödel's second incompleteness theorem. See [21].

[5]When I write "arithmetic" without any qualification, I am referring to first-order Peano Arithmetic. When I write "analysis" or "classical analysis," I am referring to second-order Peano Arithmetic with full comprehension. When I write "set theory," I mean ZFC, i.e., Zermelo-Fraenkel set theory with the axiom of choice.

[6]The transfinite ordinal number ε_0 is the supremum of the sequence $\langle \omega, \omega^\omega, \omega^{\omega^\omega}, ... \rangle$, where ω is the least transfinite ordinal number. Gentzen's result is stated relative to a particular presentation of ε_0, a significant detail that I will mostly ignore in this paper, except when it is directly relevant to the discussion at hand.

[7]Gentzen's work was foreshadowed by work of Hilbert, Bernays, Ackermann, and von Neumann. See [111].

[8]I am ignoring a technical complication here, which is that one must determine this ordinal *relative* to a sufficiently natural ordinal notation system.

[9]Voevodsky made these remarks in his lecture *What If Current Foundations of Mathematics Are Inconsistent?* at the Institute for Advanced Study. For a discussion of Voevodsky's argument, see [19].

[10]This distinction between explanations and mere certifications can be found in Aristotle [3], and was famously applied to the philosophy of mathematics by Bolzano [18, 17].

Nevertheless, an explanation can enhance one's understanding of a fact without increasing one's credence in it. For instance, exposure to the theory of gravity might help me understand why I am not floating into space, even if it does not increase my credence that I am not floating into space. The main thesis of this paper is that Gentzen's consistency proof not only certifies *that* arithmetic are consistent, it also explains *why* it is consistent.

The conception of Gentzen's proof that I am developing is not motivated by and does not seek to address skeptical considerations. In this sense I am pursuing a project similar to Nozick's [70] account of the possibility of knowledge:

> Our goal is not...to refute skepticism, to prove it is wrong or even to argue that it is wrong.... Our task here is to explain how knowledge is possible.... In doing this, we need not convince the skeptic, and we may introduce explanatory hypotheses that he would reject. What is important for our task of explanation and understanding is that we find those hypotheses acceptable or plausible.... These hypotheses are to explain to ourselves how knowledge is possible, not to prove to someone else that knowledge is possible.

Schematically, Nozick's argument is this: There are skeptics who doubt that p is true. In seeking to explain why p is true, we are not trying to convince the skeptic. Indeed, we may argue that p is true on the basis of principles that the skeptic rejects. What is important is that we argue for p on the basis of principles that enhance our understanding of why p is true. The relevance to the present analysis of consistency proofs is this: For Nozick, p is the claim that knowledge is possible, whereas for me p is the claim that arithmetic is consistent.

Throughout this paper I will contrast Gentzen's proof with another familiar argument for the consistency of arithmetic. This latter argument, which I will refer to as the *Tarskian* consistency proof,[11] goes as follows:

> All the axioms of arithmetic are true. All the rules of inference preserve truth. So every theorem of arithmetic is true. Since no contradiction is true, no contradiction is a theorem of arithmetic. That is, arithmetic is consistent.

Very few would object that the premises of the Tarskian consistency proof are false, or that the inferences are invalid. However, it is natural to object that if the consistency of arithmetic is at issue, then the soundness of arithmetic is at issue as well, so it is inappropriate to appeal to soundness when arguing for consistency. Like Moore's argument that the external world exists, the Tarskian consistency proof seems forceful only to those who are already convinced of its conclusion. I believe that this objection is fine, *insofar* as the consistency of arithmetic *is* at issue. My criticism of the Tarskian consistency proof is that it does not explain *why* arithmetic is consistent. Indeed, whenever I provide some consideration

[11] I use this terminology because Tarski [108] demonstrated how to formalize this argument by extending the induction axioms to include induction on formulas including a truth predicate.

in favor of the explanatoriness of Gentzen's proof, I will argue that the same consideration suggests that the Tarskian proof is not explanatory.

The Tarskian consistency proof has recently become a focal point of interest in the philosophy of logic, since some claim that it is a counterexample to deflationary theories of truth. According to this line of reasoing, the Tarskian consistency proof explains why arithmetic is consistent, and so constitutes a genuine explanation that relies crucially on the concept of truth. Shapiro [89] provides the following statement of this view:

> Assume that there is an effective, sound theory A ... and that A contains the rudimentary axioms of arithmetic ... Suppose that a logic teacher asserts that [Con(A)] is true, and a puzzled student asks for an explanation ... The natural reply is to point out that all of the axioms of A are true and the rules of inference preserve truth. Thus, every theorem of A is true. It follows that "0=1" is not a theorem and so A is consistent ... It seems to me that this informal version of the derivation of [Con(A)] ... is as good an explanation as there is. The argument shows why [Con(A)] is true.

Note that Shapiro claims that the Tarskian consistency proof provides the *best possible explanation* for consistency.[12] I do not doubt that the Tarskian consistency proof enjoys virtues that are not shared by Gentzen's consistency proof. Nevertheless, I will argue, contra Shapiro, that Gentzen's is the more explanatory proof.

Here is my plan for the paper. In §2 I will present some logical preliminaries, including a sketch of Gentzen's consistency proof of arithmetic. In §3 I will review the philosophical positions that have been advanced regarding the epistemic significance of consistency proofs. A number of logicians have argued that Gentzen-style consistency proofs can address skeptical concerns coming from constructivists. I will review the difficulties facing this position. In §4 I will present and defend my positive view that Gentzen's proof explains why arithmetic is consistent. I will make an effort to show that the explanitoriness of Gentzen's proof follows from *all* of the leading accounts of mathematical explanation. In §5 I will discuss some normative consequences of my view. My position implies that consistency proofs may be epistemically valuable even if they have no dialectical force against skeptics. This undermines popular arguments for certain restrictive norms on mathematical practice.

A methodological remark is in order before continuing. My position is informed by a particular practice, namely, classical mathematics. Throughout this paper, I will appeal to the ordinary standards of classical mathematical reasoning. This is not uncontroversial; classical

[12] It is not entirely clear whether Shapiro is claiming that the Tarskian proof provides the best explanation when reasoning in general about sufficiently strong effective sound theories, or if he means to claim it about some specific such theory, which could, for instance, be PA. If it is the former, then Shapiro may very well be correct. However, asking for a consistency proof for sound theories in general is a strange request, since all sound theories are obviously consistent. On the other hand, asking of a consistency proof for a *specific* sound theory makes sense, since one might question whether the theory is sound (or how we can *prove* that the theory is sound). So the latter interpretation seems more plausible.

mathematics has been attacked on philosophical grounds by finitists, constructivists, predicativists, and others. I will not defend classical mathematics here, though I will articulate a position from within it, so to speak.

7.2 Logical preliminaries

In this section I will discuss some logical preliminaries. First I will present the axiomatic theory PA. Then I will discuss some of the core concepts of Gentzen's consistency proof of arithmetic.[13]

Peano Arithmetic

Axiomatic theories are central objects of study in mathematical logic. The axioms of the theories I will consider are stated in a particular formal language, *the language of arithmetic*. The language of arithmetic contains the constant symbols 0 and 1 and the function symbols + and ×. This language is standardly interpreted as concerning the natural numbers $\mathbb{N} = \{0, 1, 2, ...\}$. When I speak of the truth or falsity of a statement in the language of arithmetic, I mean truth or falsity with respect to this interpretation. The most widely studied axiomatic theory in this language is *Peano Arithmetic* or PA. PA is axiomatized by the following axioms:

1. $x + 1 \neq 0$
2. $x + 1 = y + 1 \rightarrow x = y$
3. $x + 0 = x$
4. $x + (y + 1) = (x + y) + 1$
5. $x \times 0 = 0$
6. $x \times (y + 1) = (x \times y) + x$

   as well as an induction axiom for each formula φ :

7. $\Big(\varphi(0) \wedge \forall x\big(\varphi(x) \rightarrow \varphi(x+1)\big)\Big) \rightarrow \forall x \varphi(x)$.

[13]Gentzen produced a number of consistency proofs of arithmetic. This discussion is informed primarily by Gentzen's most famous 1938 proof. In §4.3 I will discuss other ordinally informative consistency proofs of arithmetic that emerged from research programs besides Gentzen's.

When I say that PA is *consistent*, I mean that there is no proof that begins with the axioms just listed and ends with the statement "0=1." As I mentioned in §1, there is a very simple proof that PA is consistent.

Proposition 7.2.1. PA *is consistent.*

Proof. All the axioms of PA are true. All the rules of inference preserve truth. So every theorem of PA is true. Since "0=1" is not true, it is not a theorem of PA. So PA is consistent. ❑

I will argue that there is a much better explanation for the consistency of PA, namely, Gentzen's. There are two concepts that appear in Gentzen's consistency proof that I want to emphasize. The first is *cut-elimination* and the second is the *reduction procedure*. I will describe these and briefly discuss their role in Gentzen's consistency proof of PA.

Cut-elimination

There are many proof calculi for first-order logic, but I will focus on Gentzen's *sequent calculus* [36]. The sequent calculus is characterized by its *logical axioms* and a number of *inference rules*. The initial lines in a proof are logical axioms; if some lines precede another line then the latter must follow from the former by applying one of the inference rules. The final line contains the statement that has been proved.

