Tightness and solidity in fragments of Peano Arithmetic

Piotr Gruza[*], Leszek Aleksander Kołodziejczyk[†] and Mateusz Łełyk[‡]

December 12, 2025

Abstract

It was shown by Visser that Peano Arithmetic has the property that any two bi-interpretable extensions of it (in the same language) are equivalent. Enayat proposed to refer to this property of a theory as *tightness* and to carry out a more systematic study of tightness and its stronger variants that he called neatness and solidity.

Enayat proved that not only PA, but also ZF and Z_2 are solid. On the other hand, it was shown in later work by a number of authors that many natural proper fragments of those theories are not even tight.

Enayat asked whether there is a proper solid subtheory of the theories listed above. We answer that question in the case of PA by proving that for every n, there exist both a solid theory and a tight but not neat theory strictly between $I\Sigma_n$ and PA. Moreover, the solid subtheories of PA can be required to be unable to interpret PA. We also obtain some other separations between properties related to tightness, for example by giving an example of a sequential theory that is neat but not semantically tight in the sense of Freire and Hamkins.

1 Introduction

Our aim in this paper is to show that a potential very general characterization of Peano Arithmetic (PA) as an axiomatic theory does not work.

To understand the background behind the potential characterization, recall the notion of interpretation (precise definitions of all relevant concepts will be provided in Section 2). Intuitively speaking, a structure $\mathcal{M}$ interprets a structure $\mathcal{N}$ if the universe, relations and operations of $\mathcal{N}$ can be defined in $\mathcal{M}$, where the universe of $\mathcal{N}$ can consist of tuples of elements of $\mathcal{M}$ rather than single elements, and equality in $\mathcal{N}$ can be an equivalence relation on $\mathcal{M}$ other than equality. Well-known examples include the interpretation of the field of rationals in the ring of integers, with rationals given as pairs of integers $\langle k, \ell \rangle$, where $\ell \neq 0$ and $\langle k, \ell \rangle$ is identified with $\langle p, q \rangle$ if $kq = p\ell$; and the interpretation of the field of complex numbers in the field of reals, with $a + bi$ given as the pair of reals $\langle a, b \rangle$.

An interpretation of an axiomatic theory T in a theory S is essentially a uniform recipe for interpreting a model of T in a model of S. A pair of interpretations, of T in S and of S in T, forms a bi-interpretation if the interpretations are provably mutually inverse, in the sense that each theory proves that composing the interpretations in the appropriate order gives rise to a structure that is isomorphic to the original model of

[*]University of Warsaw, Doctoral School of Natural and Exact Sciences, p.gruza3@uw.edu.pl
[†]University of Warsaw, Institute of Mathematics, lak@mimuw.edu.pl
[‡]University of Warsaw, Faculty of Philosophy, mlelyk@uw.edu.pl

the theory. The concept of bi-interpretability was originally introduced in model theory (see [1]) but plays an increasingly meaningful role in foundational investigations (cf. e.g. [24, 10]). Bi-interpretability between theories is much stronger than mutual interpretability: a well-known example is provided by the theories ZF and ZFC + GCH, which are mutually interpretable but not bi-interpretable. In contrast, PA is bi-interpretable with an appropriate formulation of finite set theory.

In general, bi-interpretability and provability do not go hand in hand: for example one easily finds examples of theories that are bi-interpretable and yet mutually inconsistent. However, as shown by Visser, [27], there are incomplete first-order theories which make bi-interpretability collapse to logical equivalence on the class of their extensions. In particular, Visser [27] proved that PA has the following curious property, later called *tightness*: if two extensions T_1 and T_2 of PA, still in the same language, are bi-interpretable, then in fact $T_1 \equiv T_2$. Note that every complete theory will be tight in this sense. In fact, tightness is a kind of internal completeness property: a complete theory is one whose models are all elementarily equivalent, while if a theory T is tight and a pair of interpretations I and J gives a bi-interpretation of T with itself, then T can prove that the structures described by I and by J are elementarily equivalent to one another, and in fact to the original model of T in which the interpretations were applied.

Enayat [4] initiated a more systematic study of tightness and related concepts. In particular, he introduced semantical variants of tightness in which one considers interpretations between models of T rather than between theories extending T. The strongest property that he considers, called *solidity*, requires that any two models $\mathcal{M}, \mathcal{N} \vDash T$ have to be definably isomorphic as soon as they satisfy a weaker form of bi-interpretability (essentially, one of the two interpretations between $\mathcal{M}$ and $\mathcal{N}$ is only assumed to be a one-sided rather than two-sided inverse of the other). Thus, solidity is an internalized categoricity property rather than mere internalized completeness.

Enayat showed that PA is not just tight but also solid, and so are other foundationally important axiom schemes such as ZF set theory and second-order arithmetic Z_2. On the other hand, it was gradually realized that natural proper fragments of those theories are not even tight. In particular:

(a) neither Zermelo set theory Z nor ZFC^- (i.e. ZFC without Power Set and with collection instead of replacement) is tight [8];

(b) for each n, the fragment Π^1_n- CA of Z_2 is not tight [9];

(c) for each n, no Π_n-axiomatized fragment of true arithmetic (thus, *a fortiori*, of PA) is tight [6].

Freire and Hamkins [8] noted that the proofs of tightness and solidity for set theory "seem to use the full strength of ZF". Similarly, Freire and Williams [9] referred to their results as "evidence that tightness characterizes Z_2 (...) in a minimal way". This situtation gave salience to a question asked already by Enayat [4, Question 3.2]: do any any of PA, ZF, Z_2 have a proper solid subtheory?

In the context of set theory, it was stated in [9] that settling Enayat's question (presumably in the negative) would amount to "a profound characterization of ZF". *Mutatis mutandis*, this would be even more true in the case of PA, due to both the very basic nature of first-order arithmetic and to the special role played by the induction scheme in tightness arguments: it was already observed in [4, just before Question 3.1] that all solid theories known up to that point that interpret a minimal amount of arithmetic also imply the full induction scheme under that interpretation.

In our view, the "right" question to ask is not quite whether the theories mentioned above have *arbitrary* solid subtheories. Taken literally, that question is vulnerable to trivial counterexamples: for instance, it is not difficult to show that "either the axioms of ZF hold, or the universe has one element and $\in$ is the empty relation" is a solid theory. Rather, one should ask whether there are solid proper subtheories that imply some reasonably strong axioms giving the scheme at hand its appropriate (arithmetical or set-theoretic) character: say $\mathrm{I}\Sigma_1$ or $\mathrm{I}\Delta_0 + \exp$ in the case of PA and Zermelo set theory in the case of ZF. (One not quite trivial solid subtheory of ZF that still seems to "leave out too much" is ZF without infinity but with the axiom that every set has a transitive closure; see [6, Theorem 18].)

Here, we show that even after such an arguably natural modification, the answer to Enayat's Question 3.2 for PA is positive; as a consequence, we show that there can be arithmetical solid theories that do not imply the full induction scheme. More precisely, we prove that for every n there is a solid theory strictly between $\mathrm{I}\Sigma_n$ and PA. Our examples have a disjunctive nature like the trivial one above, but their construction is considerably more involved: essentially, they state "either PA holds, or we are in a particular pointwise definable model of $\mathrm{I}\Sigma_n + \neg\,\mathrm{I}\Sigma_{n+1}$". To make this work, we have to ensure both that the theory of those pointwise definable models is solid, and that it has a well-calibrated interpretability strength sufficiently different from (in fact: greater than) that of PA.

The upshot of our result is that there is no hope of characterizing PA as a minimal solid theory, at least in the realm of theories ordered by logical implication. In fact, we are able to show that the same holds true for the coarser (pre-)order of interpretability as well. Nevertheless, our work points to a more subtle sense in which it remains open whether PA could be minimal solid; we discuss this briefly in Section 7 of the paper.

Another very natural question related to tightness, solidity and their cousins is whether these concepts are actually distinct. Again, a naive version of this question admits some relatively trivial positive answers, due to the fact that syntactically defined properties like tightness apply to all complete theories, whereas semantically defined ones like solidity in general do not. However, separating two syntactically defined tightness-like properties, or getting any separating example at all that would be a computably axiomatized theory subject to Gödel's theorems, was a significant challenge. We provide the first separations "of the nontrivial kind", showing for example that there are arbitrarily strong proper subtheories of PA that distinguish tightness from a property that is likewise syntactically defined but has the bi-interpretability assumption weakened as in the case of solidity.

To prove our results, we need constructions that ensure the existence of some specific interpretations, isomorphisms etc. but not of others. For such purposes, we rely on a wide variety of methods from the model theory and proof theory of arithmetic. The tools we make use of include, among other things: axiomatic truth theories, flexible formulas, pointwise definable models, and a very weak pigeonhole principle known as the cardinality scheme.

The remainder of the paper is organized as follows. We review basic background concepts and facts in Section 2. In Section 3, we discuss and develop some more advanced background material. In Section 4, we prove our results on solid proper subtheories of PA. In Section 5, we prove separations between tightness, solidity, and other similar properties. Section 6 initiates the study of proper solid subtheories of Z_2, by providing an example containing ACA_0. We summarize our work and state some open problems in Section 7.

2 Preliminaries

A few general conventions: to avoid irrelevant complications, all languages considered in this paper are finite. If a theory T is fixed or clear from the context, then $\mathcal{L}_T$ denotes the language of T. Given a formula $W(x)$, we may sometimes write $x \in W$ instead of $W(x)$, mainly in order to be able to substitute $\forall x \in W \ldots$ for the more cumbersome $\forall x\,(x \in W \to \ldots)$.

2.1 Interpretability

We expect that the reader has at least an intuitive understanding of what interpretations are in logic. As mentioned in the introduction, an interpretation of a structure $\mathcal{N}$ in a structure $\mathcal{M}$ is roughly a definition of the domain and relations of $\mathcal{N}$ inside $\mathcal{M}$, while an interpretation of a theory T in a theory S is a uniform recipe for interpreting models of T in models of S. However, since we are going to prove theorems about interpretations and interpretability, we need to be somewhat precise, and thus we provide a more formal discussion of these concepts. Those formal details can be rather boring, so the reader may consider just skimming the present subsection at first and referring back to it as needed.

For the purpose of giving an official definition of interpretations, we pretend that all languages are purely relational. There is no harm in doing so, thanks to the usual transformation of arbitrary languages into relational ones that replaces each n-ary function symbol with an $(n+1)$-ary relational symbol. It is known that this transformation applies not only to formulas, but also induces a (feasible) transformation of proofs in a theory T into proofs in its relational analogue T^{rel}; for details, see e.g. [26, Section 7.3].

Translations. The formal definition of interpretation starts with the notion of *translation*. If $\mathcal{L}_1$ and $\mathcal{L}_2$ are first-order languages, then a *translation* M from $\mathcal{L}_1$ to $\mathcal{L}_2$ is determined by specifying:

(i) a unary $\mathcal{L}_2$-formula $\delta_{\mathsf{M}}(x)$, sometimes called the *domain formula*;

(ii) for every n-ary relation symbol P of $\mathcal{L}_1$ (including equality), an $\mathcal{L}_2$-formula $P^{\mathsf{M}}(\bar{y})$ with exactly n free variables, such that $\vdash P^{\mathsf{M}}(y_1, \ldots, y_n) \to \bigwedge_{i \leq n} \delta_{\mathsf{M}}(y_i)$.

If Γ is a class of $\mathcal{L}_2$-formulas, then we say that the translation M is Γ-*restricted* if δ_{M} and all the formulas P^{M} belong to Γ.

Remark. The intention behind the definition of translation is that $\delta_{\mathcal{M}}$ defines a domain of $\mathcal{L}_1$-objects on which the formulas $P^{\mathcal{M}}$ define $\mathcal{L}_1$-relations. Our definition of translation (and the definition(s) of interpretation based on it, given below) is not the most general one possible: in particular, our translations are one-dimensional, in the sense that the domain formula always has just one free variable. For our purposes, this is inessential, because (almost) all the theories we consider support a pairing function. For an example (rather distant from our main topic) of a situation in which multi-dimensional interpretations would matter, see the Remark in Section 5.1.

The requirement that $P^{\mathsf{M}}(y_1, \ldots, y_n)$ logically imply $\delta_{\mathsf{M}}(y_i)$ for each i is a technical condition that is sometimes useful, and it can be assumed to hold without loss of generality in all contexts that will be relevant to us. So, we simplify things by including it in the definition of translation.

Given a translation M from $\mathcal{L}_1$ to $\mathcal{L}_2$, we define the translation φ^{M} of an $\mathcal{L}_1$-formula φ as follows: $(P(\bar{x}))^{\mathsf{M}}$ is $P^{\mathsf{M}}(\bar{x})$ for a relation symbol P of $\mathcal{L}_1$; the translation commutes with propositional connectives; and $(\forall x\, \psi)^{\mathsf{M}}$ is $\forall x\big(\delta_{\mathsf{M}}(x) \to \psi\big)$.

Interpretations of structures. If M is a translation of $\mathcal{L}_1$ into $\mathcal{L}_2$ and $\mathcal{N}$ is an $\mathcal{L}_2$-structure, then M is a *parameter-free interpretation in* $\mathcal{N}$ if $(\delta_{\mathsf{M}})^{\mathcal{N}}$ is nonempty and $(=^{\mathsf{M}})^{\mathcal{N}}$ is an equivalence relation on $(\delta_{\mathsf{M}})^{\mathcal{N}}$ that is a congruence w.r.t. each relation $(P^{\mathsf{M}})^{\mathcal{N}}$. In this case, M and $\mathcal{N}$ uniquely determine an $\mathcal{L}_1$-structure whose universe is the set of equivalence classes of $(=^{\mathsf{M}})^{\mathcal{N}}$ and whose relations are as determined by $(P^{\mathsf{M}})^{\mathcal{N}}$. We denote this structure by $\mathcal{N}^{\mathsf{M}}$, and we say that a model $\mathcal{M}$ is *interpreted without parameters in* $\mathcal{N}$ *via* M, if $\mathcal{M} = \mathcal{N}^{\mathsf{M}}$. In general, an *interpretation* M *in* $\mathcal{N}$ can be based on formulas δ_{M} and P^{M} that use parameters from $\mathcal{N}$: in other words, an interpretation in $\mathcal{N}$ is essentially the same thing as a parameter-free interpretation in $(\mathcal{N}, \bar{c})$ for some tuple of constants $\bar{c}$. We write $\mathsf{M} : \mathcal{N} \rhd \mathcal{M}$ to indicate that M is an interpretation of the structure $\mathcal{M}$ in $\mathcal{N}$ (so in particular $\mathcal{M} = \mathcal{N}^{\mathsf{M}}$). If the interpretation is clear from context or unimportant, we then omit the reference to it, writing simply $\mathcal{N} \rhd \mathcal{M}$ and saying that $\mathcal{M}$ is interpreted in $\mathcal{N}$. We say that a model $\mathcal{M}$ is *interpretable* (as opposed to "interpreted") in $\mathcal{N}$ is there is an interpretation M in $\mathcal{N}$ such that $\mathcal{M}$ is isomorphic to $\mathcal{N}^{\mathsf{M}}$.

Remark. One easily notices that $\mathcal{M}$ is interpretable (resp. parameter-free interpretable) in $\mathcal{N}$ if and only if there is a surjection from $\mathcal{N}$ onto $\mathcal{M}$ such that the preimage of every parameter-free $\mathcal{M}$-definable set is definable (resp. parameter-free definable) in $\mathcal{N}$. Hence our definition of interpretability is equivalent to the classical model-theoretic one, cf. e.g. [1, 14]. In our context it will be easier to work with the more syntactic approach presented above.

Remark. Note that if a translation $\mathsf{M} : \mathcal{L}_1 \to \mathcal{L}_2$ is identity preserving, which means that $x =^{\mathsf{M}} y$ is $x = y$, then M is an interpretation in $\mathcal{N}$ for every $\mathcal{L}_2$-structure $\mathcal{N}$ for which $(\delta_{\mathsf{M}})^{\mathcal{N}}$ is nonempty.