The *logical axioms* of the sequent calculus are all the implications of the form $\varphi \Rightarrow \varphi$. The sequent calculus includes many inference rules.[14] To give one example, given that Γ implies φ, infer that Γ implies $(\varphi \vee \psi)$.

$$\frac{\Gamma \Rightarrow \varphi}{\Gamma \Rightarrow (\varphi \vee \psi)}$$

However, the rule that concerns us most is the *cut rule*, which expresses the transitivity of implication: given that Γ implies φ and φ implies Δ, infer that Γ implies Δ.

$$\frac{\Gamma \Rightarrow \varphi \qquad \varphi \Rightarrow \Delta}{\Gamma \Rightarrow \Delta}$$

Note that in the first rule I mentioned, all the formulas that appear in the top line also appear in the bottom line. Every inference rule in the sequent calculus has this feature, with the exception of the cut rule; notice how the formula φ is "cut" out of the proof.

[14] Any reader familiar with the sequent calculus will notice that I am not stating these rules in their full generality. I am ignoring side formulas, for example. I am doing this to minimize clutter and make the key points more transparent.

We are interested in the sequent calculus for two reasons. First, the sequent calculus is a complete proof system for first-order logic.[15] Second, the sequent calculus enjoys a property called *cut-elimination*.

Theorem 7.2.2 (Gentzen)**.** *There is an algorithm for transforming any proof into a proof of the same conclusion within which the cut rule is never applied.*

The proof of the cut-elimination theorem exploits the symmetry of the logical connectives; indeed, cut-elimination is typically regarded as an expression of the symmetry inherent in logic. Since the cut rule is the only rule that "cuts" formulas out of proofs, cut-free proofs have the crucial *subformula property*: any formula occurring in a cut-free proof is a subformula of a formula occurring in the final line. One can straightforwardly show that no proof of a contradiction can have the subformula property.[16] By the cut-elimination theorem, this means that there is no proof of a contradiction in the sequent calculus.

Before continuing, I will note that cut-elimination is widely regarded as one of the most important features of the sequent calculus. A considerable amount of research in proof theory concerns the consequences of cut-elimination, the extension of cut-elimination to new contexts, and so on. It would be difficult to overstate the importance of cut-elimination in proof theory.

The reduction procedure

If one adds axioms (1)-(6) of PA to the sequent calculus, cut-elimination is retained. However, cut-elimination breaks down once the induction axioms are added.[17] That is, the sequent calculus with the axioms of PA in addition to the logical axioms does not enjoy cut-elimination. Nevertheless, Gentzen reduced the consistency of PA to that of a system with cut-elimination.

It is integral to the logical understanding of arithmetic that arithmetical formulas can be stratified according to their complexity. Roughly, formulas with few quantifier alternations are less complicated than those with many quantifier alternations. Gentzen discovered that one can also stratify *proofs* in PA according to their complexity. Proofs that appeal to the induction axioms are more complicated than those that don't. The proofs that do appeal

[15] That is, an implication $\Gamma \Rightarrow \Delta$ is provable in the sequent calculus if and only if the sentence $\bigwedge \Gamma \rightarrow \bigvee \Delta$ is a first-order validity.

[16] To prove a contradiction in the sequent calculus is to prove the empty sequent $\Rightarrow$. If there were a proof of the empty sequent with the subformula property, then all of the formulas in the proof would appear in the end sequent. However, there are no formulas in the end sequent, so there cannot be any proofs of the empty sequent with the subformula property.

[17] One has to be somewhat careful in how this is stated. If PA proves φ, then, by the cut-elimination theorem for first-order logic, there will be a cut-free proof from purely logical axioms of a sequent $\rho \Rightarrow \varphi$, where ρ is some finite conjunction of axioms of PA. This is different from saying that there is a cut-free proof that uses axioms of PA as initial sequents and results in the sequent $\Rightarrow \varphi$. The typical treatment of PA in the sequent calculus, with axioms (1)-(6) as initial sequents and induction as a rule, results in a loss of cut-elimination.

to the induction axioms can then be stratified according to the complexity of the formulas that appear in their induction axioms. So Gentzen used the hierarchy of formulas to stratify proofs into a hierarchy as well.

The ensuing hierarchy is transfinitely tall; in fact, it is much taller than the hierarchy of natural numbers. Thus, Gentzen assigned *ordinal numbers*—which extend the natural numbers beyond the finite—to proofs in PA to measure their complexity. The greater the complexity of a proof, the larger the ordinal assigned to it. Gentzen then demonstrated that if there were a proof of a contradiction in PA, one would be able to transform it into a proof of a contradiction in PA with a *lower ordinal rank*. Let's call this the *reduction procedure*. If there were a proof of a contradiction in PA, then that proof would have a certain ordinal α assigned to it. By Gentzen's result, we could transform it into a proof of a contradiction with a lower ordinal rank. We could then transform that into a proof of a contradiction with a still lower ordinal rank and so on, until we reach a proof of a contradiction with ordinal rank 0. But there are no proofs of contradictions of ordinal rank 0, since proofs of rank 0 enjoy cut-elimination.[18] Thus, there can be no proofs of contradictions in PA at all.

Let's take stock. Gentzen provided an algorithm to reduce any proof of a contradiction to a proof of a contradiction with ordinal rank 0. Why doesn't this qualify as finitistic? The problem is that to *prove* that Gentzen's algorithm terminates, one must employ an inference rule known as *transfinite induction along* ε_0, where ε_0 is the upper bound of the ordinals that Gentzen assigns to proofs in PA.[19] This inference rule is not obviously finitistically acceptable. Nevertheless, some have regarded this as a very small broadening of the finitistic perspective and thus argued that Gentzen's proof vindicates a version of Hilbert's program. I will take up this claim in the next section.

7.3 The constructive Hilbert program

A number of foundational programs emerged from the ashes of Hilbert's program. I will discuss one such program: the *constructive Hilbert program*. The practitioners of the constructive Hilbert program aim to prove the consistency of axiomatic theories using only constructive principles. Constructive principles are those that can be justified without appealing to the law of excluded middle and concern only those mathematical objects that have been "constructed" in intuition.[20] Typically, constructive consistency proofs appeal only to finitistic principles except for an application of transfinite induction along a constructive

[18]To be more precise, one can transform any proof of ordinal rank 0 into a proof in which cut is applied only to those formulas provided by initial sequents. I want to emphasize that this is really an application of the method of cut-elimination though, and that it is sufficient to guarantee consistency.

[19]It is not exactly correct to say that one must assume *full* transfinite induction along ε_0 to prove the consistency of arithmetic. One must assume only a restricted version of this rule. I will discuss more precisely the resources required to prove the consistency of arithmetic in §4.2.2.

[20]There is no widely accepted, mathematically precise definition of "constructive." I will return to this issue in §3.2, where I will discuss some of the most comprehensive systems that have been considered constructively acceptable.

transfinite ordinal.[21] Bernays [16] articulated the spirit of the program in the following oft-quoted passage.

> The finitary standpoint is not the only alternative to classical ways of reasoning and is not necessarily implied by the idea of proof theory.[22] An enlarging of the methods of proof theory was therefore suggested: instead of reduction to finitist methods of reasoning it was required only that the arguments be of a constructive character.

Proof theorists in this tradition have discovered consistency proofs of many axiomatic theories. Such results cannot be criticized for lack of ingenuity or rigor. It is controversial, however, what philosophical significance these results have. For instance, do the principles used in constructive consistency proofs enjoy any special epistemic status? Should constructive consistency proofs quiet reasonable doubts about a theory's consistency? As Arana [2] writes, the "philosophical interest of these projects rests largely on the extent to which 'constructive' methods are judged more valuable than non-constructive methods." In §3.1, I will revisit considerations for and against the notion that the methods used in constructive consistency proofs are uniquely evidentially secure. In §3.2, I will discuss the methodological limitations of the constructive Hilbert program.

It is worth stating explicitly that I am neither rejecting nor endorsing the constructive Hilbert program. My aim is to argue that Gentzen-style consistency proofs enjoy a certain epistemic virtue, not to argue that they don't enjoy others. Nevertheless, it will be helpful, by means of contrast, to revisit considerations in favor of and against the constructive Hilbert program.

Epistemology of the constructive Hilbert program

There is a traditional argument for the evidential security of constructive principles, descended from Hilbert's [44] arguments for the evidential security of finitistic principles. Hilbert understood finitary reasoning as reasoning that operates on concrete[23] objects presented in intuition—that is, pure Kantian intuition[24]—and that does not appeal to abstract

[21] The constructive ordinals are the order-types of recursive relations on the natural numbers.

[22] Bernays was reacting to the Gödel–Gentzen theorem that the consistency of Heyting Arithmetic, i.e., intuitionistic first-order arithmetic, implies the consistency of Peano Arithmetic. Though this demonstrated one path for convincing constructivists of the consistency of classical mathematics, Bernays was also sympathetic to Gentzen-style consistency proofs as a constructive path to executing Hilbert's program, and this passage is often cited in that context.