Interpretations of theories. Let M be a translation from $\mathcal{L}_1$ into $\mathcal{L}_2$, and for $i = 1, 2$ let T_i be an $\mathcal{L}_i$-theory. Then M is an *interpretation of* T_1 *in* T_2 if T_2 proves that $=^{\mathsf{M}}$ is an equivalence relation on δ_{M} that is a congruence w.r.t. each relation P^{M}, and T_2 also proves φ^{M} for each axiom φ of T_1 (including logical axioms, in particular the non-emptiness of the universe). We then say that T_2 *interprets* T_1 *via* M. Thus, an interpretation of T_1 in T_2 is given by a fixed set of formulas that provide a parameter-free interpretation of a model of T_1 in each $\mathcal{N} \vDash T_2$. We write $T_2 \rhd T_1$ (resp. $\mathsf{M} : T_2 \rhd T_1$) to indicate that T_2 interprets T_1 (resp. via M). We say that a translation M is an *interpretation in* T_2 if it is an interpretation of the empty theory over the appropriate language in T_2.

Composition and the identity interpretation. Interpretations between structures, or between theories, can be thought of as morphisms of a category, in that they can be composed and there is always an identity interpretation. For any language $\mathcal{L}$, the identity translation $\mathsf{id}_{\mathcal{L}}$ is given by letting $\delta_{\mathsf{id}_{\mathcal{L}}}$ be $x = x$ and translating each relation symbol of $\mathcal{L}$ to itself. For two translations $\mathsf{M}\colon \mathcal{L}_1 \to \mathcal{L}_2$ and N of $\mathcal{L}_2 \to \mathcal{L}_3$, we define $\mathsf{MN}\colon \mathcal{L}_1 \to \mathcal{L}_3$ (note the order in which we write the composition of interpretations) as follows:

- $\delta_{\mathsf{MN}} := \delta_{\mathsf{N}} \wedge (\delta_{\mathsf{M}})^{\mathsf{N}}$.
- $P^{\mathsf{MN}}(\bar{y}) := \bigwedge_{y \in \bar{y}} \delta_{\mathsf{N}}(y) \wedge (P^{\mathsf{M}}(\bar{y}))^{\mathsf{N}}$, for each $\mathcal{L}_1$-relation symbol P.

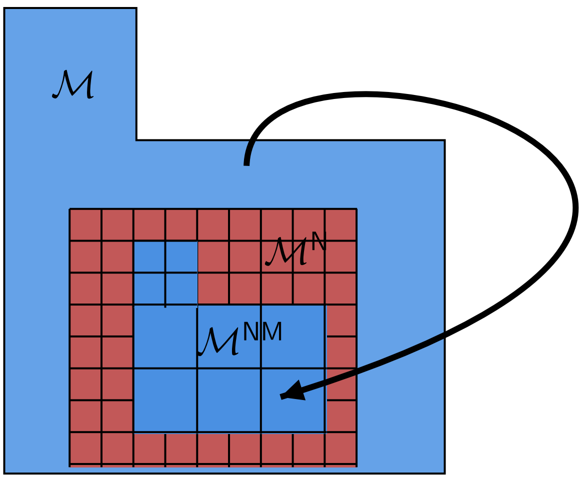
Then for every $\mathcal{L}_1$-formula φ, the formulas φ^{MN} and $(\varphi^{\mathsf{M}})^{\mathsf{N}}$ are logically equivalent. This is enough to define composition for interpretations in theories and for parameter-free interpretations in structures. If $\mathcal{M}$ is a model, N_1 is an interpretation with parameters in $\mathcal{M}$, and N_2 is an interpretation with parameters in $\mathcal{M}^{\mathsf{N}_1}$, then the composition $\mathsf{N}_1\mathsf{N}_2$ is also routinely defined, and it is *unique up to equivalence in* $\mathcal{M}$: since the parameters used by N_2 correspond to equivalence classes of the $\mathcal{M}$-definable equivalence relation $=^{\mathsf{N}_1}$, one should choose some representatives of these classes. Clearly any choice of these representatives is good, since $=^{\mathsf{N}_1}$ is a congruence w.r.t. all definable relations of $\mathcal{M}_1^{\mathsf{N}}$.

Isomorphism of interpretations. To define the crucial notion of bi-interpretation and its weaker version, retraction, we need to say what it means for interpretations to be isomorphic. Given a language $\mathcal{L}$ and two interpretations M_1, M_2 of $\mathcal{L}$-structures in a structure $\mathcal{N}$, we say that M_1, M_2 are $\mathcal{N}$*-isomorphic* (resp. *parameter-free* $\mathcal{N}$*-isomorphic*) if there is a definable (resp. parameter-free definable) relation $\iota \subseteq N^2$ such that the domain of ι is $(\delta_{\mathsf{M}_1})^{\mathcal{N}}$, the range is $(\delta_{\mathsf{M}_2})^{\mathcal{N}}$, and ι preserves all $\mathcal{L}$-predicates (including $=$). This means that ι canonically determines an isomorphism between $\mathcal{N}^{\mathsf{M}_1}$ and $\mathcal{N}^{\mathsf{M}_2}$. Note that for any interpretation K in $\mathcal{N}^{\mathsf{M}_1}$, the isomorphism ι gives rise to a corresponding interpretation $\iota[\mathsf{K}]$ given by the same formulas as K with parameters shifted by ι (so, if K involves no parameters, then $\iota[\mathsf{K}]$ is the same as K). One can define the notion of embedding between interpretations in a similar way.

Retractions and bi-interpretations of structures. Let $\mathcal{M}$ and $\mathcal{N}$ be two structures, let N be an interpretation in $\mathcal{M}$ and let M be an interpretation in $\mathcal{M}^{\mathsf{N}}$.

- We say that M and N form a *retraction in* $\mathcal{M}$ if NM is $\mathcal{M}$-isomorphic to the identity interpretation.
- We say that $\mathcal{M}$ is *a retract of* $\mathcal{N}$ if there is a retraction (N, M) in $\mathcal{M}$ such that $\mathcal{N}$ is isomorphic to $\mathcal{M}^{\mathsf{N}}$.
- We say that M and N form a *bi-interpretation in* $\mathcal{M}$ if NM is $\mathcal{M}$-isomorphic to the identity interpretation on $\mathcal{M}$ via isomorphism ι, and $\mathsf{M}\iota[\mathsf{N}]$ is $\mathcal{M}^{\mathsf{N}}$-isomorphic to the identity interpretation on $\mathcal{M}^{\mathsf{N}}$.
- We say that $\mathcal{M}$ and $\mathcal{N}$ are *bi-interpretable* if there is a bi-interpretation (N, M) in $\mathcal{M}$ such that $\mathcal{N}$ is isomorphic to $\mathcal{M}^{\mathsf{N}}$.

The picture below illustrates a retraction is illustrated in the picture below. The squares and rectangles comprising the interpreted structures $\mathcal{M}^{\mathsf{N}}$ and $\mathcal{M}^{\mathsf{NM}}$ correspond to the equivalence classes of $=^{\mathsf{N}}$ and $=^{\mathsf{NM}}$. The arrow indicates an $\mathcal{M}$-definable isomorphism between $\mathcal{M}$ and $\mathcal{M}^{\mathsf{NM}}$.



Remark. By an easy argument one can see that bi-interpretability is actually symmetric, which need not be directly obvious from the definition. Similarly, it is equivalent to the standard notion studied e.g. in [1].

Remark. We say that $\mathcal{M}$ and $\mathcal{N}$ are *parameter-free bi-interpretable* if the interpretations and isomorphism needed for the bi-interpretation are given by parameter-free formulas. It is reasonably straightforward to show that if $\mathcal{M}$ and $\mathcal{N}$ are parameter-free bi-interpretable, then the automorphism groups of $\mathcal{M}$ and $\mathcal{N}$ are isomorphic. Note that this is not true for general (parametric) bi-interpretability.

Retracts and bi-interpretability between theories. Unsurprisingly, the concepts of bi-interpretability and being a retract have their analogues for theories as well, expressing that the appropriate compositions of interpretations are isomorphic to the identity provably in the appropriate theories. Given two interpretations $\mathsf{M}_1, \mathsf{M}_2$ of a language $\mathcal{L}$ in a theory T, we say that M_1 and M_2 are *isomorphic in* T if there is a binary $\mathcal{L}_T$-formula $\iota(x,y)$ which T-provably determines an isomorphism between M_1 and M_2. That is, provably in T the relation defined by ι has δ_{M_1} as its domain, δ_{M_2} as range, and preserves all the relations of $\mathcal{L}$, including equality.

Let T_1, T_2 be theories in languages $\mathcal{L}_1, \mathcal{L}_2$, respectively.

- We say that T_1 is a *retract* of T_2 if there are translations $\mathsf{M}_1 : \mathcal{L}_1 \to \mathcal{L}_2$ and $\mathsf{M}_2 : \mathcal{L}_2 \to \mathcal{L}_1$ such that the theory T_1 interprets T_2 via M_2, the theory T_2 interprets T_1 via M_1, and $\mathsf{id}_{\mathcal{L}_1}$ is isomorphic to $\mathsf{M}_2\mathsf{M}_1$ in T_1.
- We say that T_1 and T_2 are *bi-interpretable* if there are translations $\mathsf{M}_1 : \mathcal{L}_1 \to \mathcal{L}_2$ and $\mathsf{M}_2 : \mathcal{L}_2 \to \mathcal{L}_1$ witnessing that T_1 is a retract of T_2 and T_2 is a retract of T_1.

2.2 Categoricity-like notions for first-order theories

Below we recall four categoricity- and completeness-like properties which emerged in the literature. The concepts of solidity, tightness and neatness were introduced in [4], while semantical tightness was considered for the first time in [8]. These notions can be seen to arise by making independent choices with respect to two independent questions:

1. Do we want a semantical or a syntactic property?
2. Do we want to use the notion of retraction or the notion of bi-interpretation?

Perhaps the most natural of the four properties is the one that was introduced last, semantical tightness. As the name suggests, this is a semantical property, and it is based on the notion of bi-interpretability between structures.

Definition 2.1 (Semantical tightness)**.** A theory T is *semantically tight* if whenever $\mathcal{M}$ is a model of T and (N, M) is a bi-interpretation in $\mathcal{M}$ such that $\mathcal{M}^{\mathsf{N}} \models T$, then N is $\mathcal{M}$-isomorphic to $\mathrm{id}_{\mathcal{M}}$ (and, as a consequence, M is $\mathcal{M}^{\mathsf{N}}$-isomorphic to $\mathrm{id}_{\mathcal{M}^{\mathsf{N}}}$).

Remark. We observe that the semantical tightness of a theory T entails that the bi-interpretability relation between models of T is trivial in the following sense: whenever $\mathcal{M}$ and $\mathcal{N}$ are models of T and $\mathcal{M}$ is bi-interpretable with $\mathcal{N}$, then $\mathcal{M}$ and $\mathcal{N}$ are isomorphic.

Remark. One can consider stronger and weaker notions of semantical tightness. For example, one could restrict the notion given above by insisting that the definition of the isomorphism between N and $\mathrm{id}_{\mathcal{M}}$ do not use any parameters other than the ones involved in defining the interpretations and isomorphisms that give rise to the bi-interpretation. In this paper, whenever we show the semantical tightness of a theory, it always holds in this more restrictive sense, and whenever we show failure of semantical tightness, it applies already in the weaker sense. So, our results do not depend on which of the two definitions was applied.

The original definition of semantical tightness in [8] is weaker still: the isomorphism between $\mathcal{M}$ and $\mathcal{N}$ need not be definable. We think that the definition proposed above is more in the spirit of the other notions considered in this paper (see the definition of solidity below). Moreover, our example of a theory which is neat but not semantically tight from Section 5.4 works also for definition used by [8]. For an example of a situation in which the distinction between isomorphism and definable isomorphism would matter, see the Remark in Section 5.1.

A stronger property, solidity, is also semantical but starts from notion of retraction.

Definition 2.2 (Solidity)**.** We say that T is *solid* if whenever $\mathcal{M}$ is a model of T and (N, M) is a retraction in $\mathcal{M}$ such that $\mathcal{M}^{\mathsf{N}} \models T$, then N is $\mathcal{M}$-isomorphic to $\mathrm{id}_{\mathcal{M}}$.

Remark. Analogously to the case of semantical tightness, we can observe that the solidity of a theory T trivializes the retraction relation between models of T: whenever $\mathcal{M}$ and $\mathcal{N}$ are models of T and $\mathcal{M}$ is a retract of $\mathcal{N}$, then $\mathcal{M}$ and $\mathcal{N}$ are isomorphic.

Remark. Our definition of solidity is equivalent to the original one given in [4]. However, the later paper [6] used the more restrictive definition according to which the definition of the isomorphism between N and $\mathsf{id}_{\mathcal{M}}$ can only use the parameters involved in defining N, M and the isomorphism between $\mathsf{id}_{\mathcal{M}}$ and NM. As in the case of semantical tightness, our main results do not depend on which of the two definitions is adopted (see, however, Lemma 5.9 in Section 5.3).

Remark. By an easy Löwenheim-Skolem argument, it is sufficient to verify the semantical tightness/solidity of T only on countable models of T.

Now we pass to two notions of syntactical character, which are properly speaking more "completeness-like" than "categoricity-like", as they do not imply that certain models are isomorphic but merely that they are elementarily equivalent.

Definition 2.3 (Tightness)**.** A theory T is *tight* if whenever $U \subseteq \mathcal{L}_T$ and $V \subseteq \mathcal{L}_T$ are two extensions of T which are bi-interpretable, then $U \equiv V$.

Definition 2.4 (Neatness)**.** A theory T is *neat* if whenever $U \subseteq \mathcal{L}_T$ and $V \subseteq \mathcal{L}_T$ are two extensions of T and U is a retract of V, then $U \vdash V$.

Remark. By an easy argument, it is enough to verify the tightness/neatness of T only on complete extensions of T. Hence in particular a theory is tight if and only if the bi-interpretability relation on its complete extensions coincides with the identity.

It is quite easy to see that all the arrows in the diagram below correspond to implications between the four properties:

$$\begin{tikzcd} & \text{solidity} \arrow[dl] \arrow[dr] & \\ \text{neatness} \arrow[dr] & & \text{semantical tightness} \arrow[dl] \\ & \text{tightness} & \end{tikzcd}$$

As for the question which of the implications are strict, see some results in Section 5 and the discussion in Section 7.

2.3 Basic first-order arithmetic

In this paper we focus mainly on theories in the language of arithmetic $\mathcal{L}_{\mathrm{PA}}$ (i.e. the usual language of ordered rings), though occasionally we also consider extensions of $\mathcal{L}_{\mathrm{PA}}$ by finitely many predicates (see e.g. Section 3.3) or other languages. Below, we review some basic properties of arithmetical theories that we will need later on, and we fix some notational conventions. A few more advanced topics related to first-order arithmetic are discussed in Section 3.

Standard notions and classical results in the model and proof theory of first-order arithmetic that we rely on can be found in the monographs [12] and [16].

Arithmetical theories. All the arithmetical theories that we consider extend PA^-, the finitely axiomatized theory of non-negative parts of discretely ordered rings. It was shown by Jeřábek [15] that PA^- is a *sequential* theory, which means roughly that it supports a reasonably behaved theory of finite sequences of arbitrary elements. (For a precise definition of sequentiality, an important general concept that was in fact discovered in the study of interpretability [25], see e.g. [28, Section 2.4].) Sequentiality of PA^- implies that there is an interpretation M of PA^- in PA^-, with the domain forming a *definable cut* in PA^- (i.e., provably closed downwards under $\leq$ under and successor) and the arithmetical operations translated identically, and there is a formula $y = x_z$ (intended to mean "y is the z-th element of the sequence x") such that PA^- proves the statement:

$$\forall s, x, k\, \exists s'\, \forall i, y\, \big(\delta_{\mathsf{M}}(k) \wedge i \leq k \rightarrow (y = s'_i \leftrightarrow (i < k \wedge y = s_i) \vee (i = k \wedge y = x))\big).$$

This says that a given sequence s can be extended/modified by inserting an arbitrary element x in the position indexed by an arbitrary number k from the domain of M.

Most of the systems we study extend $\mathrm{I}\Delta_0 + \exp$, also known as elementary arithmetic EA or elementary function arithmetic EFA, a well-known theory whose provably total functions are exactly the elementary computable functions. This theory extends PA^- by the induction scheme for all Δ_0 formulas, $\mathrm{I}\Delta_0$, and a single axiom stating that the exponential function is total. In general, if Γ is a class of formulas, then $\mathrm{I}\Gamma$ denotes the extension of PA^- by all instances of the induction scheme for formulas from Γ, while $B\Gamma$ denotes the extension of $\mathrm{I}\Delta_0$ by all instances of the collection scheme for formulas from Γ.

Encoding of syntax and set theory. The theory $\mathrm{I}\Delta_0 + \exp$ allows for a convenient encoding of finite sets via the Ackermann membership relation $\in_A$ ("the x-th bit in the binary notation for y is 1"). Except for Section 5.3, when a more fancy coding is required by the weaker setting, all encoding of set-theoretic and syntactical notions is done using $\in_A$. For a syntactical object o (a formula, a term, a proof), $\ulcorner o \urcorner$ denotes the Gödel code of o. We let $\mathrm{seq}(x)$ and $\mathrm{len}(s) = x$ be some fixed arithmetical formulas expressing (in terms of $\in_A$) that s is a sequence and that the length of the sequence s is x, respectively, and we let y_x or $(y)_x$ stand for the x-th element of the sequence y. Given an arithmetization of some language $\mathcal{L}$, we let $\mathrm{Form}_{\mathcal{L}}(x)$, $\mathrm{Form}^{\leq 1}_{\mathcal{L}}$, $\mathrm{Sent}_{\mathcal{L}}(x)$, $\mathrm{Term}_{\mathcal{L}}(x)$, $\mathrm{Var}(x)$ denote the arithmetical formulas expressing respectively that x is (the Gödel code of) an $\mathcal{L}$-formula, an $\mathcal{L}$-formula with at most one free variable, an $\mathcal{L}$-sentence, an $\mathcal{L}$-term, and a variable. Omitting the subscript $\mathcal{L}$ indicates that the intended language is $\mathcal{L}_{\mathrm{PA}}$. If Γ is class of formulas, then $\mathrm{Form}_{\Gamma}(x)$ denotes the arithmetical definition of Γ.

We use the notation $\mathrm{name}(x)$ to denote the the arithmetical naming function, which given a number x returns (the code of) a canonical numeral naming x, say of

$$\underbrace{(\ldots((0+1)+1)\ldots+1)}_{x\text{ additions}}.$$

We use $\mathrm{val}(t)$ for the function which given (the code of) a closed term t outputs its value. Given (the codes of) a formula φ a variable v and a term t, $\mathrm{subst}(\varphi,t)$ denotes the result of substituting the term t for all occurrences of the unique free variable of φ (preceded by the renaming of bound variables so to avoid clashes). The notation $\mathrm{subst}(\varphi,\mathrm{name}(x))$ is often simplified using the dot convention: instead of $\mathrm{subst}(\varphi,\mathrm{name}(x))$, we write $\ulcorner\varphi(\dot{x})\urcorner$. More generally, $\ulcorner\cdot\urcorner$ often indicates the application of some syntactical function on codes of formulas: for example, $\ulcorner\varphi\wedge\psi\urcorner$ denotes (the code of) the conjunction of given formulas φ and ψ.

Provability, reflection and partial truth predicates. If T is (an arithmetical definition of) a theory, then $\mathrm{Prov}_T(x)$ stands for the canonical provability predicate, i.e. provability in first-order logic with sentences from T as additional axioms, and $\mathrm{Proof}_T(z,x)$ stands for the formula stating that z is a proof of x in T. The sentence $\mathrm{Con}(T)$ is $\neg\,\mathrm{Prov}_T(\ulcorner 0\neq 0\urcorner)$. If Γ is a class of formulas, then $\Gamma\text{-}\mathrm{RFN}(T)$ denotes the theory extending $\mathrm{I}\Delta_0+\exp$ by the uniform Γ-reflection scheme for T, that is, by all axioms of the form

$$\forall x\,\big(\mathrm{Prov}_T(\ulcorner\varphi(\dot{x})\urcorner)\to\varphi(x)\big),$$

where $\varphi\in\Gamma$ (we assume that φ has at most one free variable). We write $\Gamma\text{-Con}(T)$ for the extension of $\mathrm{I}\Delta_0+\exp$ by all sentences of the form

$$\forall x\,\big(\varphi(x)\to\mathrm{Con}(T+\ulcorner\varphi(\dot{x})\urcorner)\big),$$

where $\varphi\in\Gamma$.

Remark. Let Σ_n, Π_n be the usual formula classes of the arithmetical hierarchy. It is easy to prove that $\Pi_n\text{-}\mathrm{RFN}(T)$ is equivalent to $\Sigma_n\text{-Con}(T)$, and vice versa.

For each $n\geq 1$ and $\Gamma\in\{\Sigma_n,\Pi_n\}$ there is a partial satisfaction predicate $\mathrm{Sat}_\Gamma(\varphi,x)$ which satisfies the usual inductive Tarskian truth conditions for formulas from Γ provably in $\mathrm{I}\Delta_0+\exp$. As a consequence, for each $\varphi(x)\in\Gamma$,

$$\mathrm{I}\Delta_0+\exp\vdash\forall x\,\big(\mathrm{Sat}_\Gamma(\ulcorner\varphi\urcorner,x)\leftrightarrow\varphi(x)\big).$$

$\mathrm{Tr}_\Gamma(\varphi)$ denotes the canonical truth predicate based on Sat_Γ, that is the formula $\mathrm{Sat}_\Gamma(\varphi,0)$ applied to sentences φ (so that the second argument, in this case fixed to be 0, does not matter).

Thanks to the partial truth predicates, for each $n\geq 1$ and $\Gamma\in\{\Sigma_n,\Pi_n : n\in\mathbb{N}\}$ the theory $\Gamma\text{-}\mathrm{RFN}(T)$ can be finitely axiomatized, using a fixed finite axiomatization of $\mathrm{I}\Delta_0+\exp$ and the sentence

$$\forall\varphi\,\big(\mathrm{Form}_\Gamma(\varphi)\wedge\mathrm{Prov}_T(\varphi)\to\mathrm{Tr}_\Gamma(\varphi)\big).$$

See [2, Lemma 2.7] for details.

Ehrenfeucht's Lemma. We recall a classical fact about models of PA (originally due to [3], a proof can also be found e.g. in [18, Theorem 1.7.2]).

Assume that $\mathcal{M}\vDash\mathrm{PA}$ and a,b are distinct elements of $\mathcal{M}$ such that b is definable from a – in other words, b is unique such that $\mathcal{M}\vDash\varphi(b,a)$, where $\varphi(x,y)$ is a formula with no free variables other than x,y. Then $\mathrm{tp}^{\mathcal{M}}(a)\neq\mathrm{tp}^{\mathcal{M}}(b)$, where the notation $\mathrm{tp}^{\mathcal{M}}(\cdot)$ refers to the complete type of an element in $\mathcal{M}$.

Models of fragments of PA. We briefly summarize some well-known constructions of models of $\mathrm{I}\Sigma_n + \exp + \neg\,\mathrm{B}\Sigma_{n+1}$ and $\mathrm{B}\Sigma_n + \exp + \neg\,\mathrm{I}\Sigma_n$. A detailed presentation can be found in [12, Chapter IV.1(d)]. In our main arguments in Section 4 and 5, we will rely heavily on the arithmetization of these constructions.

A typical method of building a model of $\mathrm{I}\Sigma_n + \neg\,\mathrm{B}\Sigma_{n+1}$ is to use *pointwise definable* structures. Given $\mathcal{M} \vDash \mathrm{PA}^-$, the substructure $\mathcal{K}_{n+1}(\mathcal{M})$ consists of those elements of $\mathcal{M}$ which are definable in $\mathcal{M}$ by a Σ_{n+1} formula. If $\mathcal{M} \vDash \mathrm{I}\Sigma_n$, then $\mathcal{K}_{n+1}(\mathcal{M}) \preccurlyeq_{n+1} \mathcal{M}$ (that is, the extension is elementary with respect to Σ_{n+1} formulas), so $\mathcal{K}_{n+1}(\mathcal{M})$ is a model of $\mathrm{I}\Sigma_n$, and it satisfies exp if $\mathcal{M}$ does. Assuming $\mathcal{K}_{n+1}(\mathcal{M}) \vDash \exp$, we also have $\mathcal{K}_{n+1}(\mathcal{M}) \vDash \neg\,\mathrm{B}\Sigma_{n+1}$ unless $\mathcal{K}_{n+1}(\mathcal{M})$ coincides with the standard model.

A typical method of building a model of $\mathrm{B}\Sigma_n + \neg\,\mathrm{I}\Sigma_n$ for $n \geq 1$ is to use a sufficiently elementary proper initial segment of a model of enough induction. If $\mathcal{M} \vDash \mathrm{I}\Sigma_n$, where $n \geq 1$, and $\mathcal{J} \preccurlyeq_{n-1} \mathcal{M}$ is a proper initial segment of $\mathcal{M}$, then $\mathcal{J} \vDash \mathrm{B}\Sigma_n$.

To get $\mathcal{J} \vDash \neg\,\mathrm{I}\Sigma_n$, we can for instance ensure that $\mathcal{J}$ is nonstandard but has a Σ_n-definition of the standard cut. One way of doing that is to let $\mathcal{J}$ be the closure of $[0, a]$, where a is a nonstandard element of $\mathcal{M}$, under the witness-bounding function for the universal Σ_{n-1} formula (that is, the function that on input x outputs the smallest y such that all Σ_{n-1} sentences (whose codes are) smaller than x are witnessed either below y or not at all). If a is a nonstandard Σ_{n-1}-definable element of $\mathcal{M}$, then the segment $\mathcal{J}$ thus obtained coincides with the structure called $\mathcal{H}_{n-1}(\mathcal{M})$ in [12], but we will reserve the letter $\mathcal{H}$ for structures of a different kind (Henkin models).

3 Groundwork

In this section, we discuss three topics in first-order arithmetic which are still of essentially preliminary character but require more extensive treatment. In some cases, this is because we need to refine standard formulations of presumably rather familiar results, in others because the results themselves and the concepts underlying them might not be very widely known.

First, we give a hierarchical version of the well-known argument showing that a model of PA is an initial segment of any model of arithmetic that it interprets. Then, we recall the notion of flexible formulas and prove the existence of particular variants of flexible formulas that we will later make use of. Finally, we discuss the topic of axiomatic truth theories with multiple nested truth predicates.

3.1 The formalized categoricity argument

It is well-known that by formalizing the classical argument used to prove that the (second-order) Dedekind-Peano axiomatization of the standard natural numbers is categorical – or to prove that the standard numbers form an initial segment of any nonstandard model of arithmetic – one can show that every model of PA embeds as an initial segment into any model of arithmetic that it can interpret. In [4], this observation is attributed to Feferman [7].

We need a hierarchical version of that result, in which the ground model might not satisfy full PA, but at the same time we have control over the complexity of the interpretation. This version is proved by mimicking the usual argument.

Definition 3.1. An interpretation M (in an $\mathcal{L}_{\mathrm{PA}}$-theory or in a structure for $\mathcal{L}_{\mathrm{PA}}$) is Σ_n-*restricted* if the formula δ_{M} and all the formulas P^{M}, for P a symbol of the interpreted language, are Σ_n.

Lemma 3.2. *Let $n \geq 1$. Suppose that $\mathcal{M} \models \mathrm{I}\Delta_0 + \exp$ and that $I \subseteq_e \mathcal{M}$ is a cut in $\mathcal{M}$ such that for every Σ_n formula $\varphi(x)$ (possibly with parameters from $\mathcal{M}$) it holds that*

$$\mathcal{M} \models \varphi(0) \wedge \forall x \left(\varphi(x) \to \varphi(x+1)\right) \Rightarrow \text{for each } a \in I \text{ it holds that } \mathcal{M} \models \varphi(a). \qquad (\dagger)$$

Suppose further that $\mathcal{N} \models \mathrm{PA}^-$ is interpreted in $\mathcal{M}$ via a Σ_n-restricted interpretation N.

Then there exists an $\mathcal{M}$-definable relation $\iota(x, y)$ such that $\iota \cap (I \times \mathcal{N})$ is an embedding of $\mathcal{L}_{\mathrm{PA}}$-structures $I \hookrightarrow \mathcal{N}$. Moreover, $\iota[I]$ is an initial segment of $\mathcal{N}$, and the definition of ι refers only to the parameters used by N.

Proof. Fix n, $\mathcal{M}$, I and N as above. In particular $\mathcal{M}^{\mathsf{N}} = \mathcal{N} \models \mathrm{PA}^-$ and

$$\delta_{\mathsf{N}}, +^{\mathsf{N}}, \times^{\mathsf{N}}, 0^{\mathsf{N}}, 1^{\mathsf{N}}, <^{\mathsf{N}}, =^{\mathsf{N}}$$

are given by Σ_n formulas. For the purpose of this proof, we introduce the following abbreviation: if $\varphi(x)$ is a Σ_n formula, then $s\colon \varphi(x)$ denotes the formula in two free variables s and x resulting from $\varphi(x)$ by deleting the leftmost existential quantifier (or quantifier block) and substituting the variable s for the variable bound by that quantifier. Hence if $\varphi(x)$ is $\exists y\, \psi(x, y)$, then $s\colon \varphi(x)$ is $\psi(x, s)$.

We define $\iota(x, y)$ to hold if:

$$\exists s, t \left[\mathrm{seq}(s, t) \wedge \mathrm{len}(s) = \mathrm{len}(t) = x + 1 \right.$$
$$\left. \wedge\, s_0 =^{\mathsf{N}} 0^{\mathsf{N}} \wedge \forall i \leq x\, (t_i\colon (s_{i+1} =^{\mathsf{N}} s_i +^{\mathsf{N}} 1_{\mathsf{N}})) \wedge s_x =^{\mathsf{N}} y\right].$$

Since the formula $t_i\colon (s_{i+1} =^{\mathsf{N}} s_i +_{\mathsf{N}} 1_{\mathsf{N}})$ is Π_{n-1}, the relation ι is Σ_n-definable. We claim that for every $a, a' \in I$ the following holds in $\mathcal{M}$:

$$\exists y\, \iota(a, y), \qquad (1)$$
$$\forall y\, \forall y' \left(\iota(a, y) \wedge \iota(a, y') \to y =^{\mathsf{N}} y'\right), \qquad (2)$$
$$\forall y\, \forall y' \left(\iota(a, y) \wedge \iota(a', y') \wedge y =^{\mathsf{N}} y' \to a = a'\right). \qquad (3)$$

The proof of (1) is a straightforward application of (†), since $\exists y\, \iota(x, y)$ is a Σ_n formula, and the subset of $\mathcal{M}$ it defines is closed under successor. The latter follows from the fact that the successor operation is provably total in PA^- and that we can always extend a given sequence by one element.

To prove (2), consider $a \in I$ and s, s' such that $s\colon \iota(a, y)$ and $s'\colon \iota(a, y')$, and apply (†) to the formula $(s_x =^{\mathsf{N}} s'_x) \vee x > a$ to prove $s_i =^{\mathsf{N}} s'_i$ for all $i \leq a$.

Finally, (3) can be proved by using (†) to simulate the Σ_n least number principle up to elements of I: if $a \in I$ is the smallest number for which there is a' witnessing that ι is not injective with respect to $=^{\mathsf{N}}$, we reach an easy contradiction by considering $a - 1$ and $a' - 1$.