[23] I am not using "concrete" here in the metaphysician's sense, so concreteness does not exclude abstractness. Rather, the concreteness of these objects consists in their primitiveness and sensibility in intuition. For discussion of Hilbert's finitist epistemology see [49] or [110].

[24] Hilbert's conception of intuition may have evolved throughout his career. In particular, though Hilbert eventually embraced a Kantian conception of intuition, Mancosu has argued that Hilbert originally supported finitism on the basis of an empiricist conception of intuition. See [63].

concepts concerning infinite sets. Hilbert argued that finitary reasoning was so basic and unimpeachable that it was necessary for all logical and scientific thought. Hilbert believed that a finitary consistency proof of a theory T would confer the same unimpeachable epistemic status upon T.[25] After proving the consistency of arithmetic by finitary methods and transfinite induction along ε_0, Gentzen [35] argued that transfinite induction up to ε_0 is constructively justified. Takeuti [107] argued that constructive reasoning operates on objects in intuition and enjoys something like the special epistemic status Hilbert ascribed to finitary reasoning. The combined arguments of Gentzen and Takeuti imply that transfinite induction up to ε_0 enjoys the type of special evidential security Hilbert praised.

Skepticism of Kantian intuition in the philosophy of mathematics has a long history, and I hardly need to contribute to it here. Nevertheless, objections to the use of Kantian intuition to justify the principles used in consistency proofs suggest themselves immediately. Bolzano and Frege famously criticized Kantian intuition on the grounds that we cannot practically intuit arbitrarily large numbers. In the context of ordinal analysis the question is even more pressing; can we really intuit—whether in Kant's sense or in any other sense—transfinite ordinals? Moreover, Hilbert insisted that consistency proofs require intuitions of indeterminate objects. Can we really have intuitions of indeterminate objects?[26] Perhaps there are other arguments to the effect that constructive principles enjoy a privileged epistemic status, but such arguments have not, to my knowledge, been brought to bear on the epistemology of consistency proofs.

It is worth noting that consistency proofs for axiomatic theories stronger than PA require principles that are increasingly difficult to consider constructive. There are two sources of this difficulty. First, as the transfinite ordinals used in the consistency proofs become larger and more remote from ordinary mathematical practice, the proofs become less obviously constructive. Perhaps, the objection goes, transfinite induction along ε_0 is constructively justified but transfinite induction along larger ordinals is not. Second, consistency proofs for strong theories have involved the use of heavy set-theoretic machinery. For instance, Rathjen's [79] consistency proof of Π^1_2-CA_0 made use of ordinal notation systems that were defined using large cardinals, which are quintessentially non-constructive objects.[27] Rathjen

[25]Hilbert did not identify any axiomatic theory with finitary reasoning. According to an influential conceptual analysis due to Tait [103], finitary reasoning coincides with Primitive Recursive Arithmetic, or PRA. It is worth noting that Tait's analysis has been challenged. Another influential argument, due to Kreisel, equates finitary with visualizable. Kreisel argues on this basis that finitary reasoning coincides with PA. See [53]. By Gödel's theorem, PRA is insufficient for establishing the consistency of any theory containing PRA. However, Gentzen's result demonstrates that it is sometimes possible to prove consistency statements by augmenting PRA with transfinite induction along a constructive ordinal. PRA is often vast overkill for accomplishing this task. EA^+, i.e., elementary arithmetic augmented with the totality of superexponentiation, suffices for the consistency proof of PA.

[26]The problem is this. Consistency proofs require induction on a Π_1 property, namely, consistency. To avoid finitistic meaninglessness, Hilbert suggested that one induct instead on an open formula. But working with formulas with free variables is tantamount to working with an "indeterminate object" in intuition. For discussion, see [73] and [42].

[27]Π^1_2-CA_0 is the strongest subsystem of analysis whose consistency has been proved by means of ordinal

argued that the use of large cardinals was heuristic and eliminable from these proofs in principle, but the mathematical community trusts Rathjen's proofs because of their faith in the cogency of the non-constructive principles he used.[28] As Feferman [27] put it, "it is not at all clear what they contribute to an extended [Hilbert's program] in the sense envisioned by Bernays."

Methodological limitations

Whatever the epistemic virtues of the constructive Hilbert program, there are reasons to believe that a constructive consistency proof of analysis is not forthcoming. To circumscribe the limits of constructivism, one would have to precisely articulate a conception of constructively acceptable mathematics. Martin-Löf has articulated and formalized a framework of this sort, his *constructive type theory*. Constructive type theory is perhaps the most general framework for constructively acceptable mathematics. With a precise framework in place, one can assess its limitations. Martin-Löf [67] gave the following assessment.

> What we see in front of us at this stage is some kind of abyss, or chasm, which we do not seem able to pass... The original aim was to obtain a constructive consistency proof for classical analysis...but we have now so much information that we know that this is out of our reach, and why? Well, if this is to be a constructive consistency proof, it will have to use constructively acceptable principles...we have at present exhausted the principles for which we can claim evidence.

Martin-Löf claimed that Π^1_2-CA_0 is on the other side of the abyss. Rathjen [81] argued for the weaker claim that currently known constructively acceptable principles cannot secure the consistency of theories stronger than Δ^1_3-CA_0. In any case, Rathjen's consistency proof of Π^1_2-CA_0 already appeals to controversial principles,[29] and the details of the proof are widely regarded as unusually cumbersome. Indeed, Rathjen [80] claimed that his proof "taxes the limits of human tolerance." Given these difficulties, it is reasonable to expect that constructively acceptable consistency proofs will not advance much further.[30]

It is a great testament to the ingenuity and perseverance of the proof theory community that they have pushed constructively acceptable consistency proofs to their very limits. Nevertheless, it seems that constructive consistency proofs have hit a wall.

---

analysis.

[28] Feferman makes a similar point in [27].

[29] Martin-Löf argued that these principles were already on the other side of the abyss. In particular, Martin-Löf argued that the use of large cardinals in the consistency proofs was evidence of postulating what was necessary to complete the proof rather than building objects from below.

[30] It is worth noting that constructive type theory is supposed to be an open-ended framework. Nevertheless, to extend the framework in a constructively acceptable way, any extension must be accompanied with an argument certifying the extension's constructivist credentials.

A number of proof theorists, including Girard [38] and Feferman [27], have argued that the constructive Hilbert program was wrongheaded and epistemically dubious. In fact, many of the proof theorists involved in ordinal analysis emphasize other benefits of ordinal analysis, such as its utility in measuring the strength of theories.[31] There is nothing wrong with using ordinal analysis to measure strength, but it is too philosophically modest to claim that this exhausts its significance. Moreover, it would be premature to dismiss Gentzen-style consistency proofs as philosophically insignificant because of the difficulties facing the constructive Hilbert program. In the next section I will articulate what I take the philosophical significance of consistency proofs to be.

7.4 Consistency proofs as explanations

In this section I will articulate and defend a conception of consistency proofs that is not motivated by and does not seek to address skeptical considerations. In this sense, as mentioned in §1, I am pursuing a project similar to that of Nozick [70] with respect to skepticism about knowledge. I will assume that we are justified in believing that arithmetic is consistent; indeed, we have a very simple semantic consistency proof that arithmetic is consistent. I do not expect that Gentzen's proof will increase anyone's credence that PA is consistent but I will argue that it has enhanced our understanding of why PA is consistent.

In §4.1, I will provide a survey of the leading theories of mathematical explanation and discuss the heuristics that they yield for identifying explanatory proofs. In §4.2, I will argue that Gentzen's proof is explanatory. Rather than argue for this claim on the basis of any one theory of explanation, I will argue that Gentzen's meets the conditions for explanatoriness identified by all the leading theories of mathematical explanation.

Mathematical explanation

There is no consensus among philosophers as to what constitutes an explanation. Some philosophers [58, 86] have endorsed theories of explanation that rely crucially on counterfactual dependence or causation. Such theories cannot easily account for mathematical explanation. Mathematical facts are necessarily true, so it is difficult to explain them by appealing to counterfactual dependence. Moreover, mathematical objects are causally inert, so mathematical facts cannot be explained by appealing to the causal properties of mathematical objects. One could deny that there are explanations within mathematics at all, but a cursory glance at mathematical practice reveals that mathematicians place a great deal of emphasis on the relative explanatory weight of proofs, so these views are not consonant with mathematical practice.[32] In any case, I will assume that there are explanations within

[31] Others still emphasize the computational information engendered by proof theory [52, 4].

[32] Mancosu demonstrates this with many examples from mathematical practice. See [64]. I do not believe that this is a knock-down argument in favor of the cogency of mathematical explanation. Nevertheless, it suggests that the notion merits serious consideration.

mathematics.

Some of the most prominent figures in the philosophy of mathematics have argued that some but not all proofs are explanatory. These philosophers argue that all proofs certify the truth of the statement proved, but only some do so on the basis of those facts in virtue of which it holds. Many of these philosophers have argued that the explanans, i.e., the facts in virtue of which the proved statement holds, need not be more obvious or certain than the proved statement itself. For instance, Russell [84] argued that the basic laws of arithmetic should be established by recourse to less evident but more primitive logical laws, and the logicist program was supposed to elucidate this. I believe that the situation with consistency proofs is analogous. Though the consistency of arithmetic is evident, and perhaps more evident than the hypotheses Gentzen invoked to prove it, the order of explanation proceeds from these hypotheses to consistency.