This completes the proof that ι is an embedding from I into $\mathcal{M}$. Clearly, the definition of ι does not make use of any parameters beyond the ones used in N. It remains to check that the range of ι is an initial segment of $\mathcal{N}$. To this end, consider $a \in I$ and y, s, t, z such that that $s, t\colon \iota(a, y)$ and $z <^{\mathsf{N}} y$. Apply (†) to find $i < a$ such that $s_i \leq^{\mathsf{N}} z$ and $s_{i+1} >^{\mathsf{N}} z$. Then it is easy to check that in fact $s_i =^{\mathsf{N}} z$. □

Remark. Lemma 3.2 is stated for models of $\mathrm{I}\Delta_0 + \exp$, because the proof of the lemma officially relies on sequence coding by means of the Ackermann interpretation. However, using the sequentiality of PA^- and essentially the same proof as above but with appropriately modified sequence coding, one can obtain the following variant of the lemma:

Suppose that $\mathcal{M} \models \mathrm{PA}^-$ and that there is a shortest $\mathcal{M}$-definable cut $I \subseteq_e \mathcal{M}$. Suppose further that $\mathcal{N} \models \mathrm{PA}^-$ is interpreted in $\mathcal{M}$ via an interpretation N. Then there exists an $\mathcal{M}$-definable relation $\iota(x, y)$ such that $\iota \cap (I \times \mathcal{N})$ is an embedding $I \hookrightarrow \mathcal{N}$. Moreover, the $\iota[I]$ is an initial segment of $\mathcal{N}$, and the definition of ι refers only to the parameters used by N.

Corollary 3.3. *Let $n \geq 1$. Suppose that $T \supseteq \mathrm{I}\Sigma_n$ and that N is a Σ_n-restricted interpretation of PA^- in T. Then, provably in T, there is an embedding of id_T into N whose range is an $\leq^{\mathsf{N}}$-initial segment of δ_{N}.*

Proof. Fix any model $\mathcal{M} \models T$ and apply the proof of Lemma 3.2 to $I = M$. □

The corollary below is a syntactical incarnation of the well-known fact that the truth of Π_1 formulas is preserved in initial substructures.

Corollary 3.4. *Let $n \geq 1$. Suppose that $T \supseteq \mathrm{I}\Sigma_n$ and that N is a Σ_n-restricted interpretation of PA^- in T. Then for every Π_1-sentence φ*

$$T \vdash \varphi^{\mathsf{N}} \to \varphi.$$

3.2 Flexible formulas

Flexible formulas, or formulas whose truth values that are "as undetermined as possible", were introduced in their basic form by Mostowski [23] and Kripke [19]. The theory of flexible formulas was then developed by a number of authors, in recent years *inter alios* by Woodin, Hamkins, Blanck and Enayat. Flexible formulas of various kinds play an important technical role in our arguments. Below we introduce yet another variant, quite similar to one from [20] and well-suited to applications to arithmetical theories.

Definition 3.5. Let T be a theory, and let Θ, Γ be two classes of $\mathcal{L}_T$-formulas.

The formula $\xi(x)$ is Γ-*flexible over* T if for every formula $\varphi(x)$ in Γ, the theory $T + \forall x\, (\xi(x) \leftrightarrow \varphi(x))$ is consistent.

We say that ξ is (Θ, Γ)-*flexible over* T if for every formula $\varphi(x)$ in Γ and every $\mathcal{M} \models T$ there is a Θ-elementary extension $\mathcal{N}$ of $\mathcal{M}$ such that $\mathcal{N} \models T + \forall x\, (\xi(x) \leftrightarrow \varphi(x))$.

We note that if T is consistent and ξ is (Σ_n, Σ_k)-flexible over T, then obviously ξ is Σ_k-flexible over T as well. However, for each $k \geq 1$ one can construct a Σ_k-flexible formula which is itself Σ_k, whereas for instance a Σ_1 formula can never even be (Σ_0, Σ_1)-flexible (recall that the satisfaction of Σ_1 formulas is preserved upwards under extensions of models of $\mathrm{I}\Delta_0 + \exp$).

Theorem 3.6 below states that for every sufficiently strong consistent r.e. theory T, flexible formulas of both kinds always exist. The first part of the theorem is due to [22], and the proof of the second part is inspired by the proof of the first part as given in Lindström's book [21, Chapter 2.3, Theorem 11].

Theorem 3.6. *Let $n \geq 0, k \geq 1$, and let $T \supseteq I\Delta_0 + \exp$ be a consistent r.e. theory.*

(a) There is a Σ_k formula $\xi(x)$ that is Σ_k-flexible over T. Moreover, the statement

$$\mathrm{Con}(T) \to \forall \varphi\, \big(\mathrm{Form}_{\Sigma_k}(\varphi) \to \mathrm{Con}(T + \forall x\, (\xi(x) \leftrightarrow \varphi(x)))\big)$$

is provable in $\mathrm{I}\Delta_0 + \exp$.

(b) *There is a* $\Sigma_{\max(n+2,k)}$ *formula* $\xi(x)$ *that is* (Σ_n, Σ_k)*-flexible over* U*. Moreover, the statement*

$$\forall\varphi\left(\mathrm{Form}_{\Sigma_k}(\varphi) \to \Pi_n\text{-}\mathrm{Con}(T + \forall x\,(\xi(x) \leftrightarrow \varphi(x)))\right)$$

is provable in $\Sigma_{\max(n+3,k+1)}$-$\mathrm{RFN}(T)$*.*

Proof. We begin with, and give more details of, the proof of (b), which is a bit more technically complicated. Fix T, n, and $k \geq 1$. Let $\mathrm{Form}^{\leq 1}_{\Sigma_k}(\varphi)$ express that φ is a Σ_k formula with at most one free variable. Let $\rho'(\theta, \psi, \varphi, z)$ (where each of θ, ψ, φ is a variable) abbreviate

$$\mathrm{Form}^{\leq 1}(\theta) \wedge \mathrm{Tr}_{\Sigma_{n+1}}(\psi) \wedge \mathrm{Form}^{\leq 1}_{\Sigma_k}(\varphi) \wedge \mathrm{Proof}_T(z, \ulcorner \psi \to \neg\forall x\,(\theta(x) \leftrightarrow \varphi(x))\urcorner).$$

Let $\rho(x, y)$ stand for

$$\rho'(x, (y)_1, (y)_2, (y)_3) \wedge \forall w{<}y\, \neg\rho'(x, (w)_1, (w)_2, (w)_3).$$

Thus, intuitively, $\rho(x, y)$ says "x is a formula with one free variable, and y provides the smallest witness that some true Σ_{n+1} sentence T-provably implies that x is not equivalent to some particular Σ_k formula". We note that ρ is a $\Sigma_{n+1} \wedge \Pi_{n+1}$ formula.

By the parametric version of the diagonal lemma (see e.g. [12, Theorem III.2.1(2)]) there is a formula $\xi(x)$ such that

$$\mathrm{I}\Delta_0 + \exp \vdash \forall x\left[\xi(x) \leftrightarrow \exists y\left(\rho(\ulcorner\xi\urcorner, y) \wedge \mathrm{Sat}_{\Sigma_k}((y)_2, x)\right)\right].$$

Furthermore, we can choose ξ so that it is a $\Sigma_{\max(n+2,k)}$ formula.

We claim that $\xi(x)$ is (Σ_n, Σ_k)-flexible over T. To prove the claim, consider $\mathcal{M} \vDash T$ and a Σ_k formula $\varphi(x)$, and assume for the sake of contradiction that there is no Σ_n-elementary extension of $\mathcal{M}$ satisfying $T + \forall x\,(\xi(x) \leftrightarrow \varphi(x))$. By compactness (and an obvious pairing argument), it follows that there is a Π_n-sentence $\psi'(a)$ with a parameter $a \in \mathcal{M}$ such that $\mathcal{M} \vDash \psi'(a)$ and $T \vdash \psi'(a) \to \neg\forall x\,(\xi(x) \leftrightarrow \varphi(x))$. Since $T \cup \{\xi, \varphi\} \subseteq \mathcal{L}_{\mathrm{PA}}$, we can quantify out the parameter and conclude that T proves the following implication:

$$\exists v\, \psi'(v) \to \neg\forall x\,(\xi(x) \leftrightarrow \varphi(x)).$$

Note that this implication is a $\Sigma_{\max(n+3,k+1)}$ statement.

Let ψ be $\exists v\, \psi'(v)$, and let p be a number coding a T-proof of $\psi \to \neg\forall x\,(\xi(x) \leftrightarrow \varphi(x))$. Then

$$\mathcal{M} \vDash \rho'(\ulcorner\xi\urcorner, \ulcorner\psi\urcorner, \ulcorner\varphi\urcorner, p).$$

The number $\ell = \langle\ulcorner\varphi\urcorner, \ulcorner\psi\urcorner, p\rangle$ is standard, so by external induction we may assume that it is the smallest number witnessing $\mathcal{M} \vDash \exists y\, \rho'(\ulcorner\xi\urcorner, (y)_1, (y)_2, (y)_3)$ (importantly, whatever the actual smallest witness for $\exists y$ is, its middle coordinate is a standard Σ_k formula which is not equivalent to ξ in any Σ_n-elementary extension of $\mathcal{M}$; by slight abuse of notation, we may continue calling that formula φ).

By the minimality of ℓ, we have $\mathcal{M} \vDash \forall y\,(\rho(\ulcorner\xi\urcorner, y) \leftrightarrow y = \ell\big)$. Hence, since $\mathcal{M} \vDash \mathrm{I}\Delta_0 + \exp$, the choice of ξ implies that $\mathcal{M} \vDash \forall x\,(\xi(x) \leftrightarrow \mathrm{Sat}_{\Sigma_k}(\ulcorner\varphi\urcorner, x))$. Using the "It's snowing"-it's snowing lemma for Sat_{Σ_k}, we conclude that $\mathcal{M} \vDash \forall x\,\big(\xi(x) \leftrightarrow \varphi(x)\big)$. However, ξ cannot be equivalent to φ in any Σ_n-elementary extension of $\mathcal{M}$, including $\mathcal{M}$ itself. Thus, we have arrived at a contradiction, which proves the claim that $\xi(x)$ is (Σ_n, Σ_k)-flexible over T.

We still have to argue that the above proof can be formalized in $\Sigma_{\max(n+3,k+1)}$- $\mathrm{RFN}(T)$. This is generally unproblematic, with the following modifications:

- $\mathcal{M} \models \delta$ is replaced by $\mathrm{Sat}_{\Sigma_{\max(n+3,k+1)}}(\delta)$, for any statement δ;
- the nonexistence of a Σ_n-elementary extension of $\mathcal{M}$ satisfying T together with the equivalence of ξ and γ is replaced by $\neg\Pi_n\text{-Con}(T + \forall x\,(\xi(x) \leftrightarrow \varphi(x)))$, so instead of a sentence $\psi'(a)$ implying the inequivalence we have $\psi'(\dot{a})$ for some number a, where the numeral naming a does not have to be quantified out;
- to ensure the existence of a least triple $\langle \varphi, \psi, p\rangle$ satisfying the Σ_{n+1} formula $\rho'(\ulcorner\xi\urcorner, \varphi, \psi, p)$, we invoke $\mathrm{I}\Sigma_{n+1}$, which is already a consequence of Σ_{n+2}- $\mathrm{RFN}(\mathrm{I}\Delta_0 + \exp)$;
- finally, towards the end of the argument we conclude the inequivalence of ξ and φ (and thus reach a contradiction) by invoking $\Sigma_{\max(n+3,k+1)}$- $\mathrm{RFN}(T)$.

The proof of part (a) is as in [21], and it is similar to the argument given above but simpler and purely syntactic. One defines a Δ_0 formula $\rho'(\theta, \varphi, z)$ as above but without any mention of the Σ_{n+1} sentence ψ, so the diagonal formula $\xi(x)$ can be Σ_k rather than $\Sigma_{\max(n+2,k)}$. Then one argues like in the proof of (b), but referring only to the provability of various statements in T (or its subtheories) rather than their truth in $\mathcal{M}$. At the end of the argument we conclude that $\forall x\,(\xi(x) \leftrightarrow \varphi(x))$ is both provable and disprovable in T, which contradicts the consistency of T. $\square$

Corollary 3.7. *Suppose T is an r.e. theory which extends $\mathrm{I}\Delta_0 + \exp$ and assume that $\xi(x)$ is a Σ_1-formula that is Σ_1-flexible over T and witnesses Theorem 3.6(a). Then* $\mathrm{I}\Delta_0 + \exp \vdash \mathrm{Con}(T) \to \forall x\,\neg\xi(x)$.

Proof. Fix T and $\xi(x)$ as in the assumptions. Working in $\mathrm{I}\Delta_0 + \exp$, assume that $\exists x\,\xi(x)$. Then, by provable Σ_1-completeness, $\mathrm{Prov}_T(\ulcorner\exists x\,\xi(x)\urcorner)$. However, that implies

$$\neg\mathrm{Con}(T + \forall x\,(\xi(x) \leftrightarrow x \neq x)),$$

and $x \neq x$ is a Σ_1 formula. By the "moreover" part of Theorem 3.6(a), we obtain $\neg\mathrm{Con}(T)$. $\square$

3.3 Theories of iterated Tarskian truth

In the proofs of our main result, we will need to have access to a pair of theories which are themselves solid but additionally do not interpret models of each other in a "nice" way: specifically, no model of one of the theories should be a retract of a model of the other. It turns out that one way of securing such a property is to use Tarski's undefinability of truth theorem. Consequently, one example of not just a pair, but a whole family of such theories is supplied by the following canonical theories of truth over arithmetic with varying numbers of hierarchically nested truth predicates.

Definition 3.8. For $n \in \omega$, the theory $\mathrm{CT}^n[\mathrm{PA}]$ is formulated in the language $\mathcal{L}_n$ which extends $\mathcal{L}_{\mathrm{PA}}$ with fresh predicates $P_1, \dots, P_n$ (we assume that $\mathcal{L}_0 = \mathcal{L}_{\mathrm{PA}}$). The theories are defined inductively: $\mathrm{CT}^0[\mathrm{PA}] = \mathrm{PA}$, and $\mathrm{CT}^{n+1}[\mathrm{PA}]$ extends $\mathrm{CT}^n[\mathrm{PA}]$ by the induction scheme for all $\mathcal{L}_{n+1}$-formulas and the following axioms:

(i) $\forall t \in \mathrm{Term}\,\big(P_{n+1}(\mathrm{subst}(\ulcorner P_i(x)\urcorner, t)) \leftrightarrow P_i(\mathrm{val}(t))\big)$, for each $i = 1, \dots, n$.

(ii) $\forall s, t \in \mathrm{Term}\,\big(P_{n+1}(\ulcorner s = t\urcorner) \leftrightarrow \mathrm{val}(s) = \mathrm{val}(t)\big)$.

(iii) $\forall \varphi \in \mathrm{Sent}_{\mathcal{L}_n}\,\big(P_{n+1}(\ulcorner\neg\varphi\urcorner) \leftrightarrow \neg P_{n+1}(\varphi)\big)$.

(iv) $\forall \varphi, \psi \in \mathrm{Sent}_{\mathcal{L}_n} \big(P_{n+1}(\ulcorner \varphi \wedge \psi \urcorner) \leftrightarrow (P_{n+1}(\varphi) \wedge P_{n+1}(\psi))\big)$.

(v) $\forall \varphi \in \mathrm{Form}^{\leq 1}_{\mathcal{L}_n} \forall v \in \mathrm{Var} \big(P_{n+1}(\ulcorner \forall v\, \varphi \urcorner) \leftrightarrow \forall y\, P_{n+1}(\mathrm{subst}(\varphi, \mathrm{name}(y)))\big)$.

One usually writes CT[PA] instead of CT^1[PA], and we will occasionally write P instead of P_1. The theories CT^n[PA] are sometimes also called $\mathrm{RT}^{<n+1}$ (for example in [13]).

Remark. By induction on formula complexity inside CT^n[PA], we can show that for all $1 \leq i \leq j \leq n$, P_j agrees with P_i on $\mathcal{L}_{i-1}$, provably in CT^n[PA]. More precisely, for every $1 \leq i \leq j \leq n$ the following is provable in CT^n[PA]:

$$\forall \varphi \big(\mathrm{Sent}_{\mathcal{L}_{i-1}}(\varphi) \rightarrow (P_j(\varphi) \leftrightarrow P_i(\varphi)\big).$$

The lemma below can be seen as a generalization of the result that all the theories CT^n[PA] are solid [6]. The proof of the lemma combines a few simple but important observations concerning models of the full induction scheme and theories of iterated Tarskian truth.