One could argue that Gentzen's proof is explanatory by adopting a specific theory of mathematical explanation and then checking that Gentzen's proof conforms to its strictures. Opinions have not converged on any specific account of mathematical explanation, however, so any argument that is tied to a specific theory is bound to be controversial. It would be more dialectically secure to argue that *whichever* of the leading theories theories of mathematical explanation one adopts, Gentzen's proof conforms to its strictures. Ultimately, this is what I will attempt.

Theories of mathematical explanation can be partitioned into *local* and *holistic* theories (Mancosu proposes this classification is his survey article [65]). According to local theories, explanatoriness is a property of individual arguments; according to holistic theories, explanatoriness is a property primarily of theoretical frameworks and only derivatively of individual arguments. I will briefly review the most popular versions of these theories of mathematical explanation: Steiner's local theory and Kitcher's holistic theory. According to Mancosu [64], "there are mainly two philosophical accounts of mathematical explanation available," namely, Steiner's and Kitcher's. While reviewing these theories, I will highlight the heuristics they yield for identifying explanatory proofs. In §4.2, I will argue that Gentzen's proof is explanatory according to Steiner's theory. Then in §4.3, I will argue that Gentzen's proof is explanatory according to Kitcher's theory. Finally, in §4.4, I will argue that these considerations suggest that Gentzen's proof should be explanatory according to *any* reasonable theory of mathematical explanation.

Steiner's theory of mathematical explanation

The quintessential local theory of mathematical explanation is Steiner's [102]. Steiner's theory is guided by the maxim "that to explain the behaviour of an entity, one deduces the behavior from" those properties that characterize the entity. Steiner identifies characterizing properties as those that are "unique to a given entity or structure within a family or domain of such entities or structures." According to Steiner, an explanatory proof of a claim φ must exhibit *how* φ depends on properties that characterize an entity mentioned in φ.

How can we *recognize* explanatory proofs? Steiner suggests the following test:

> It must be evident ... that if we substitute in the proof a different object of the same domain, the theorem collapses; more, we should be able to see as we vary the object how the theorem changes in response.

This provides a heuristic for recognizing whether a proof is explanatory in Steiner's sense. To see if a proof explains why a particular entity a has a property, we first identify a family $\mathcal{F}$ to which a belongs such that a is uniquely characterized among the entities in $\mathcal{F}$ by enjoying a certain property ψ mentioned in the proof. Second, we substitute different elements of the family $\mathcal{F}$ into the proof, and we see that the proof breaks down precisely where the property ψ is invoked. Finally, we check whether the same proof form yields analogous theorems about other entities in the family $\mathcal{F}$ when we shift our attention to their characterizing properties.

Kitcher's theory of mathematical explanation

The quintessential holistic theory of mathematical explanation is Kitcher's [50]. Kitcher provides the following description of his own view:

> Successful explanations earn that title because they belong to a set of explanations, the explanatory store ... Intuitively, the explanatory store associated with science at a particular time contains those derivations which collectively provide the best systematization of our beliefs. Science supplies us with explanations whose worth cannot be appreciated by considering them one-by-one but only by seeing how they form part of a systematic picture of the order of nature.

According to Kitcher, *systematic* accounts of nature render nature more intelligible, and arguments are explanatory only insofar as they contribute to the systematization of our beliefs. Accordingly, Kitcher identifies theoretical unification as the hallmark of explanation. He emphasizes that theoretical unification reduces the number of argument patterns required for comprehensive understanding of reality:

> Science advances our understanding of nature by showing us how to derive descriptions of many phenomena, using the same patterns of derivation again and again.

Thus, according to Kitcher, to recognize the explanatoriness of a proof it suffices to recognize that the proof pattern generalizes fruitfully within the context of a systematic and comprehensive theoretical framework.

Continuing ahead

I will argue that, whether we follow Steiner or Kitcher, we ought to regard Gentzen's proof as explanatory. Note that, despite their profound differences, the two most prominent theories of mathematical explanation both emphasize that *conduciveness to generalization* is

a characteristic mark of explanatory proofs. Indeed, the conduciveness of good explanations to generalization is widely noted in the philosophical literature on both mathematical and non-mathematical explanation. Gentzen's proof, as I will demonstrate in the following subsections, generalizes along several dimensions. I will first discuss these generalizations in a way that emphasizes that they depend on the characterizing properties of the entities involved, in accord with Steiner's theory. I will then discuss these generalizations in a way that emphasizes how they have facilitated theoretical unification, in accord with Kitcher's theory.

Gentzen's proof and Steiner's theory of explanation

Gentzen's proof is an argument for the consistency of PA. There are two dimensions along which the proof could conceivably be generalized:

(i) to concern theories other than PA;

(ii) to concern properties other than consistency.

I will argue that Gentzen's proof generalizes along both dimensions. Along the way I will identify a family to which PA belongs and a family to which the property *consistency* belongs. I also identify properties that characterize PA and the property *consistency* within their respective families. When arguing that Gentzen's proof generalizes along a dimension, I will pay attention to the characterizing properties of *other* entities in the family, and show how substituting those properties into the proof yields analogous results about those entities.

Along the way, I will contrast Gentzen's proof with the Tarskian proof. Insofar as the Tarskian proof generalizes along either dimension, it does not generalize by shifting attention to the characterizing properties of other entities in the family. Indeed, whenever the Tarskian proof goes through, consistency (or an analogous property) is derived from the same property, namely, soundness; that is, the Tarskian proof exhibits no sensitivity to the theory being analyzed or the property being established. Moreover, the Tarskian proof does *not* generalize as widely as Gentzen's proof, since there are consistent but unsound theories.

Generalizing to other theories

PA is not the only theory that has been subject to a Gentzen-style consistency proof; similar analyses have been carried out for a wide array of systems. The methods introduced by Gentzen, namely cut-elimination and the reduction procedure, play a central role in each such analysis. Nevertheless, generalizing Gentzen's methods beyond PA requires innovation. When one approaches a new system T, one has to determine *how* to stratify T so that the reduction procedure can be carried out. Ordinal analysis is *inter alia* how the appropriate stratifications are developed.

Following Steiner, it will be useful to regard PA as a member of a family. PA is a *reasonable theory*, i.e., a consistent, effectively axiomatized theory that interprets a modicum

of arithmetic. Standard proof-theoretic questions[33] can be intelligibly asked only about reasonable theories. Since I am currently focusing on such questions, I will regard PA's family as comprising the reasonable theories. There is a stratification of PA's proofs of order-type ε_0, and, insofar as proof-theoretic questions are concerned, PA is characterized by this stratification.

For any reasonable T, finding a similar stratification of T involves calculating the proof-theoretic ordinal of T. Once this is accomplished, one can develop an analogous consistency proof for T. These consistency proofs will be different from Gentzen's insofar as they involve induction along smaller or larger ordinals. I claim that for any reasonable theory T, T is characterized among reasonable theories by its stratification that facilitates a Gentzen-style consistency proof. Put another way, for any reasonable theory T, T is characterized among reasonable theories by its proof-theoretic ordinal.[34] I take it that this perspective is continuous with the perspective of working proof-theorists, such as Pohlers [75], who claims that by calculating the proof-theoretic ordinal of a theory T, we thereby gain a *total* understanding of T's proof-theoretic properties.

To check that Steiner's criteria for explanatoriness are met, we must check that (i) when we substitute different elements of PA's family into Gentzen's proof, the proof breaks down precisely where the PA's characterizing property is invoked and (ii) the same proof form yields analogous theorems about other entities in PA's family when we shift our attention to their characterizing properties.

Regarding (i), note that Gentzen's proof is insufficient for proving the consistency of theories stronger than PA, *precisely* because these theories have different proof-theoretic ordinals. Indeed, if the proof-theoretic ordinal of a theory T is some $\alpha > \varepsilon_0$, then induction on ε_0 is insufficient for proving the consistency of T.[35] Indeed, if we stratify T into some hierarchy of height ε_0, then the crucial reduction procedure cannot be carried out.

Regarding (ii), I reiterate that Gentzen-style consistency proofs have been developed for many theories, including theories with proof-theoretic ordinals other than ε_0. By calculating the proof-theoretic ordinal of a theory T, we may thereby discover a stratification of T that facilitates a Gentzen-style consistency proof of T. Thus, Gentzen's proof form yields analogous theorems about other axiomatic theories when we shift our attention to their proof-theoretic ordinals.

---

[33]Proof theory is a broad area, but here I am referring to those questions concerning consistency strength, ordinal analysis, reflection principles, provably recursive functions, and the like.

[34]Throughout, when I speak of proof-theoretic ordinals, I really mean canonical representations of proof-theoretic ordinals. Discovering the latter is effectively equivalent to discovering a stratification of T that facilitates a Gentzen-style consistency proof.

[35]Indeed, each $\beta < \alpha$, T will prove transfinite induction along α. The claim then follows by Gödel's second incompleteness theorem. It is worth pointing out that all of these results are sensitive to manner in which ordinals are represented. Nevertheless, extensive investigation has provided ample empirical evidence for the following two claims. First, for any *natural* presentation of α, induction along α will suffice to prove the consistency of T. Second, for any *natural* presentation of an ordinal $\beta < \alpha$, induction along β will *not* suffice to prove the consistency of arithmetic.

The Tarskian consistency proof is notably not conducive to generalization in this manner. If we consider any member T of PA's family that is *not* sound, then the Tarskian proof cannot be carried out. Insofar as the Tarskian proof *does* generalize, it does so only by invoking soundness, a property which is too coarse to characterize any reasonable theory with respect to questions of proof-theoretic interest. Indeed, a Tarskian consistency proof of a theory T, when it can be carried out, can be carried out no matter what T's proof-theoretic ordinal is; thus, the Tarskian proof a theory's consistency does not depend on the theory's characteristic features.

Generalizing to other properties

Gentzen's proof is conducive to generalization even in the case of PA, since it generalizes to prove that PA enjoys other properties of interest. In §2.3 I mentioned that it is integral to the logical understanding of arithmetic that arithmetical formulas can be stratified into a hierarchy based on their complexity. Roughly, statements with fewer quantifier alternations are less complicated than statements with many quantifier alternations. This engenders the following hierarchy of formulas:

$$\Pi_1 \subset \Pi_2 \subset ... \subset \Pi_n \subset ...$$

The Π_n formulas are roughly those formulas that have n many quantifier alternations, starting with a universal quantifier.[36]

The statement Con(PA) that expresses the consistency of arithmetic is the quintessential Π_1 sentence that is not provable in arithmetic. Indeed, the sentence says "for *all* numbers x, x does not encode a proof of '0=1' from the axioms of arithmetic." Con(PA) is often called a *reflection principle* for PA, which is to say that it expresses that PA meets some minimal standards of correctness. Other reflection principles for PA have been widely studied. For instance, we say that PA is n-*consistent* if PA is consistent with all true Π_n sentences. This yields the following hierarchy of statements:

$$\mathsf{Con}(\mathsf{PA}), 1\mathsf{Con}(\mathsf{PA}), 2\mathsf{Con}(\mathsf{PA}), ..., n\mathsf{Con}(\mathsf{PA}), ...$$

where for each n, nCon(PA) expresses the n-consistency of PA. For present purposes, it is important to note that for each n, nCon(PA) is a Π_{n+1} sentence. We may regard the claims

$$\{\mathsf{Con}(\mathsf{PA}), 1\mathsf{Con}(\mathsf{PA}), 2\mathsf{Con}(\mathsf{PA}), ..., n\mathsf{Con}(\mathsf{PA}), ...\}$$

as forming a family of entities each of which is distinguished from the others by its logical complexity (e.g., being Π_1, Π_2, etc.).

[36]Here is a more precise definition. We assume that all formulas are in prenex normal form, i.e., all quantifiers appear in the beginning of the formula. A formula with only bounded quantifiers is both Σ_0 and Π_0. If φ is Π_n and ψ is Σ_n, then $\exists x\varphi$ is Σ_{n+1} and $\forall x\psi$ is Π_{n+1}. We can extend the definition of these classes so that Π_n includes all formulas that are EA provably equivalent to a Π_n formula, in the old sense. This delivers the inclusions as described above.

Steiner's heuristic suggests that if Gentzen's proof is explanatory, then it should be possible to prove that PA is n-consistent by modifying Gentzen's proof by invoking a characteristic property of n-consistency instead of the characteristic property of consistency. Indeed, this is the case. Gentzen proved the consistency of PA by finitary methods and an application of a certain inference rule: ε_0 induction for Π_1 predicates. The same argument, slightly modified so that we instead employ ε_0 induction for Π_{n+1} predicates, yields a proof that PA is n-consistent.[37]

Steiner's heuristic also suggests that it should be clear that the proofs of n-consistency must *depend* on the characteristic property of n-consistency; if we substitute in the characteristic property of some other reflection principle, then the proof should break down. To see that Gentzen's proof and its generalizations truly depend on the precise ε_0 induction rule that is applied, it will be useful to discuss the *reversals* of these results. To prove the n-consistency of arithmetic by Gentzen's methods, we use finitary techniques plus an application of ε_0 induction to a Π_{n+1} formula. This is the best possible result, by the following theorem (note that by a *non-nested* application of ε_0 I just mean an application of ε_0 induction to a formula which was not itself derived by ε_0 induction).[38]

Theorem 7.4.1 (Beklemishev)**.** *Any statement that is provable by finitary methods and non-nested applications of ε_0 induction to Π_{k+1} predicates can also be proved by finitary methods from the assumption that* PA *is k-consistent.*[39]

Thus, ε_0 induction for Π_{n+1} predicates is both *necessary* and *sufficient* for proving the n-consistency of arithmetic.[40] In particular, one *cannot* prove the n-consistency of arithmetic by applying induction to any Π_n predicates. So Gentzen's proof is couched at a level of grain that emphasizes the importance of the quantifier complexity, i.e., the characteristic property, of the claims being proved.

Before continuing, it is worth noting that the Tarskian proof is not articulated at a level of grain that is conducive to results of this sort. Certainly one can prove the n-consistency of arithmetic by appealing to the soundness of arithmetic. However, one cannot "reverse" the Tarskian proof so as to prove the soundness of arithmetic from the consistency of arithmetic. The assumption of soundness is not necessary for proving consistency; accordingly, the Tarskian proof of consistency does not depend in any way on consistency's characteristic property.

---

[37] For details see [8].

[38] Beklemishev essentially proves this theorem in [8], where he works over the base theory PRA. He weakens the base theory to EA^+ in [9]. EA^+ is stated in the language of arithmetic augmented with a primitive symbol for exponentiation. EA^+ has the usual recursive axioms for addition, multiplication, and exponentiation, as well as an axiom stating the totality of super-exponentiation. EA^+ is the weakest canonical theory in which various meta-mathematical results, such as the cut-elimination theorem and its corollaries, can be proven and hence in which various meta-mathematical arguments can be carried out. EA^+ is widely regarded as finitistically acceptable.

[39] Here's a more precise statement. Over EA^+, k-Con(PA) is deductively equivalent to the inference rule for non-nested applications of ε_0 induction for Π_{k+1} predicates.

[40] Once again, this means necessary and sufficient modulo finitary methods.

Gentzen's proof and Kitcher's theory of explanation

The preceding remarks demonstrate that Gentzen's proof pattern generalizes. I have emphasized along the way that these generalizations meet Steiner's criteria for explanatoriness. Have these same considerations demonstrated that Gentzen's proof is explanatory in Kitcher's sense?

Recall that, according to Kitcher [50], explanatory frameworks are those in which we can "derive descriptions of many phenomena, using the same patterns of derivation again and again" and that strike a balance "between minimizing the number of patterns of derivation employed and maximizing the number of conclusions generated." Gentzen's proof is arguably the seminal result of proof theory, and we have already seen that Gentzen's proof generalizes to derive other claims within the framework of proof theory. Therefore, the preceding considerations provide evidence that Gentzen's proof is explanatory in Kitcher's sense.

However, there is more evidence that Gentzen's proof is explanatory in Kitcher's sense. Generalizations of Gentzen's argument have played important roles in systematizing our understanding of other logical aspects, not yet discussed, of arithmetic and its kin. A Gentzen-style consistency proof of a theory T often leads *inter alia* to (i) a characterization of the functions[41] that are provably total in T, (ii) a characterization of the transfinite induction principles that are provable in T, and (iii) the discovery of combinatorial principles that are true but not provable in T. Moreover, systematic connections have been discovered between Gentzen style consistency proofs and other streams of research within mathematical logic, including the study of functional interpretations [45], indicators [5], and modal logics of provability [8]. This is all to say that Gentzen's proof draws our attention to "connections, common patterns, in what initially appeared to be different situations," as Kitcher [50] claims explanations do.

It is worth noting that Tarskian consistency proofs did not and *could not* facilitate these discoveries. Indeed, Tarskian consistency proofs appeal essentially to the soundness of a theory, and that information is too coarse-grained to deliver these results, which are sensitive to fine-grained information. For instance, (i) above requires knowledge of the Π_2 consequences of a theory, but these cannot be encoded into the soundness claim. Knowing that a theory is sound does not yield any information about the theory's Π_2 consequences.[42] By contrast, one of the modifications of Gentzen's proof discussed in the previous subsection does deliver this result.[43]

[41] Of course, here I mean recursive functions on the natural numbers.

[42] None of this is to say that *semantic techniques* are inappropriate for studying consistency. Indeed, one can establish these corollaries using semantic techniques. However, one can do so only by examining *non-standard models*, since all total recursive functions are total in the standard model, all ordinals are well-ordered in the standard model, and all true combinatorial principles hold in the standard model.