Lemma 3.9. *Fix $m \in \omega$. Let $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ be models in a language extending $\mathcal{L}_{\mathrm{PA}}$ such that*

(i) $\mathcal{M}_i \models \mathrm{PA}^-$, for $i = 1, 2, 3$,

(ii) $\mathsf{M}_2 : \mathcal{M}_1 \rhd \mathcal{M}_2$ and $\mathsf{M}_3 : \mathcal{M}_2 \rhd \mathcal{M}_3$.

Suppose further that for each $i = 1, 2, 3$ we are given an interpretation N_i of CT^m[PA] in $\mathcal{M}_i$ such that

(a) the domain of N_i is the shortest definable cut in $\mathcal{M}_i$, and

(b) there exists an $\mathcal{M}_1$-definable isomorphism from $\mathcal{M}_1^{\mathsf{N}_1}$ onto $\mathcal{M}_3^{\mathsf{N}_3}$.

Then, there are $\mathcal{M}_i$-definable isomorphisms between $\mathcal{M}_i^{\mathsf{N}_i}$ and $\mathcal{M}_{i+1}^{\mathsf{N}_{i+1}}$.

Proof. Since we are going to apply N_i and M_{i+1} only in $\mathcal{M}_i$, we shall abbreviate $\mathcal{M}_i^{\mathsf{N}_i}, \mathcal{M}_i^{\mathsf{M}_{i+1}}$ as N_i, M_{i+1}, respectively. Observe that since the domain of N_i is the shortest cut in $\mathcal{M}_i$, we know that N_i satisfies induction with respect to all $\mathcal{M}_i$-definable properties. We shall refer to this feature as the $\mathcal{M}_i$-inductiveness of N_i. Thanks to $\mathcal{M}_i$-inductiveness, we can repeat the argument from Section 3.1 so as to conclude that for each $i \leq 2$ there is a $\mathcal{M}_i$-definable embedding h_i of the reduct $\mathcal{M}_i^{\mathsf{N}_i}\restriction_{\mathcal{L}_{\mathrm{PA}}}$ onto an initial segment of the reduct $\mathcal{M}_{i+1}^{\mathsf{N}_{i+1}}\restriction_{\mathcal{L}_{\mathrm{PA}}}$.

The reasoning thus far was independent of m. To prove the lemma, we use induction on m. Assume first that $m = 0$. Let j be the $\mathcal{M}_1$-definable isomorphism from N_1 onto N_3 (or, more precisely from the point of view of $\mathcal{M}_1$, onto $\mathsf{M}_2\mathsf{M}_3\mathsf{N}_3$). Since $m = 0$, in order to prove that N_i and N_{i+1} are $\mathcal{M}_i$-definably isomorphic we only need to show that both the embeddings h_i are onto N_{i+1} for their respective i. Suppose otherwise: then $(h_2 \circ h_1)(\mathsf{N}_1)$ is an $\mathcal{M}_1$-definable proper initial segment of $\mathsf{M}_2\mathsf{M}_3\mathsf{N}_3$, and thus $(j^{-1} \circ h_2 \circ h_1)(\mathsf{N}_1)$ is an $\mathcal{M}_1$-definable proper initial segment of N_1, contradicting the $\mathcal{M}_1$-inductiveness of N_1.

Now fix $m > 0$ and assume that the lemma holds for $m-1$. The inductive assumption tells us that for $i = 1, 2$, the map h_i is an isomorphism between $\mathsf{N}_i\restriction_{\mathcal{L}_{n-1}}$ and $\mathsf{N}_{i+1}\restriction_{\mathcal{L}_{n-1}}$. Let P^i_m be the m-th truth predicate of N_i. We argue that $h_i[P^i_m] = P^{i+1}_m$, which will complete the proof. Consider $h_i^{-1}[P^{i+1}_m]$. This is an $\mathcal{M}_i$-definable subset of N_i. Since h_i is an isomorphism between $\mathsf{N}_i\restriction_{\mathcal{L}_{m-1}}$ and $\mathsf{N}_{i+1}\restriction_{\mathcal{L}_{m-1}}$, we see that $h_i^{-1}[P^{i+1}_m]$ actually satisfies the axioms of $\mathrm{CT}^m[\mathrm{PA}]$, cf. Definition 3.8. In N_i, define I be the set of logical depths of those sentences for which P^i_m coincides with $h_i^{-1}[P^{i+1}_m]$. In other words, I consists of those $x \in \mathsf{N}_i$ such that $\mathcal{M}_i$ satisfies

$$\forall \varphi \in \mathsf{N}_i \Big((\varphi \in \mathrm{Form}_{\mathcal{L}_{m-1}} \wedge \mathsf{dpt}(\varphi) \leq x)^{\mathsf{N}_i} \to \big(P^i_m(\varphi) \leftrightarrow \varphi \in h_i^{-1}[P^{i+1}_m] \big) \Big) .$$

Since both P^i_m and $h_i^{-1}[P^{i+1}_m]$ satisfy conditions (i)–(ii) of Definition 3.8, we know that $0 \in I$, and since both satisfy the inductive conditions (iii)–(v), we also know that I is closed under successor. Thus, I is a $\mathcal{M}_i$-definable cut contained in N_i, and since the latter is the shortest cut in $\mathcal{M}_i$, we conclude that P^i_m actually coincides with $h_i^{-1}[P^{i+1}_m]$. □

Corollary 3.10. *For every m, the theory $\mathrm{CT}^m[\mathrm{PA}]$ is solid.*

Proof. This is a special case of Lemma 3.9 in which N_i is the identity interpretation on $\mathcal{M}_i$. □

4 Solidity below PA

This section contains the proof of our main result. We begin with the proof of a no-frills version, which simply says that there are arbitrarily strong solid proper subtheories of PA. Then, in Section 4.2, we discuss some aspects of the proof on a more abstract level. This lets us obtain two improvements of the basic result: firstly, that the solid theories can be weaker than PA in terms of not just provability, but also interpretability; secondly, that they can be made weak enough that extending them by any single true sentence, or even by all true Π_n sentences for a fixed n, will still be insufficient to derive PA. These improvements are presented in Sections 4.3 and 4.4, respectively.

4.1 The basic construction

Most of this subsection is devoted to a proof of the following theorem:

Theorem 4.1. *For every $n \in \mathbb{N}$, there exists an r.e. solid subtheory of* PA *that contains* $\mathrm{I}\Sigma_n + \exp$ *but not* $\mathrm{B}\Sigma_{n+1}$.

To prove Theorem 4.1, we will gradually introduce the main concepts involved in our argument and derive a series of lemmas about those concepts.

Fix $n \in \mathbb{N}$. We want to construct a theory $T_n \supseteq \mathrm{I}\Sigma_n + \exp$ that is a proper subtheory of PA but is nevertheless solid. To that end, we will define an auxiliary theory $\mathrm{IT}(n) \supseteq \mathrm{I}\Sigma_n + \exp$ that is inconsistent with $\mathrm{B}\Sigma_{n+1}$ but rather strong from the perspective of interpretability. Furthermore, we will define an interpretation K_n of a model of $\mathrm{IT}(n)$ in the standard model $(\mathbb{N}, \mathrm{Th}(\mathbb{N}))$ of CT[PA], and an interpretation N_n in the other direction. Our eventual theory T_n will essentially say that we are either in a universe satisfying PA or in one satisfying $\mathrm{IT}(n)$.

Remark. The interpretations K_n and N_n will in fact witness the bi-interpretability of $\mathrm{IT}(n)$ and CT[PA], which we will show as a separate proposition after the proof of Theorem 4.1. Hence the abbreviation $\mathrm{IT}(n)$, which stands for *interpreting truth.*

We now proceed to define our concepts more precisely, beginning with K_n. Let $\xi_n(x)$ be a Σ_1 formula that is Σ_1-flexible over $\mathrm{I}\Sigma_{n+1}$ and witnesses Theorem 3.6(a). The interpretation K_n describes the following process, as carried out in the standard model $(\mathbb{N}, \mathrm{Th}(\mathbb{N}))$ of CT[PA]:

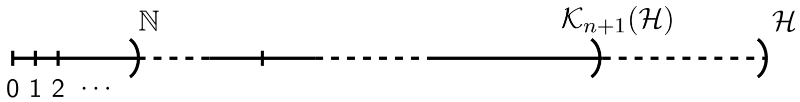


Figure 4.1: Construction of the model $(\mathbb{N}, \mathrm{Th}(\mathbb{N}))^{\mathsf{K}_n}$ of $\mathrm{IT}(n)$. The solid horizontal lines represent $(\mathbb{N}, \mathrm{Th}(\mathbb{N}))^{\mathsf{K}_n}$, which is the pointwise Σ_{n+1}-definable substructure of the Henkin structure $\mathcal{H}$. The dashed horizontal lines represent the rest of $\mathcal{H}$.

- Consider a canonically defined binary tree whose paths correspond to complete consistent henkinized extensions of the theory

$$\mathrm{I}\Sigma_{n+1} \cup \{\xi_n(\underline{k}) : k \in \mathrm{Th}(\mathbb{N})\} \cup \{\neg\xi_n(\underline{k}) : k \in \mathbb{N} \setminus \mathrm{Th}(\mathbb{N})\}. \tag{4}$$

- Take the Henkin model, say $\mathcal{H}$, given by the leftmost path through that tree.
- Take $\mathcal{K}_{n+1}(\mathcal{H})$, that is the submodel of $\mathcal{H}$ consisting of the Σ_{n+1}-definable elements.

Since $\xi_n(x)$ is a Σ_1-flexible formula over $\mathrm{I}\Sigma_{n+1}$, the compactness theorem implies that the theory in (4) is consistent. Thus, when applied in $(\mathbb{N}, \mathrm{Th}(\mathbb{N}))$, the interpretation K_n indeed produces a structure $\mathcal{K} := \mathcal{K}_{n+1}(\mathcal{H})$. The construction of $\mathcal{K}$, that is, of $(\mathbb{N}, \mathrm{Th}(\mathbb{N}))^{\mathsf{K}_n}$, is schematically presented in Figure 4.1.

Lemma 4.2. *$\mathcal{K}$ is a Σ_{n+1}-elementary substructure of $\mathcal{H}$ satisfying* $\mathrm{I}\Sigma_n + \exp + \neg\,\mathrm{B}\Sigma_{n+1}$.

Proof. It follows directly from the well-known properties of pointwise Σ_{n+1}-definable structures discussed at the end of Section 2 that $\mathcal{K} \preccurlyeq_{n+1} \mathcal{H}$ and, as a consequence, that $\mathcal{K} \vDash \mathrm{I}\Sigma_n + \exp$. We can also conclude that $\mathcal{K} \vDash \neg\,\mathrm{B}\Sigma_{n+1}$ unless $\mathcal{K}$ is the standard model.

However, since $\mathcal{H} \vDash \exists x\, \xi_n(x)$, by Σ_{n+1}-elementarity we get $\mathcal{K} \vDash \exists x\, \xi_n(x)$, which ensures nonstandardness by Corollary 3.7. $\square$

The standard cut $\mathbb{N}$ is definable in $\mathcal{K}$ as the set of those x that satisfy the formula $\delta_n(x)$: "there exists an element without a Σ_{n+1} definition smaller than x". Clearly then, $\mathbb{N}$ is the smallest definable cut of $\mathcal{K}$. Moreover, since $\xi_n(x)$ is a Σ_1 formula and $\mathcal{K} \preccurlyeq_{n+1} \mathcal{H}$, for each standard k it holds that $\mathcal{K} \vDash \xi_n(k)$ if and only if k is (the code of) an arithmetical sentence true in $\mathbb{N}$.

Thus, $(\mathbb{N}, \mathrm{Th}(\mathbb{N}))$ can be interpreted in $\mathcal{K}$ by the interpretation N_n in which the domain is defined by δ_n, the arithmetical operations are unchanged, and P is given by $\xi_n(x)$.

Lemma 4.3. *There is a $\mathcal{K}$-definable isomorphism i_n between $\mathsf{N}_n\mathsf{K}_n$ and the identity interpretation of $\mathcal{K}$ in itself, and there is an* $(\mathbb{N}, \mathrm{Th}(\mathbb{N}))$*-definable isomorphism j_n between $\mathsf{K}_n\mathsf{N}_n$ and the identity interpretation of* $(\mathbb{N}, \mathrm{Th}(\mathbb{N}))$ *in itself.*

Proof. The isomorphism i_n takes an element y of $\mathcal{K}$, finds the least Σ_{n+1} definition x of y (where x is necessarily a standard number, because $\mathcal{K}$ is in fact pointwise Σ_{n+1}-definable), and maps y to the element defined by x in the structure obtained according to K_n.

The isomorphism j_n is a special case of the map appearing in Lemma 3.2: it takes $x \in \mathbb{N}$ to the x-th smallest element of $\mathcal{K}$. By the construction of $\mathcal{K}$ and the definition of N_n, the range of j_n is exactly $\mathcal{K}^{\mathsf{N}_n} = (\mathbb{N}, \mathrm{Th}(\mathbb{N}))^{\mathsf{K}_n\mathsf{N}_n}$ and the isomorphism of arithmetical structures extends to the truth predicate P. $\square$

Note that $\mathsf{N}_n, \mathsf{K}_n, i_n, j_n$ are all definable without parameters in the respective structures. (Which is in any case obvious since both $(\mathbb{N}, \mathrm{Th}(\mathbb{N}))$ and $\mathcal{K}$ are pointwise definable.)

We let $\mathrm{IT}(n)$ be a theory axiomatizing some salient properties of $\mathcal{K}$. Namely, the axioms of $\mathrm{IT}(n)$ are:

(i) $\mathrm{I}\Sigma_n + \exp + \neg\,\mathrm{B}\Sigma_{n+1}$,

(ii) "δ_n defines a cut which is the shortest definable cut",

(iii) $\mathsf{N}_n \models \mathrm{CT}[\mathrm{PA}]$,

(iv) "$i_n \colon \mathrm{id} \to \mathsf{N}_n\mathsf{K}_n$ is an isomorphism",

(v) $\mathsf{N}_n \models$ "$j_n \colon \mathsf{id} \to \mathsf{K}_n\mathsf{N}_n$ is an isomorphism".

We note that (ii) and (iii) are infinite collections of sentences.

Finally, we let T_n be the following theory:

$$\mathrm{I}\Delta_0 + \exp \cup \{\mathrm{B}\Sigma_{n+1} \to \mathrm{I}\Sigma_k : k \in \mathbb{N}\} \cup \{\neg\,\mathrm{B}\Sigma_{n+1} \to \varphi : \varphi \in \mathrm{IT}(n)\}.$$

In other words, T_n is defined by cases: if $\mathrm{B}\Sigma_{n+1}$ holds, then PA holds, and if $\mathrm{B}\Sigma_{n+1}$ fails, then $\mathrm{IT}(n)$ holds.

Lemma 4.4. *T_n contains $\mathrm{I}\Sigma_n + \exp$ but not $\mathrm{B}\Sigma_{n+1}$. Thus, it is a proper subtheory of* PA.

Proof. By the construction of the model $\mathcal{K}$ described above, and the facts summarized in Lemmas 4.2 and 4.3, $\mathrm{IT}(n)$ is a consistent theory.

By definition, $\mathrm{IT}(n)$ contains $\mathrm{I}\Sigma_n + \exp + \neg\,\mathrm{B}\Sigma_{n+1}$, and each model of T_n is either a model of PA or one of $\mathrm{IT}(n)$. □

To prove Theorem 4.1, we need to show the solidity of T_n. This requires analyzing a number of cases dependent on the theories satisfied by models forming a potential counterexample to solidity. We prove two more lemmas, the first of which rules out a counterexample consisting of models of $\mathrm{IT}(n)$, while the other will be helpful in ruling out counterexamples in which models of $\mathrm{IT}(n)$ and of PA alternate.