[43] One might wonder whether a generalization of the Tarskian proof might deliver this information, but it is difficult to see how such a generalization could go. One might try to extract this information by, say, attempting to prove the Π_2 soundness of PA by induction on the length of proofs, in the Tarskian style. However, the induction step of such an argument would not work if the inference considered increases the quantifier complexity of the formulas considered. To get around this subtlety, one would invoke the

Gentzen's proof and other theories of mathematical explanation

Before discussing the normative consequences of the preceding arguments, it is worth considering other theories of mathematical explanation. As stated in §7.4, Mancosu [64] wrote in 2001 that Steiner's and Kitcher's are effectively the "two philosophical accounts of mathematical explanation available." In the intervening years there has been an expansion in the theories of explanation available, though Steiner's and Kitcher's remain two of the most influential contenders. Is Gentzen's proof explanatory according to the other theories?

Even if we do not follow Steiner or Kitcher, there remains strong evidence that Gentzen's proof is explanatory. The extant theories of mathematical explanation all predict that explanatory proofs tend to enjoy certain qualities, which I will call *marks of explanation*. Indeed, theories of mathematical explanation, old and new, have identified the following as candidate marks of explanation: conduciveness to generalization [50, 102], methodological purity [22, 23], facilitation of theoretical unification [32, 50], tracking of dependency relations [102, 48, 74], and exploitation of symmetry [57]. It is remarkable that Gentzen's proof bears *each* of these marks. Perhaps some new theory of mathematical explanation will identify some mark of explanation that Gentzen's proof lacks, but at present we lack evidence that there is any such mark.

The standard theories of mathematical explanation do not, as far as I know, propose any necessary conditions for explanatoriness that do not apply to Gentzen's proof. As we saw in the previous section, some of them (including the two leading theories) propose sufficient conditions for explanatoriness that *do* apply to Gentzen's proof. Moreover, whenever any of the extant theories of mathematical explanation predicts that some marks accompany explanations, Gentzen's proof bears these marks. Therefore, the claim that Gentzen's proof is explanatory falls within the *overlapping consensus* of the extant theories of mathematical explanation. Accordingly, one could argue for the explanatoriness of Gentzen's proof abductively, without embracing any particular theory of mathematical explanation: Whichever theory of mathematical explanation is correct, Gentzen's proof bears the marks that the theory predicts accompany explanatory proofs; the best explanation of this empirical observation is that Gentzen's proof *is* explanatory.

7.5 Normative consequences

In this section I will discuss the normative upshot of the preceding analysis of Gentzen's proof. In particular, I will argue against the widely held view that consistency proofs are

cut-elimination theorem. However, there remains a problem: the induction axioms of PA have unbounded syntactic complexity, so one must use the reduction procedure to get the cut-elimination theorem to do the required work. By attempting to modify the Tarskian proof, we converge on a Gentzen-style proof. The failure of the Tarskian proof to generalize here follows from a general consideration. The Tarskian proof is an ordinary induction (not transfinite induction) on the length of proofs. Any ordinary induction with an arithmetical induction formula can be carried out in PA. Thus, no such argument can yield results about what is not provable in PA, since such results imply the consistency of PA.

epistemically interesting only if they use the limited methods of constructive mathematics. I will then discuss a line of research that my position suggests.

Restricted methodology

A prevalent view among logicians is that, for any axiomatic theory T, mathematical techniques that are less evidentially secure than those codified in T should not appear in consistency proofs of T. Already such restrictions were articulated by Gentzen [37]:

> In a consistency proof we can use only forms of inference that count as considerably more secure than the forms of inference of the theory whose consistency is to be proved ... In carrying out this consistency proof we must make certain...that the inferences and derived concepts used in the proof *itself* are *indisputable* or at least considerably more reliable than the doubtful forms of inference of elementary number theory.

Which concepts and inferences are "considerably more reliable" than those of arithmetic? Some have argued that consistency proofs should use only the limited means of constructive mathematics. For instance, Zach [112] remarks that for an assignment of an ordinal α_{T} to a theory T "to have any foundational significance it is also required that one can give a constructive argument for transfinite induction up to α_{T}." Likewise, Rathjen [83] writes that "consistency proofs should be given relative to 'constructive' theories."

If one wants to convince skeptics, then perhaps one ought to restrict one's methods to those accepted by the skeptic. However, foundational significance should not be identified with propensity to convince skeptics. Are the methodological constraints identified by Gentzen, Rathjen, and Zach appropriate constraints on the task of explaining *why* axiomatic theories are consistent? If one avails oneself of arbitrary resources when proving a consistency statement, then one might reach for the Tarskian proof, which, I have argued, is not explanatory. Thus, it is reasonable to abide by some methodological constraints when trying to prove consistency. However, just because *some* methodological constraints are appropriate it does not follow that using non-constructivist resources is inappropriate. Without the motivation of addressing constructivists, it is not clear that this constraint is justified; defending the constraint would require a substantial argument.

Which principles can be appealed to in explanatory consistency proofs? It seems hopelessly difficult to delimit, with formally precise criteria, the principles to which one can appeal in explanatory consistency proofs. On the other hand, the epistemic virtues that I described in §7.4 can provide evidence for the correctness of a methodology. That is, if a particular approach to proving consistency can be systematically generalized, facilitates theoretical unification, and so on, this is evidence of the explanatory power of that approach.

Incomparable systems

A related methodological constraint on consistency proofs arises from the second incompleteness theorem. From Gödel's theorem, some philosophers have inferred that a consistency proof of a theory T can be philosophically interesting only if, like Gentzen's, it goes through in a theory that is *deductively incomparable* with T.[44] What is the source of this constraint? Smith [95] introduces the constraint as follows:

> Trying to prove the consistency of PA by appeal to a *stronger* theory which already contains PA might well not seem to be a good strategy if we want to quiet doubts about PA's consistency ... And the Second Theorem shows that it is impossible to prove PA's consistency by appeal to a *weaker* theory which is contained inside PA...there's another possibility: maybe we can prove PA's consistency by appeal to an attractive theory which is weaker than PA in some respects but stronger in others.

Smith is explicit that the goal of quieting doubts about consistency motivates this constraint. I do not believe that the constraint holds up when this motivation is abandoned. Of course, some consistency proofs given in strictly stronger theories are not interesting. Indeed, the Tarskian proof is given in a strictly stronger theory, and it lacks many of the epistemic virtues of Gentzen's proof. However, it would be rash to infer from this example that no such consistency proof is explanatory. To illustrate what an informative consistency proof given in a strictly stronger theory might look like, I will consider a specific example: Tait's resolution of Takeuti's conjecture.

Takeuti [106] proved, using only finitistic means, that if second-order logic enjoys cut-elimination, then analysis is consistent.[45] Takeuti then conjectured that second-order logic enjoys cut-elimination. Tait [104] proved that second-order logic enjoys cut-elimination, resolving Takeuti's conjecture.[46] Tait's proof used highly non-constructive techniques. Indeed, Tait's proof invokes a very strong form of the assumption that analysis is consistent.[47]

[44]The thought is this. Neither one of $\mathsf{PRA} + \Pi_1\text{-TI}_{\varepsilon_0}$ or PA is contained in the other. Only the former proves Con(PA) and only the latter proves any of the canonical induction axioms $I\Sigma_{n+1}$.

[45]By "second-order logic" I mean the standard sequent proof system for second-order logic with full comprehension. For details, see [107], where Takeuti calls the system LK^+.

[46]Indeed, Takeuti conjectured that full simple type theory enjoys cut-elimination. This was known as Takeuti's Fundamental Conjecture. After Tait partially resolved the conjecture, Takahashi and Prawitz independently resolved the full conjecture in the positive; see [105] and [78]. Both proofs used non-constructive semantic techniques. It is worth mentioning that cut-elimination for simple type theory also follows from Girard's strong normalization theorem for the system F. See [39].

[47]Specifically, Tait relied on the assumption that every relation on the naturals is an element of a countable ω model of second-order arithmetic, where an ω model of second-order arithmetic is a model of second-order arithmetic whose first-order domain is the set $\mathbb{N}$ of natural numbers and whose second order domain is some subset of $\mathcal{P}(\mathbb{N})$. The statement that analysis is consistent is equivalent to the statement that the successor relation is contained in a countable ω model of second-order arithmetic. It is important to remember that Tait's assumption, though very strong, is still true.

Proving the consistency of analysis by assuming a strong form of the consistency of analysis might seem circular. Rathjen [82] comments that though Tait's theorem is a "very interesting result...as far as instilling trust in the consistency of [analysis]...is concerned, the...proofs are just circular since they blatantly use the very comprehension principles formalized in [analysis] (and a bit more)." I agree with Rathjen; however, if one seeks explanation instead of justification, then the charge of circularity is not as strong. Takeuti's result—that cut-elimination for second-order logic implies the consistency of analysis—shows that the consistency of analysis can be understood in terms of cut-elimination, as long as second-order logic *does* enjoy cut-elimination. This is already an impressive achievement. Experience in proof theory suggests that consistency can often be fruitfully understood in terms of cut-elimination; verifying that this is so in the case of analysis brings us closer to integrating the consistency of analysis into the current theoretical landscape. Tait's proof, though it relies on the assumption that analysis is consistent, nevertheless certifies that second-order logic enjoys cut-elimination. The combined work of Takeuti and Tait shows that the consistency of analysis follows from cut-elimination for a canonical, natural proof calculus. Given the central role of cut-elimination in proof theory, it is difficult to resist the conclusion that these results have brought us closer to an explanation of the consistency of analysis.