Lemma 4.5. $\mathrm{IT}(n)$ *is solid.*

Proof. Let $\mathcal{M}_1 \rhd \mathcal{M}_2 \rhd \mathcal{M}_3$ be models of $\mathrm{IT}(n)$ such that there is an $\mathcal{M}_1$-definable isomorphism from $\mathcal{M}_1$ onto $\mathcal{M}_3$. For each $i \in \{1,2,3\}$, let $\mathcal{N}_i$ be the model of $\mathrm{CT}[\mathrm{PA}]$ obtained by applying the interpretation N_n in $\mathcal{M}_i$, and let $\mathcal{M}'_i$ be the model of $\mathrm{IT}(n)$ obtained by applying K_n in $\mathcal{N}_i$. Note that the domain of each $\mathcal{N}_i$ is the smallest definable cut of $\mathcal{M}_i$, by axioms (ii) of $\mathrm{IT}(n)$, and that there is an $\mathcal{M}_1$-definable isomorphism from $\mathcal{N}_1$ onto $\mathcal{N}_3$. See Figure 4.2.

We can use Lemma 3.9 for $m = 1$ to infer that there is an $\mathcal{M}_1$-definable isomorphism between $\mathcal{N}_1$ and $\mathcal{N}_2$. This isomorphism in turn clearly gives rise to an $\mathcal{M}_1$-definable isomorphism between $\mathcal{M}'_1$ and $\mathcal{M}'_2$.

By axioms (iv) of $\mathrm{IT}(n)$, for each i there is an $\mathcal{M}_i$-definable (hence $\mathcal{M}_1$-definable) isomorphism between $\mathcal{M}_i$ and $\mathcal{M}'_i$. Combining this with the isomorphism between $\mathcal{M}'_1$ and $\mathcal{M}'_2$, we obtain an $\mathcal{M}_1$-definable isomorphism between $\mathcal{M}_1$ and $\mathcal{M}_2$, which completes the proof. □

Lemma 4.6. *No model of* PA *(as an $\mathcal{L}_{\mathrm{PA}}$-structure) is a retract of a model of* $\mathrm{CT}[\mathrm{PA}]$. *In other words, if $\mathcal{M}$, $\mathcal{M}'$ are models of* PA, *$\mathcal{K}$ is a model of* $\mathrm{CT}[\mathrm{PA}]$, *and $\mathcal{M} \rhd \mathcal{K} \rhd \mathcal{M}'$, then there is no $\mathcal{M}$-definable isomorphism from $\mathcal{M}$ onto $\mathcal{M}'$.*

Similarly, no model of $\mathrm{CT}[\mathrm{PA}]$ *is a retract of a model of* PA.

$$
\begin{tikzcd}[row sep=large, column sep=large]
\mathcal{M}_1 \arrow[r, phantom, "\rhd"] \arrow[d, phantom, "\triangledown"] \arrow[rr, bend left] \arrow[dd, bend right, "i_n"'] & \mathcal{M}_2 \arrow[r, phantom, "\rhd"] \arrow[d, phantom, "\triangledown"] \arrow[dd, bend right, "i_n"'] & \mathcal{M}_3 \arrow[d, phantom, "\triangledown"] \\
\mathcal{N}_1 \arrow[d, phantom, "\triangledown"] \arrow[rr, dashed, bend left] \arrow[r, dashed, bend right] & \mathcal{N}_2 \arrow[d, phantom, "\triangledown"] & \mathcal{N}_3 \arrow[d, phantom, "\triangledown"] \\
\mathcal{M}'_1 \arrow[r, dashed, bend right] & \mathcal{M}'_2 & \mathcal{M}'_3
\end{tikzcd}
$$

Figure 4.2: The proof of Lemma 4.5. The solid arrows represent isomorphisms given directly by the assumptions about $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$, and the dashed arrows represent isomorphisms shown to exist during the argument. Composing the arrows gives an isomorphism from $\mathcal{M}_1$ onto $\mathcal{M}_2$.

Proof. We prove only the first part of the statement, as the proof of the other part is very similar. Suppose that $\mathcal{M}, \mathcal{M}' \vDash \mathrm{PA}$ and $\mathcal{K} \vDash \mathrm{CT}[\mathrm{PA}]$ with $\mathcal{M} \rhd \mathcal{K} \rhd \mathcal{M}'$, but there is an $\mathcal{M}$-definable isomorphism from $\mathcal{M}$ onto $\mathcal{M}'$. Then, by Lemma 3.9 for $m = 0$, there exists an $\mathcal{M}$-definable isomorphism from $\mathcal{M}$ onto the $\mathcal{L}_{\mathrm{PA}}$-reduct of $\mathcal{K}$.

We claim that such an isomorphism would make it possible to define a satisfaction predicate for $\mathcal{M}$ in $\mathcal{M}$, contradicting Tarski's theorem on undefinability of truth. Indeed, let $\mathsf{K} : \mathcal{M} \rhd \mathcal{K}$, and let j be the isomorphism from $\mathcal{M}$ onto $\mathcal{K}\restriction_{\mathcal{L}_{\mathrm{PA}}}$. Then, since $\mathcal{K} \vDash \mathrm{CT}[\mathrm{PA}]$, the formula

$$\sigma(x, y) := \exists x' \, \exists y' \left[j(x) =^{\mathsf{K}} x' \wedge j(y) =^{\mathsf{K}} y') \wedge \big(\mathsf{K} \vDash P(\mathrm{subst}(x', \mathrm{name}(y')))\big) \right],$$

evaluated in $\mathcal{M}$, correctly determines whether a (standard) $\mathcal{L}_{\mathrm{PA}}$-formula x is satisfied in $\mathcal{M}$ by an (arbitrary) element $y \in \mathcal{M}$.

Note that the formula $\sigma(x, y)$ may involve parameters from $\mathcal{M}$ required to define the interpretation K or the isomorphism j. However, by Tarski's theorem, not even a formula with parameters can be a definition of *satisfaction* for formulas with free variables, in contrast to merely being a definition of *truth* for sentences. $\square$

We can now complete the proof of Theorem 4.1.

Proof of Theorem 4.1. By the definition of T_n and Lemma 4.4, we already know that T_n is an r.e. subtheory of PA containing $\mathrm{I}\Sigma_n + \exp + \neg\, \mathrm{B}\Sigma_{n+1}$. It remains to show that T_n is solid.

So, let $\mathcal{M}_1 \rhd \mathcal{M}_2 \rhd \mathcal{M}_3$ be models of T_n such that there is an $\mathcal{M}_1$-definable isomorphism between $\mathcal{M}_1$ and $\mathcal{M}_3$. We need to prove that there is an $\mathcal{M}_1$-definable isomorphism between $\mathcal{M}_1$ and $\mathcal{M}_2$.

By the definition of T_n, each $\mathcal{M}_i$ satisfies either PA or $\mathrm{IT}(n)$. Moreover, clearly $\mathcal{M}_1 \equiv \mathcal{M}_3$. This leaves four cases to consider.

1° Each $\mathcal{M}_i$ satisfies PA. Then an $\mathcal{M}_1$-definable isomorphism between $\mathcal{M}_1$ and $\mathcal{M}_2$ exists by the solidity of PA.

2° Each $\mathcal{M}_i$ satisfies $\mathrm{IT}(n)$. Then the isomorphism exists by Lemma 4.5.

3° $\mathcal{M}_1 \vDash \mathrm{PA}$ and $\mathcal{M}_2 \vDash \mathrm{IT}(n)$. Let $\mathcal{N}_2$ be the model of CT[PA] obtained by applying the interpretation N_n in $\mathcal{M}_2$, and let $\mathcal{M}'_2$ be the model of $\mathrm{IT}(n)$ obtained by applying K_n in $\mathcal{N}_2$. Note that i_n applied in $\mathcal{M}_2$ is an isomorphism between $\mathcal{M}_2$ and $\mathcal{M}'_2$, by axiom (iv) of $\mathrm{IT}(n)$. Let $\mathcal{M}'_3$ be the model of PA obtained by applying in $\mathcal{M}'_2$ the interpretation provided by the formulas defining the interpretation of $\mathcal{M}_3$ in $\mathcal{M}_2$, but with all parameters of the latter replaced by their $(i_n)^{\mathcal{M}_2}$-images.

$$\begin{tikzcd}
\mathcal{M}_1 \arrow[rrrr, bend left] & \rhd & \mathcal{M}_2 \arrow[dddd, bend left, "i_n"] & \rhd & \mathcal{M}_3 \arrow[dddd, dashed, bend left] \\
 & & \triangledown & & \\
 & & \mathcal{N}_2 & & \\
 & & \triangledown & & \\
 & & \mathcal{M}'_2 & \rhd & \mathcal{M}'_3
\end{tikzcd}$$

Composing interpretations, we see that $\mathcal{M}_1 \rhd \mathcal{N}_2 \rhd \mathcal{M}'_3$. Moreover, by assumption there is an $\mathcal{M}_1$-definable isomorphism between $\mathcal{M}_1$ and $\mathcal{M}_3$, and the isomorphism $(i_n)^{\mathcal{M}_2}$ between $\mathcal{M}_2$ and $\mathcal{M}'_2$ induces an $\mathcal{M}_2$-definable (thus, $\mathcal{M}_1$-definable) isomorphism between $\mathcal{M}_3$ and $\mathcal{M}'_3$. Composing the isomorphisms from $\mathcal{M}_1$ onto $\mathcal{M}_3$ and from $\mathcal{M}_3$ onto $\mathcal{M}'_3$, we obtain an $\mathcal{M}_1$-definable isomorphism between $\mathcal{M}_1$ and $\mathcal{M}'_3$. However, then the triple of models $\mathcal{M}_1 \rhd \mathcal{N}_2 \rhd \mathcal{M}'_3$ witnesses that $\mathcal{M}_1$ is a retract of $\mathcal{N}_2$, contradicting Lemma 4.6.

4° $\mathcal{M}_1 \vDash \mathrm{IT}(n)$ and $\mathcal{M}_2 \vDash \mathrm{PA}$. This case is similar to the previous one but slightly more subtle. Let $\mathcal{N}_1 \vDash \mathrm{CT}[\mathrm{PA}]$ be obtained by applying N_n in $\mathcal{M}_1$, let $\mathcal{M}'_1 \vDash \mathrm{IT}(n)$ be obtained by applying K_n in $\mathcal{N}_1$, and let $\mathcal{N}'_1 \vDash \mathrm{CT}[\mathrm{PA}]$ be obtained by applying N_n in $\mathcal{M}'_1$. Note that $(i_n)^{\mathcal{M}_1}$ is an isomorphism between $\mathcal{M}_1$ and $\mathcal{M}'_1$, by axiom (iv) of $\mathrm{IT}(n)$, while $(j_n)^{\mathcal{N}_1}$ is an isomorphism between $\mathcal{N}_1$ and $\mathcal{N}'_1$, by axiom (v). Let $\mathcal{M}'_2 \vDash \mathrm{PA}$ be obtained by applying in $\mathcal{M}'_1$ the interpretation of $\mathcal{M}_2$ in $\mathcal{M}_1$, but with parameters moved by $(i_n)^{\mathcal{M}_1}$. Let $\mathcal{M}'_3 \vDash \mathrm{IT}(n)$ be obtained by applying in $\mathcal{M}'_2$ the interpretation of $\mathcal{M}_3$ in $\mathcal{M}_2$, again with parameters moved by $(i_n)^{\mathcal{M}_1}$. Finally, let $\mathcal{N}'_3 \vDash \mathrm{CT}[\mathrm{PA}]$ be obtained by applying N_n in $\mathcal{M}'_3$.

$$\begin{tikzcd}
\mathcal{M}_1 \arrow[rrrr, bend left] \arrow[dddd, bend left, "i_n"] & \rhd & \mathcal{M}_2 & \rhd & \mathcal{M}_3 \\
\triangledown & & & & \\
\mathcal{N}_1 \arrow[dddd, bend right, "j_n"'] & & & & \\
\triangledown & & & & \\
\mathcal{M}'_1 \arrow[rrrr, dashed, bend right] & \rhd & \mathcal{M}'_2 & \rhd & \mathcal{M}'_3 \\
\triangledown & & & & \triangledown \\
\mathcal{N}'_1 \arrow[rrrr, dashed, bend right] & & & & \mathcal{N}'_3
\end{tikzcd}$$

Composing interpretations, we see that $\mathcal{N}_1 \rhd \mathcal{M}'_2 \rhd \mathcal{N}'_3$. Moreover, there is an $\mathcal{M}'_1$-definable, and thus $\mathcal{N}_1$-definable, isomorphism from $\mathcal{M}'_1$ onto $\mathcal{M}'_3$, which induces an $\mathcal{N}_1$-definable isomorphism from $\mathcal{N}'_1$ onto $\mathcal{N}'_3$. This can be composed with the isomorphism $(j_n)^{\mathcal{N}_1}$ to give an $\mathcal{N}_1$-definable isomorphism between $\mathcal{N}_1$ and $\mathcal{N}'_3$. However, then the triple of models $\mathcal{N}_1 \rhd \mathcal{M}'_2 \rhd \mathcal{N}'_3$ witnesses that $\mathcal{N}_1$ is a retract of $\mathcal{M}'_2$, contradicting Lemma 4.6.

This concludes the proof that T_n is solid, and thus also the proof of Theorem 4.1. $\square$

To conclude the subsection, we prove a result that was already announced above: the interpretations N_n and K_n witness not only the bi-interpretability of the specific models $(\mathbb{N}, \mathrm{Th}(\mathbb{N}))$ and $\mathcal{K}$, but work in a more general axiomatic context.

Proposition 4.7. *For each* $n \geq 1$, $\mathrm{IT}(n)$ *is bi-interpretable with* $\mathrm{CT}[\mathrm{PA}]$.

Proof. We will show that N_n and K_n witness the bi-interpretability. Axioms (iii) and (iv) of $\mathrm{IT}(n)$ explicitly state that N_n is an interpretation of $\mathrm{CT}[\mathrm{PA}]$ and that $\mathsf{N}_n\mathsf{K}_n$ is isomorphic to the identity interpretation. It remains to show that K_n is really an interpretation of $\mathrm{IT}(n)$ in $\mathrm{CT}[\mathrm{PA}]$, not merely in $(\mathbb{N}, \mathrm{Th}(\mathbb{N}))$, and that $\mathsf{K}_n\mathsf{N}_n$ is isomorphic to the identity interpretation provably in $\mathrm{CT}[\mathrm{PA}]$. This is a somewhat routine verification that $\mathrm{CT}[\mathrm{PA}]$ is strong enough to carry out various constructions involved in the definitions of $\mathsf{N}_n, \mathsf{K}_n, i_n, j_n$. We provide some details.

The first step is to check that the construction of the models $\mathcal{H}$ and $\mathcal{K}$ prescribed by K_n formalizes in $\mathrm{CT}[\mathrm{PA}]$. We begin by verifying that $\mathrm{CT}[\mathrm{PA}]$ proves the consistency of the $\mathcal{L}_{\mathrm{PA}} \cup \{P\}$-definable theory

$$U := \mathrm{I}\Sigma_{n+1} \cup \{\xi_n(\underline{k}) : P(k)\} \cup \{\neg\xi_n(\underline{\ell}) : \neg P(\ell)\}.$$

(note that this is the formalized version of the theory appearing in (4) in the definition of K_n). Indeed, work in $\mathrm{CT}[\mathrm{PA}]$ and assume that U is inconsistent. Then for some disjoint finite sets c, d of numbers we have

$$\neg\mathrm{Con}\left(\mathrm{I}\Sigma_{n+1} + \bigwedge_{k \in c} \xi_n(\underline{k}) \wedge \bigwedge_{\ell \in d} \neg\xi_n(\underline{\ell})\right)$$

However, then also

$$\neg\mathrm{Con}\left(\mathrm{I}\Sigma_{n+1} + \forall x \left(\xi_n(x) \equiv \bigvee_{k\in c} x = \underline{k}\right)\right).$$

Since ξ_n is provably Σ_1-flexible over $\mathrm{I}\Sigma_{n+1}$ in the sense of Theorem 3.6(a), and $\bigvee_{k\in c} x = \underline{k}$ is a Σ_1 formula (quantifier-free, in fact), the "moreover" part of Theorem 3.6(a) gives us $\neg\mathrm{Con}(\mathrm{I}\Sigma_{n+1})$. But we are working in CT[PA], so we do have $\mathrm{Con}(\mathrm{I}\Sigma_{n+1})$, and thus we reach a contradiction. This concludes the proof of $\mathrm{Con}(U)$ in CT[PA].

From now on it will be convenient to assume that we are working with a given model $\mathcal{M} \vDash \mathrm{CT[PA]}$. We already know that $\mathcal{M} \vDash \mathrm{Con}(U)$, so we can apply the usual construction associated with the arithmetized completeness theorem (see e.g. [16, Theorem 13.13]) to U in $\mathcal{M}$. We only need the construction in its most basic form: produce a ($\Delta_1(P)$-definable) henkinized consistent extension of T and take the ($\Delta_2(P)$-definable) leftmost path through the infinite binary tree whose paths correspond to complete consistent extensions of the henkinization. Thus we obtain a model of U, say $\mathcal{H}^{\mathcal{M}}$, definable in $\mathcal{M}$ by a formula that by abuse of notation we could call $\mathcal{H}$. In fact, the structure $\mathcal{H}^{\mathcal{M}}$ has an $\mathcal{M}$-definable satisfaction relation (or, in other words, $\mathcal{M}$-definable elementary diagram). Thus, we can easily define the structure $\mathcal{M}^{\mathsf{K}_n} := (\mathcal{K}^{n+1})^{\mathcal{M}}(\mathcal{H}^{\mathcal{M}})$ consisting of those elements of $\mathcal{H}^{\mathcal{M}}$ that are definable in $\mathcal{H}^{\mathcal{M}}$ by Σ_{n+1} formulas in the sense of $\mathcal{M}$. In contrast to $\mathcal{H}^{\mathcal{M}}$, from the point of view of $\mathcal{M}$ the structure $\mathcal{M}^{\mathsf{K}_n}$ is only a partial model in the sense that $\mathcal{M}$ cannot define the full satisfaction relation for $\mathcal{M}^{\mathsf{K}_n}$ but only its universe and operations. Of course, that is already enough to define satisfaction for Σ_m formulas for any fixed m. In other words, we have $\mathcal{M}$-definable predicates $\mathcal{H} \vDash \varphi(x)$ and $\mathsf{K}_n \vDash_m \varphi(x)$, for any $m \in \omega$, which agree with satisfaction in $\mathcal{H}^{\mathcal{M}}$ resp. $\mathcal{M}^{\mathsf{K}_n}$ for atomic formulas and satisfy the usual inductive clauses of a definition of satisfaction for all $\mathcal{M}$-formulas resp. for all $\mathcal{M}$-formulas that belong to the class Σ_m.

We still have to check that K_n provides an interpretation of $\mathrm{IT}(n)$, or in other words, that $\mathcal{M}^{\mathsf{K}_n} \vDash \mathrm{IT}(n)$. In the process, we will also check that K_n and N_n give rise to a bi-interpretation.

The verification that $\mathcal{M}^{\mathsf{K}_n}$ is a model of $\mathrm{I}\Sigma_n + \exp$ and that Σ_{n+1}-elementarity holds between $\mathcal{M}^{\mathsf{K}_n}$ and $\mathcal{H}^{\mathcal{M}}$, i.e. that

$$\mathcal{M} \vDash \forall \varphi \in \mathrm{Form}_{\Sigma_{n+1}}\, \forall x \in \mathsf{K}_n \left(\mathsf{K}_n \vDash_{n+1} \varphi(x) \leftrightarrow \mathcal{H} \vDash \varphi(x)\right),$$

is straightforward.

Recall the map named j_n in Lemma 4.3 and first introduced in the proof of Lemma 3.2, namely the one taking $k \in \mathcal{M}$ to the k-th smallest element of $\mathcal{M}^{\mathsf{K}_n}$. By the argument from Section 3.1, the map j_n is an $\mathcal{M}$-definable embedding of $\mathcal{M}\restriction_{\mathcal{L}_{\mathrm{PA}}}$ onto an initial segment of $\mathcal{M}^{\mathsf{K}_n}$. Let $\mathcal{J}$ be the ($\mathcal{M}$-definable) image of j_n. We know that $\mathcal{J}$ is also an initial segment of $\mathcal{H}^{\mathcal{M}}$, since every element of $\mathcal{H}^{\mathcal{M}}$ that is below $j_n(k)$ is named by a numeral from $\mathcal{M}$, so it is Σ_{n+1}-definable in the sense of $\mathcal{M}$. Moreover, $\mathcal{J}$ is a proper cut in $\mathcal{M}^{\mathsf{K}_n}$: otherwise, $\mathcal{M}^{\mathsf{K}_n}$ would be isomorphic to $\mathcal{M}\restriction_{\mathcal{L}_{\mathrm{PA}}}$, which cannot happen by Corollary 3.7, because $\mathcal{M} \vDash \mathrm{Con}(\mathrm{I}\Sigma_{n+1})$, while $\mathcal{H}^{\mathcal{M}}$ and as a consequence $\mathcal{M}^{\mathsf{K}_n}$ both satisfy $\exists x\, \xi_n(x)$.

By the definition of $\mathcal{M}^{\mathsf{K}_n}$ and Σ_{n+1}-elementarity, each element of $\mathcal{M}^{\mathsf{K}_n}$ can be Σ_{n+1}-defined by a formula in $\mathcal{J}$. This lets us carry out the usual argument showing that $\mathcal{M}^{\mathsf{K}_n} \vDash \neg\,\mathrm{B}\Sigma_{n+1}$, so $\mathcal{M}^{\mathsf{K}_n}$ validates axiom (i) of $\mathrm{IT}(n)$.

Clearly, $\mathcal{J}$ is the smallest $\mathcal{M}$-definable cut in $\mathcal{M}^{\mathsf{K}_n}$ (and thus, also the smallest $\mathcal{M}^{\mathsf{K}_n}$-definable cut), because otherwise $\mathcal{M}$ would define its own proper cut. This implies in particular that $\mathcal{J} \subseteq \delta_n^{\mathcal{M}^{\mathsf{K}_n}}$. But we also have $\delta_n^{\mathcal{M}^{\mathsf{K}_n}} \subseteq \mathcal{J}$, because each element of $\mathcal{M}^{\mathsf{K}_n}$ has a Σ_{n+1} definition in $\mathcal{J}$. So, $\delta_n^{\mathcal{M}^{\mathsf{K}_n}} = \mathcal{J}$, and thus $\mathcal{M}^{\mathsf{K}_n}$ satisfies axioms (ii) of $\mathrm{IT}(n)$.

For each $k \in \mathcal{M}$, we have that $P(k)$ holds in $\mathcal{M}$ exactly if $\xi_n(j_n(k))$ holds in $\mathcal{H}^{\mathcal{M}}$. This is because $\mathcal{M} \models (\mathcal{H} \models U)$ and $j_n(k)$ is the element named by the numeral $\underline{k}$ in $\mathcal{H}^{\mathcal{M}}$. The truth values of ξ_n are the same in $\mathcal{M}^{\mathsf{K}_n}$ as in $\mathcal{H}^{\mathcal{M}}$, by Σ_{n+1}-elementarity. So, by the definition of N_n, we indeed have $\mathcal{M}^{\mathsf{K}_n\mathsf{N}_n} \models \mathrm{CT[PA]}$, which means that $\mathcal{M}^{\mathsf{K}_n}$ satisfies axioms (iii). Moreover, we have just shown that j_n is an isomorphism between $\mathcal{M}$ and $\mathcal{M}^{\mathsf{K}_n\mathsf{N}_n}$. This also implies that (the map defined by the same formula as) j_n is an isomorphism between $\mathcal{M}^{\mathsf{K}_n\mathsf{N}_n}$ and $\mathcal{M}^{\mathsf{K}_n\mathsf{N}_n\mathsf{K}_n\mathsf{N}_n}$, so $\mathcal{M}^{\mathsf{K}_n}$ satisfies axiom (v).

Finally we argue that $\mathcal{M}^{\mathsf{K}_n}$ satisfies (iv). Since $\mathcal{J} = \delta_n^{\mathcal{M}^{\mathsf{K}_n}}$ is the shortest cut in $\mathcal{M}^{\mathsf{K}_n}$, and each element of $\mathcal{M}^{\mathsf{K}_n}$ is definable via a Σ_{n+1}-definition from $\mathcal{J}$, the definition of i_n makes sense: each element of $\mathcal{M}^{\mathsf{K}_n}$ has a least Σ_{n+1}-definition. To verify that i_n is indeed an isomorphism between $\mathcal{M}^{\mathsf{K}_n}$ and $\mathcal{M}^{\mathsf{K}_n\mathsf{N}_n\mathsf{K}_n}$, one uses the fact that $\mathcal{M}$ and $\mathcal{M}^{\mathsf{K}_n\mathsf{N}_n}$ are isomorphic via j_n. □

4.2 Modularizing the construction

In the proof of Theorem 4.1 we made use of Lemma 4.6: no model of PA can be a retract of a model of CT[PA], and vice versa. We now carry out a more general study of families of theories with this extreme form of non-bi-interpretability property. Infinite families of this kind will be needed in our proofs of refinements of Theorem 4.1.

Definition 4.8. We say that the family $\{U_k\}_{k\in\omega}$ of theories is *retract-disjoint* if for any $k, n \in \omega$ the following holds: if $\mathcal{M} \models U_k$ and $\mathcal{N} \models U_n$ and $\mathcal{M}$ is a retract of $\mathcal{N}$, then $k = n$.

Definition 4.9. Let $\{U_k\}_{k\in\omega}$ be a sequence of theories and $\{\varphi_k\}_{k\in\omega}$ be a sequence of sentences. The symbol $\bigoplus_k(U_k|\varphi_k)$ denotes the theory $\{\varphi_k \to \psi : k \in \omega, \psi \in U_k\}$.

Proposition 4.10. *Suppose that $\{U_k\}_{k\in\omega}$ is a family of solid theories, that V is a theory, and that $\{\varphi_k\}_{k\in\omega}$ is a sequence of sentences with the following properties:*

1. *the sentences φ_k for $k \in \omega$ are pairwise inconsistent,*
2. *the theory $V \cup \{\neg\varphi_k : k \in \omega\}$ is solid,*
3. *the family $\{V \cup \{\neg\varphi_k : k \in \omega\}\} \cup \{U_k\}_{k\in\omega}$ is retract-disjoint.*

Then the theory $V \cup \bigoplus_k(U_k|\varphi_k)$ is solid.

Proof. Assume that $\mathcal{M} \models V \cup \bigoplus_k(U_k|\varphi_k)$ and there is a retraction N, M such that $\mathcal{M}^{\mathsf{N}} \models V \cup \bigoplus_k(U_k|\varphi_k)$. Put $\mathcal{N} := \mathcal{M}^{\mathsf{N}}$. By our assumptions, exactly one of the following two cases holds:

1° there is $k \in \omega$ such that both $\mathcal{M}$ and $\mathcal{N}$ are models of U_k,

2° both $\mathcal{M}$ and $\mathcal{N}$ are models of $V + \{\neg\varphi_k : k \in \omega\}$.

By solidity of each of the relevant theories, in either case there is an $\mathcal{M}$-definable isomorphism from $\mathcal{M}$ onto $\mathcal{N}$. □

Lemma 4.11. *The family $\{\mathrm{CT}^k[\mathrm{PA}]\}_{k\in\omega}$ is retract-disjoint.*

Proof. The argument is a slight generalization of the proof of Lemma 4.6. Fix $\mathcal{M}_1 \models \mathrm{CT}^k$ and $\mathcal{M}_2 \models \mathrm{CT}^m$ and assume that $\mathcal{M}_1$ is a retract of $\mathcal{M}_2$ as witnessed by the retraction $(\mathsf{M}_2, \mathsf{M}_1)$. Without loss of generality assume that $\mathcal{M}_2 = \mathcal{M}_1^{\mathsf{M}_2}$. Aiming at a contradiction assume further that $k \neq m$. Let $\ell := \min\{k, m\}$. By Lemma 3.9, we conclude that for $i \leq 2$ the $\mathcal{L}_\ell$-reducts of $\mathcal{M}_i$ and $\mathcal{M}_{i+1}$ are isomorphic via an $\mathcal{M}_i$-definable isomorphism.

Now we can assume without loss of generality that $\ell = k < m$. Arguing like in the proof of Lemma 4.6, we can use the $(k+1)$-th truth predicate of $\mathcal{M}_2$ to define satisfaction for $\mathcal{L}_k$-formulas in $\mathcal{M}_1$ within $\mathcal{M}_1$, contradicting undefinability of truth. $\square$

The proposition below and its consequence, Corollary 4.13, will play a key role in verifying retract disjointness of various families of theories. The argument is presented in the diagram below, and it is probably most convenient to follow the formulation of the proposition and its proof while looking at the diagram. We use the following local convention: whenever our assumptions imply that a structure $\mathcal{A}$ is interpretable in a structure $\mathcal{B}$, the symbol $\mathsf{A}_{\mathcal{B}}$ stands for an interpretation of (an isomorphic copy of) $\mathcal{A}$ in $\mathcal{B}$ witnessing this fact and having whatever additional properties are postulated by the assumptions.

Proposition 4.12. *Suppose that $\mathcal{M}, \mathcal{N}, \mathcal{K}, \mathcal{G}$ are structures such that:*

- *$\mathcal{M}$ is a retract of $\mathcal{N}$ via the retraction $(\mathsf{N}_{\mathcal{M}}, \mathsf{M}_{\mathcal{N}})$,*
- *$\mathcal{M}$ is bi-interpretable with $\mathcal{K}$ via the bi-interpretation $(\mathsf{K}_{\mathcal{M}}, \mathsf{M}_{\mathcal{M}})$,*
- *$\mathcal{N}$ is a retract of $\mathcal{G}$ via the retraction $(\mathsf{G}_{\mathcal{N}}, \mathsf{N}_{\mathcal{G}})$.*

Then $\mathcal{K}$ is a retract of $\mathcal{G}$. Moreover, the retraction $(\mathsf{G}_{\mathcal{K}}, \mathsf{K}_{\mathcal{G}})$ witnessing this is such that the isomorphism between $\mathcal{M}^{\mathsf{N}_{\mathcal{M}}\mathsf{G}_{\mathcal{N}}}$ and $\mathcal{M}^{\mathsf{K}_{\mathcal{M}}\mathsf{G}_{\mathcal{K}}}$ is $\mathcal{M}$-definable.

In the statement of the proposition, the mere fact that $\mathcal{K}$ is a retract of $\mathcal{G}$ can be obtained from the transitivity of retracts, which is a special case of the proposition with a considerably simpler proof. However, a more involved argument seems needed to obtain the definability relation expressed in the "moreover" part of the statement.