Tait's proof was carried out in a theory strictly stronger than analysis, but the objection that Tait assumes the consistency of analysis misses the point. The combined work of Takeuti and Tait won't convince any skeptics, but that does not imply that their work will not contribute to a satisfying explanation of the consistency of analysis.

Towards an explanation of the consistency of analysis

None of this is to say that our understanding of the consistency of analysis is satisfactory. Tait's result, though non-trivial and interesting, does not bear all of the marks of explanation I mentioned in §7.4. Perhaps the most glaring issue is this. There are other noteworthy consistency proofs of analysis, but the relationship between these proofs is not understood. One such consistency proof comes from Spector's [97] work extending Gödel's *Dialectica* interpretation by the bar recursion principle. Another comes from Girard's [39] strong normalization theorem for the system F. In the case of first-order arithmetic, there is a highly developed and compelling network of theorems explaining the consistency of arithmetic. The connections between different consistency proofs—including those using functional interpretations and those using cut-elimination—are well-understood. Thus, different methods converge on the same explanation, and the results reinforce each other with respect to the correctness of this explanation. This is not yet the situation for analysis.[48] Thus, it seems that it would

[48]It is not understood, for example, how to show that every provably recursive function of analysis is represented by a bar-recursive term by appealing to cut-elimination for second-order logic. Of course, some connections are understood, but they are not so illuminating. To prove the consistency of fragments of second-order arithmetic with Γ-comprehension, one can appeal to Tait's cut-elimination argument, which will use Γ comprehension and slightly more. Similarly, Spector's functional interpretation uses ever higher

be illuminating to develop—unbeholden to constructive or finitistic scruples—connections between these three proofs. I expect that the discovery of such connections would greatly enhance our understanding of the consistency of analysis.

bar recursive functionals to match increases in the complexity of comprehension. So there is some connection between the complexity of bar recursive functionals and the complexity of the amount of comprehension used to eliminate cuts.

Bibliography

[1] Uri Andrews, Mingzhong Cai, David Diamondstone, Steffen Lempp, and Joseph S Miller. On the structure of the degrees of relative provability. *Israel Journal of Mathematics*, 207(1):449–478, 2015.

[2] Andrew Arana. Proof Theory in Philosophy of Mathematics. *Philosophy Compass*, 5(4):336–347, 2010.

[3] Aristotle. *Posterior Analytics*, volume 1. Harvard University Press, 1960.

[4] Jeremy Avigad. The computational content of classical arithmetic. *arXiv preprint arXiv:0901.2551*, 2009.

[5] Jeremy Avigad and Richard Sommer. A model-theoretic approach to ordinal analysis. *Bulletin of Symbolic Logic*, 3(01):17–52, 1997.

[6] Lev Beklemishev. Iterated local reflection versus iterated consistency. *Annals of Pure and Applied Logic*, 75(1-2):25–48, 1995.

[7] Lev Beklemishev. Proof-theoretic analysis by iterated reflection. *Archive for Mathematical Logic*, 42(6):515–552, 2003.

[8] Lev Beklemishev. Provability algebras and proof-theoretic ordinals, I. *Annals of Pure and Applied Logic*, 128(1-3):103–123, 2004.

[9] Lev Beklemishev. Reflection principles and provability algebras in formal arithmetic. *Russian Mathematical Surveys*, 60(2):197, 2005.

[10] Lev Beklemishev. Veblen hierarchy in the context of provability algebras. In P. Hájek, L. Valdés-Villanueva, and D. Westerståhl, editors, *Logic, Methodology and Philosophy of Science, Proceedings of the Twelfth International Congress*, pages 65–78. Kings College Publications, London, 2005. Preprint: Logic Group Preprint Series 232, Utrecht University, June 2004.

[11] Lev Beklemishev. Calibrating provability logic: From modal logic to reflection calculus. *Advances in modal logic*, 9:89–94, 2012.

[12] Lev Beklemishev. Positive provability logic for uniform reflection principles. *Annals of Pure and Applied Logic*, 165(1):82–105, 2014.

[13] Lev Beklemishev. Reflection calculus and conservativity spectra. *Russian Mathematical Surveys*, 73(4):569–613, aug 2018.

[14] Lev Beklemishev and Fedor Pakhomov. Reflection algebras and conservation results for theories of iterated truth, 2019. arXiv:1908.10302.

[15] Lev D Beklemishev. Provability logics for natural turing progressions of arithmetical theories. *Studia Logica*, 50(1):107–128, 1991.

[16] Paul Bernays. David Hilbert. *Encyclopedia of Philosophy*, 3:496–505, 1967.

[17] Bernard Bolzano. Contributions to a better-grounded presentation of mathematics. *Ewald (1996)*, 1:174–224, 1810.

[18] Bernard Bolzano. *The Theory of Science: Die Wissenschaftslehre oder Versuch einer Neuen Darstellung der Logik*. Univ of California Press, 1972.

[19] Timothy Y Chow. The consistency of arithmetic. *The Mathematical Intelligencer*, 41(1):22–30, 2019.

[20] Evgenij Dashkov. On the positive fragment of the polymodal provability logic GLP. *Mathematical Notes*, 91(3-4):318–333, 2012.

[21] Michael Detlefsen. On interpreting Gödel's second theorem. *Journal of Philosophical Logic*, 8(1):297–313, 1979.

[22] Jean Dieudonné. *Linear algebra and geometry*. Hermann, 1969.

[23] Doug Hensley Edgar, Gerald A. and Douglas B. West. Problem 10830. *Edgar, Gerald A., Doug Hensley and Douglas B. West "" American Mathematical Monthly 109/4 (): 396–397*, 109(4):396–397, 2002.

[24] Ali Enayat and Fedor Pakhomov. Truth, disjunction, and induction. *Archive for Mathematical Logic*, 58(5):753–766, Aug 2019.

[25] Herbert B Enderton and Hilary Putnam. A note on the hyperarithmetical hierarchy. *The Journal of Symbolic Logic*, 35(3):429–430, 1970.

[26] Solomon Feferman. Transfinite recursive progressions of axiomatic theories. *The Journal of symbolic logic*, 27(3):259–316, 1962.

[27] Solomon Feferman. *What rests on what? The proof-theoretic analysis of mathematics*. 1992.

[28] Solomon Feferman and Clifford Spector. Incompleteness along paths in progressions of theories. *The Journal of Symbolic Logic*, 27(4):383–390, 1962.

[29] David Fernández-Duque. Worms and spiders: Reflection calculi and ordinal notation systems, 2016. arXiv:1605.08867.

[30] David Fernez-Duque. Impredicative consistency and reflection, 2015.

[31] Harvey Friedman. Uniformly defined descending sequences of degrees. *The Journal of Symbolic Logic*, 41(2):363–367, 1976.

[32] Michael Friedman. Explanation and scientific understanding. *The Journal of Philosophy*, 71(1):5–19, 1974.

[33] Sy-David Friedman, Michael Rathjen, and Andreas Weiermann. Slow consistency. *Annals of Pure and Applied Logic*, 164(3):382–393, 2013.

[34] Emanuele Frittaion. Uniform reflection in second order arithmetic. `https://drive.google.com/file/d/19-25_Gr5wGE6beQD5ho_k52slivagTjU/view` (accessed 31.01.2020), 2019.

[35] Gerhard Gentzen. Die gegenwärtige Lage in der mathematischen Grundlagenforschung. Neue Fassung des Widerspruchsfreiheitsbeweises für die reine Zahlentheorie. *Bull. Amer. Math. Soc. 45 (1939), 812-813*, pages 0002–9904, 1939.

[36] Gerhard Gentzen. Investigations into logical deduction. *American Philosophical Quarterly*, 1(4):288–306, 1964.

[37] Gerhard Gentzen. The Consistency of Elementary Number Theory. *The Collected Papers of Gerhard Gentzen*, 1969.

[38] J.-Y. Girard. *Proof Theory and Logical Complexity*. Bibliopolis, 1987.

[39] Jean-Yves Girard. *Interprétation fonctionelle et élimination des coupures de l'arithmétique d'ordre supérieur*. PhD thesis, Université Paris VII, 1972.

[40] Kurt Gödel. Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für mathematik und physik*, 38(1):173–198, 1931.

[41] Petr Hájek and Pavel Pudlák. *Metamathematics of first-order arithmetic*, volume 3. Cambridge University Press, 2017.

[42] Bob Hale. The problem of mathematical objects. *Foundational Theories of Classical and Constructive Mathematics*, 2011.

[43] Joseph Harrison. Recursive pseudo-well-orderings. *Transactions of the American Mathematical Society*, 131(2):526–543, 1968.

[44] David Hilbert. Neubegründung der Mathematik. Erste Mitteilung. In *Gesammelte Abhandlungen*, pages 157–177. Springer, 1970.

[45] William Alvin Howard. Assignment of ordinals to terms for primitive recursive functionals of finite type. *Studies in Logic and the Foundations of Mathematics*, 60:443–458, 1970.

[46] Joost J Joosten. Turing–Taylor expansions for arithmetic theories. *Studia Logica*, 104(6):1225–1243, 2016.

[47] Juliette Kennedy. Did the Incompleteness Theorems Refute Hilbert's Program. *Stanford Encyclopedia of Philosophy*, 2015.

[48] Jaegwon Kim. Explanatory knowledge and metaphysical dependence. *Philosophical Issues*, 5:51–69, 1994.

[49] Philip Kitcher. Hilbert's epistemology. *Philosophy of Science*, pages 99–115, 1976.

[50] Philip Kitcher. Explanatory unification and the causal structure of the world. *Scientific explanation*, 13:410–505, 1989.

[51] Peter Koellner. Independence and large cardinals. *Stanford Encyclopedia of Philosophy*, 2010.

[52] Georg Kreisel. Mathematical significance of consistency proofs. *The Journal of Symbolic Logic*, 23(2):155–182, 1958.

[53] Georg Kreisel et al. Ordinal logics and the characterization of informal concepts of proof. In *Proceedings International Congress of Mathematicians*, pages 14–21, 1958.

[54] Georg Kreisel and Azriel Lévy. Reflection principles and their use for establishing the complexity of axiomatic systems. *Mathematical Logic Quarterly*, 14(7-12):97–142, 1968.

[55] Georg Kreisel, J Shoenfield, and Hao Wang. Number theoretic concepts and recursive well-orderings. *Archiv für mathematische Logik und Grundlagenforschung*, 5(1-2):42–64, 1960.

[56] S Kripke. The Collapse of the Hilbert Program: Why a System Cannot Prove its Own 1-consistency. *The Bulletin of Symbolic Logic*, 15(2):229–230, 2009.

[57] Marc Lange. Aspects of mathematical explanation: Symmetry, unity, and salience. *Philosophical Review*, 123(4):485–531, 2014.

[58] David Lewis. Causation. *The Journal of Philosophy*, 70(17):556–567, 1974.

[59] Per Lindström. *Aspects of incompleteness*, volume 10. Cambridge University Press, 2017.

[60] Patrick Lutz and James Walsh. Incompleteness and jump hierarchies, 2019. arXiv:1909.10603.

[61] Penelope Maddy. Believing the axioms I. *The Journal of Symbolic Logic*, 53(2):481–511, 1988.

[62] Penelope Maddy. Believing the axioms II. *The Journal of Symbolic Logic*, 53(3):736–764, 1988.

[63] Paolo Mancosu. *From Brouwer to Hilbert*. Oxford, 1997.

[64] Paolo Mancosu. Mathematical explanation: Problems and prospects. *Topoi*, 20(1):97–117, 2001.

[65] Paolo Mancosu. Explanation in mathematics. *Stanford Encyclopedia of Philosophy*, 2008.

[66] Alberto Marcone and Antonio Montalbán. The Veblen functions for computability theorists. *The Journal of Symbolic Logic*, 76(2):575–602, 2011.

[67] Per Martin-Löf. The Hilbert-Brouwer controversy resolved? In *One Hundres Years of Intuitionism (1907–2007)*, pages 243–256. Springer, 2008.

[68] Antonio Montalbán and James Walsh. On the inevitability of the consistency operator. *The Journal of Symbolic Logic*, 84(1):205–225, 2019.

[69] Carl Mummert and Stephen G Simpson. An incompleteness theorem for β_n-models. *The Journal of Symbolic Logic*, 69(2):612–616, 2004.

[70] Robert Nozick. *Philosophical Explanations*. Harvard University Press, 1981.

[71] Fedor Pakhomov and James Walsh. Reflection ranks and ordinal analysis, 2018. arXiv:1805.02095v1.

[72] Fedor Pakhomov and James Walsh. Reflection ranks and ordinal analysis. *arXiv preprint arXiv:1805.02095*, 2018.

[73] Charles Parsons. Finitism and intuitive knowledge. *The Philosophy of Mathematics Today*, 1998.

[74] Christopher Pincock. The unsolvability of the quintic: A case study in abstract mathematical explanation. *Philosophers' Imprint*, 15(3):1–19, 2015.

[75] Wolfram Pohlers. A short course in ordinal analysis. *Proof Theory (Leeds, 1990)*, pages 27–78, 1992.

[76] Wolfram Pohlers. *Proof Theory: The First Step Into Impredicativity*. Springer Science, 2008.

[77] Marian Boykan Pour-El and Saul Kripke. Deduction-preserving "recursive isomorphisms" between theories. *Fundamenta Mathematicae*, 61:141–163, 1967.

[78] Dag Prawitz. Hauptsatz for higher order logic. *The Journal of Symbolic Logic*, 33(03):452–457, 1968.

[79] Michael Rathjen. Recent Advances in Ordinal Analysis: Π^1_2-CA and Related Systems. *Bulletin of Symbolic Logic*, 1(04):468–485, 1995.

[80] Michael Rathjen. The realm of ordinal analysis. In *Sets and Proofs*, pages 219–279. Cambridge University Press, 1999.

[81] Michael Rathjen. The constructive Hilbert program and the limits of Martin-Löf type theory. In *Logicism, Intuitionism, and Formalism*, pages 397–433. Springer, 2009.

[82] Michael Rathjen. Proof theory: From arithmetic to set theory. *Annals of Pure and Applied Logic*, 165:563–572, 2014.

[83] Michael Rathjen and Wilfried Sieg. Proof Theory. *Stanford Encyclopedia of Philosophy*, 2018.

[84] Bertrand Russell. The regressive method of discovering the premises of mathematics. *Essays in analysis*, pages 272–83, 1973.

[85] Gerald E Sacks. *Degrees of Unsolvability*. Princeton University Press, 1963.

[86] Wesley Salmon. *Scientific Explanation and the Causal Structure of the World*. 1984.

[87] Ulf R. Schmerl. A fine structure generated by reflection formulas over primitive recursive arithmetic. *Studies in Logic and the Foundations of Mathematics*, 97:335–350, 1979.

[88] Ulf R. Schmerl. Iterated reflection principles and the ω-rule. *The Journal of Symbolic Logic*, 47(4):721–733, 1982.

[89] Stewart Shapiro. Proof and truth: Through thick and thin. *The Journal of Philosophy*, 95(10):493–521, 1998.

[90] V Yu Shavrukov and Albert Visser. Uniform density in Lindenbaum algebras. *Notre Dame Journal of Formal Logic*, 55(4):569–582, 2014.

[91] Joseph R Shoenfield. On a restricted ω-rule. *The Journal of Symbolic Logic*, 1969.

[92] Stephen G. Simpson. *Subsystems of Second Order Arithmetic*, volume 1. Cambridge University Press, 2009.

[93] Stephen G. Simpson and Rick L. Smith. Factorization of polynomials and Σ10 induction. *Annals of Pure and Applied Logic*, 31(2):289–306, 1986.

[94] Theodore A Slaman and John R Steel. Definable functions on degrees. In *Cabal Seminar 81–85*, pages 37–55. Springer, 1988.

[95] Peter Smith. *An Introduction to Gödel's Theorems*. Cambridge University Press, 2013.

[96] Craig Smorynski. *Self-Reference and Modal Logic*. Springer Science, 2012.

[97] Clifford Spector. Provably recursive functionals of analysis: a consistency proof of analysis by an extension of principles formulated in current intuitionistic mathematics. *Recursive function theory*, pages 1–27, 1962.

[98] John Steel. Descending sequences of degrees. *The Journal of Symbolic Logic*, 40(1):59–61, 1975.

[99] John R Steel. Forcing with tagged trees. 1978.

[100] John R Steel. A classification of jump operators. *The Journal of Symbolic Logic*, 47(2):347–358, 1982.

[101] John R Steel. Gödel's program. *Interpreting Gödel: Critical Essays*, pages 153–79, 2014.

[102] Mark Steiner. Mathematical explanation. *Philosophical Studies*, 34(2):135–151, 1978.

[103] William W Tait. Finitism. *The Journal of Philosophy*, 78(9):524–546, 1981.

[104] William W Tait et al. A non constructive proof of Gentzen's Hauptsatz for second order predicate logic. *Bulletin of the American Mathematical Society*, 72(6):980–83, 1966.

[105] Moto-o Takahashi. A proof of cut-elimination theorem in simple type-theory. *Journal of the Mathematical Society of Japan*, 19(4):399–410, 1967.

[106] Gaisi Takeuti. On a generalized logic calculus. In *Japanese journal of mathematics: transactions and abstracts*, volume 23, pages 39–96. The Mathematical Society of Japan, 1953.

[107] Gaisi Takeuti. *Proof Theory*, volume 81. Courier Corporation, 2013.

[108] Alfred Tarski. The concept of truth in formalized languages. *Logic, semantics, metamathematics*, 2:152–278, 1956.

[109] Alan Mathison Turing. Systems of logic based on ordinals. *Proceedings of the London mathematical society*, 2(1):161–228, 1939.

[110] Richard Zach. Hilbert's program. *Stanford Encyclopedia of Philosophy*, 2003.

[111] Richard Zach. The practice of finitism: Epsilon calculus and consistency proofs in Hilbert's program. *Synthese*, 137(1):211–259, 2003.

[112] Richard Zach. Hilbert's program then and now. *Philosophy of Logic*, 5:411–447, 2006.