Proof. Note that, by our convention, $\mathcal{N}$ is isomorphic to $\mathcal{M}^{\mathsf{N}_{\mathcal{M}}}$ and $\mathcal{K}$ is isomorphic to $\mathcal{M}^{\mathsf{K}_{\mathcal{M}}}$; the interpretations $\mathsf{N}_{\mathcal{M}}$ and $\mathsf{M}_{\mathcal{N}}$ witness that $\mathcal{M}$ is a retract of $\mathcal{N}$; etc. Below, we will identify $\mathcal{N}$ with $\mathcal{M}^{\mathsf{N}_{\mathcal{M}}}$, $\mathcal{K}$ with $\mathcal{M}^{\mathsf{K}_{\mathcal{M}}}$, and $\mathcal{G}$ with $\mathcal{N}^{\mathsf{G}_{\mathcal{N}}}$ to simplify the notation. Also for the sake of notational simplicity, we assume that all the interpretations and isomorphisms involved are definable without parameters; otherwise, nothing substantial would change but we would have to keep track of how the parameters are mapped by various isomorphisms.

$$
\begin{tikzcd}
\mathcal{M} \arrow[rrrr, bend left, "\iota"] \arrow[dddd, bend right, "i"'] & \rhd^{\mathsf{N}_{\mathcal{M}}} & \mathcal{N} & \rhd^{\mathsf{M}_{\mathcal{N}}} & \mathcal{N}^{\mathsf{M}_{\mathcal{N}}} \\
\triangledown & & & & \\
\mathcal{K} \arrow[dddd, bend right] & & & & \\
\triangledown & & & & \\
\mathcal{M}_1 \arrow[rrrr, bend left, "\widehat{\iota}"] & \rhd^{\mathsf{N}_{\mathcal{M}}} & \mathcal{N}_1 \arrow[dddd, bend right, "k"'] & \rhd^{\mathsf{M}_{\mathcal{N}}} & \mathcal{N}_1^{\mathsf{M}_{\mathcal{N}}} \arrow[dddd, "k_1"] \\
\triangledown & & \triangledown & & \\
\mathcal{M}_1^{\mathsf{K}_{\mathcal{M}}} \arrow[ddddrrrr, dashed, bend right] & & \mathcal{N}_1^{\mathsf{G}_{\mathcal{N}}} & & \\
& & \triangledown & & \\
& & \mathcal{N}_2 & \rhd^{\mathsf{M}_{\mathcal{N}}} & \mathcal{M}_2 \\
& & & & \triangledown \\
& & & & \mathcal{M}_2^{\mathsf{K}_{\mathcal{M}}}
\end{tikzcd}
$$

Define

$$\mathcal{M}_1 := \mathcal{K}^{\mathsf{M}_{\mathcal{K}}},\ \mathcal{N}_1 := \mathcal{M}_1^{\mathsf{N}_{\mathcal{M}}}$$

(note that $\mathcal{M}_1$ is $\mathcal{M}^{\mathsf{K}_{\mathcal{M}}\mathsf{M}_{\mathcal{K}}}$ and $\mathcal{N}_1$ is $\mathcal{M}^{\mathsf{K}_{\mathcal{M}}\mathsf{M}_{\mathcal{K}}\mathsf{N}_{\mathcal{M}}}$). By our assumptions, $\mathsf{K}_{\mathcal{M}}$ and $\mathsf{M}_{\mathcal{K}}$ witness that $\mathcal{K}$ is a retract of $\mathcal{M}$, so there is an $\mathcal{M}$-definable isomorphism $i : \mathcal{M} \to \mathcal{M}_1$. This isomorphism induces further $\mathcal{M}$-definable isomorphisms $\mathcal{N} \to \mathcal{N}_1$ and $\mathcal{N}^{\mathsf{M}_{\mathcal{N}}} \to \mathcal{N}_1^{\mathsf{M}_{\mathcal{N}}}$. Thus $\mathcal{G}$ is ($\mathcal{M}$-definably) isomorphic to $\mathcal{N}_1^{\mathsf{G}_{\mathcal{N}}}$, and the interpretations $\mathsf{G}_{\mathcal{N}}$ and $\mathsf{N}_{\mathcal{G}}$ witness that $\mathcal{N}_1$ is a retract of $\mathcal{N}_1^{\mathsf{G}_{\mathcal{N}}}$. Moreover, since $\mathcal{N}$ and $\mathcal{N}^{\mathsf{G}_{\mathcal{N}}\mathsf{N}_{\mathcal{G}}}$ are $\mathcal{N}$-definably isomorphic, there is an $\mathcal{N}_1$-definable isomorphism $k : \mathcal{N}_1 \to \mathcal{N}_1^{\mathsf{G}_{\mathcal{N}}\mathsf{N}_{\mathcal{G}}}$. Let $\mathcal{N}_2$ be $\mathcal{N}_1^{\mathsf{G}_{\mathcal{N}}\mathsf{N}_{\mathcal{G}}}$. Composing the interpretations $\mathsf{M}_{\mathcal{K}}$, $\mathsf{N}_{\mathcal{M}}$, and $\mathsf{G}_{\mathcal{N}}$ witnesses that $\mathcal{N}_1^{\mathsf{G}_{\mathcal{N}}}$ is interpretable in $\mathcal{K}$. Also, k induces an $\mathcal{N}_1$-definable isomorphism k_1 between $\mathcal{N}_1^{\mathsf{M}_{\mathcal{N}}}$ and $\mathcal{N}_2^{\mathsf{M}_{\mathcal{N}}} =: \mathcal{M}_2$, and $\mathcal{M}_2^{\mathsf{K}_{\mathcal{M}}}$ is interpretable in $\mathcal{N}_1^{\mathsf{G}_{\mathcal{N}}}$ via the composition of $\mathsf{N}_{\mathcal{G}}$, $\mathsf{M}_{\mathcal{N}}$, and $\mathsf{K}_{\mathcal{M}}$.

So, we have interpretations of $\mathcal{N}_1^{\mathsf{G}_{\mathcal{N}}}$ in $\mathcal{K}$ and of $\mathcal{M}_2^{\mathsf{K}_{\mathcal{M}}}$ in $\mathcal{N}_1^{\mathsf{G}_{\mathcal{N}}}$. We want to show that there is a $\mathcal{K}$-definable isomorphism between $\mathcal{K}$ and $\mathcal{M}_2^{\mathsf{K}_{\mathcal{M}}}$. Note firstly that, by the choice of $\mathsf{M}_{\mathcal{N}}$ and $\mathsf{N}_{\mathcal{M}}$, there is an $\mathcal{M}$-definable isomorphism $\iota : \mathcal{M} \to \mathcal{N}^{\mathsf{M}_{\mathcal{N}}}$. Thus there is also an $\mathcal{M}_1$-definable isomorphism $\widehat{\iota} : \mathcal{M}_1 \to \mathcal{N}_1^{\mathsf{M}_{\mathcal{N}}}$. Composed with k_1, this gives an $\mathcal{M}_1$-definable isomorphism $\mathcal{M}_1 \to \mathcal{M}_2$, which induces an $\mathcal{M}_1$-definable (thus, $\mathcal{K}$-definable) isomorphism $\mathcal{M}_1^{\mathsf{K}_{\mathcal{M}}} \to \mathcal{M}_2^{\mathsf{K}_{\mathcal{M}}}$. But $\mathsf{K}_{\mathcal{M}}$ and $\mathsf{M}_{\mathcal{K}}$ are chosen so as to witness the bi-interpretability of $\mathcal{M}$ and $\mathcal{K}$, so there is a $\mathcal{K}$-definable isomorphism $\mathcal{K} \to \mathcal{M}_1^{\mathsf{K}_{\mathcal{M}}}$. Composed with the isomorphism $\mathcal{M}_1^{\mathsf{K}_{\mathcal{M}}} \to \mathcal{M}_2^{\mathsf{K}_{\mathcal{M}}}$, this gives the desired $\mathcal{K}$-definable isomorphism $\mathcal{K} \to \mathcal{M}_2^{\mathsf{K}_{\mathcal{M}}}$.

This completes the proof that $\mathcal{K}$ is a retract of $\mathcal{G}$. For the "moreover" part, notice that $\mathsf{G}_{\mathcal{K}}$, the interpretation of (a copy of) $\mathcal{G}$ in $\mathcal{K}$ in the retraction, is such that $\mathcal{K}^{\mathsf{G}_{\mathcal{K}}} = \mathcal{M}^{\mathsf{K}_{\mathcal{M}}\mathsf{G}_{\mathcal{K}}}$ is $\mathcal{N}_1^{\mathsf{G}_{\mathcal{N}}}$. On the other hand, $\mathcal{M}^{\mathsf{N}_{\mathcal{M}}\mathsf{G}_{\mathcal{N}}}$ is $\mathcal{G}$, and we have already shown that $\mathcal{G}$ is $\mathcal{M}$-definably isomorphic to $\mathcal{N}_1^{\mathsf{G}_{\mathcal{N}}}$. □

Corollary 4.13. *Suppose that $\{U_k\}_{k\in\omega}$ and $\{V_k\}_{k\in\omega}$ are two sequences of theories such that for each k, the theory U_k is bi-interpretable with V_k. Then if $\{U_k\}_{k\in\omega}$ is retract-disjoint, then $\{V_k\}_{k\in\omega}$ is retract-disjoint.*

Proof. Assume that $m \neq n$ and that $\mathcal{M} \vDash V_m$ is a retract of $\mathcal{N} \vDash V_n$. By the assumption on bi-interpretability between theories, there are $\mathcal{K} \vDash U_m$ and $\mathcal{G} \vDash U_n$ which are bi-interpretable with $\mathcal{M}$ and $\mathcal{N}$, respectively. By Proposition 4.12, $\mathcal{K}$ is a retract of $\mathcal{G}$, so the family $\{U_k\}_{k\in\omega}$ is not retract-disjoint. □

The fact expressed in the corollary below was first observed in [4]. Here we derive it using the argument from the proof of Proposition 4.12.

Corollary 4.14. *If U and V are bi-interpretable theories and U is solid, then V is solid.*

Proof. Let U and V be as above, and let $\mathsf{V} : U \rhd V$ and $\mathsf{U} : V \rhd U$ witness the bi-interpretability. Assume that $\mathcal{M} \vDash V$ and let (N, M) be a retraction in $\mathcal{M}$ such that $\mathcal{M}^{\mathsf{N}} \vDash V$. Put $\mathcal{N} = \mathcal{M}^{\mathsf{N}}$.

Applying the argument from the proof of Proposition 4.12 to $\mathcal{M}$, $\mathcal{N}$, $\mathcal{K} := \mathcal{M}^{\mathsf{U}}$, and $\mathcal{G} := \mathcal{N}^{\mathsf{U}}$, we can fix a retraction (G, K) witnessing that $\mathcal{K}$ is a retract of $\mathcal{G}$ and the isomorphism between $\mathcal{M}^{\mathsf{NU}}$ $(= \mathcal{G})$ and $\mathcal{M}^{\mathsf{UG}}$ $(= \mathcal{K}^{\mathsf{G}})$ is $\mathcal{M}$-definable.

By the solidity of U, there is a $\mathcal{K}$-definable isomorphism between $\mathcal{K}$ and $\mathcal{K}^{\mathsf{G}}$, which induces an $\mathcal{M}$-definable isomorphism between $\mathcal{M}^{\mathsf{UV}}$ $(= \mathcal{K}^{\mathsf{V}})$ and $\mathcal{N}^{\mathsf{UV}}$ $(= \mathcal{G}^{\mathsf{V}})$. However, $\mathcal{M}^{\mathsf{UV}}$ is $\mathcal{M}$-definably isomorphic to $\mathcal{M}$, and $\mathcal{N}^{\mathsf{UV}}$ is $\mathcal{N}$-definably (hence, $\mathcal{M}$-definably) isomorphic to $\mathcal{N}$, so there is an $\mathcal{M}$-definable isomorphism between $\mathcal{M}$ and $\mathcal{N}$.

□

4.3 A solid theory not interpreting PA

The solid theories T_n constructed in Section 4.1 are weak in the sense that they do not prove PA. However, they are relatively strong in the sense of interpretability, and in particular they clearly interpret PA – for each n the $\mathcal{L}_{\mathrm{PA}}$-part of N_n interprets PA in $\mathrm{IT}(n)$, and then an interpretation of PA in T_n is obtained by an obvious case distinction.

So, it is quite natural to ask whether there are solid subtheories of PA that are also strictly weaker than PA in terms of interpretability. This subsection contains a proof of the following theorem, which provides a positive answer to that question.

Theorem 4.15. *For every $n \in \mathbb{N}$, there exists a r.e. solid subtheory of* PA *that contains* $\mathrm{I}\Sigma_n$ *but not* $\mathrm{B}\Sigma_{n+1}$ *and that does not interpret* PA.

Let $\mathrm{IT}(n)$ be defined as in Section 4.1. We observe that

Corollary 4.16. *For each $n \geq 1$, there is no Σ_n-restricted interpretation of* PA *in* $\mathrm{IT}(n)$.

Proof. Assume that M is a Σ_n-restricted interpretation of $\mathrm{I}\Delta_0 + \exp$ in $\mathrm{IT}(n)$. By definition, $\mathrm{IT}(n)$ contains $\mathrm{I}\Sigma_n$, and by the discussion from Section 4.1, it also proves $\exists x\, \xi_n(x)$ for the the Σ_1-flexible formula ξ_n. So, by Corollary 3.4, we have $\mathrm{IT}(n) \vdash (\exists x\, \xi_n(x))^{\mathsf{M}}$. But $\mathrm{PA} \vdash \mathrm{Con}(\mathrm{I}\Sigma_n)$, whereas $\mathrm{I}\Delta_0 + \exp + \exists x\, \xi_n(x) \vdash \neg\mathrm{Con}(\mathrm{I}\Sigma_n)$ by Corollary 3.7. Hence, M is not an interpretation of PA. $\square$

Now, the intuition is that the theory we are looking for is roughly the following one ("p" stands for "proto-"):

$$\mathrm{pT}_n := \mathrm{I}\Sigma_n + \exp + \bigoplus_{k \geq n} (\mathrm{IT}(k) | \mathrm{I}\Sigma_k \wedge \neg\, \mathrm{I}\Sigma_{k+1}).$$

Lemma 4.17. *The theory* pT_n *does not interpret* PA.

Proof. Suppose that M is an interpretation of PA in pT_n. Let $k \geq n$ be such that M is Σ_k-restricted. Then M is also a Σ_k-restricted interpretation of PA in the consistent theory $\mathrm{pT}_n + \mathrm{I}\Sigma_k + \neg\, \mathrm{I}\Sigma_{k+1}$. However, the latter theory coincides with $\mathrm{IT}(k)$, which contradicts Corollary 4.16. $\square$

The problem with pT_n is that the family $\{\mathrm{IT}(k)\}_{k\in\omega}$ is not retract-disjoint. As a result, there are models $\mathcal{M}, \mathcal{N} \vDash \mathrm{pT}_n$ such that $\mathcal{M}$ is a retract of $\mathcal{N}$ but the two structures satisfy pT_n "for different reasons" and thus cannot be isomorphic. For example, consider the models of $\mathrm{IT}(2)$ and $\mathrm{IT}(3)$, respectively, obtained by applying the interpretations K_2 and K_3 from Section 4.1 in $(\mathbb{N}, \mathrm{Th}(\mathbb{N}))$. These models are not even elementarily equivalent, but they are both bi-interpretable with $(\mathbb{N}, \mathrm{Th}(\mathbb{N}))$, so they are not only retracts but in fact bi-interpretable with one another.

We will now show how to improve the theories $\mathrm{IT}(n)$ so as to eliminate such patterns. Below we use the same interpretations K_n and N_n as before, in Section 4.1.

Define the theory W_n by

$$W_n := \{\sigma \in \mathcal{L}_{\mathrm{PA}} : \quad \mathrm{IT}(n) \vdash \sigma^{\mathsf{N}_n}\},$$

and let $W = \bigcup_{n\in\omega} W_n$. Since $\mathbb{N} \vDash W_n$ for each n, we know that W is a consistent theory. It is also clearly r.e. Let $\zeta(x)$ be a Σ_1-flexible formula over W. Let $\mathrm{ITD}(n) := \mathrm{IT}(n) + (\forall x\, (\zeta(x) \leftrightarrow x = \underline{n}))^{\mathsf{N}_n}$ (here "D" stands for "disjoint"). By flexibility of ζ, the theory $\mathrm{ITD}(n)$ is consistent for each n.

Lemma 4.18. *For each $n \in \omega$, the theory* $\mathrm{ITD}(n)$ *is bi-interpretable with* $\mathrm{CT}[\mathrm{PA}] + \forall x\,(\zeta(x) \leftrightarrow x = n)$.

Proof. As in the proof of Proposition 4.7, we once again use the interpretations N_n and K_n. By the definition of $\mathrm{ITD}(n)$, these interpretations witness the mutual interpretability of $\mathrm{ITD}(n)$ and the extension of $\mathrm{CT}[\mathrm{PA}]$ by $\forall x\,(\zeta(x) \leftrightarrow x = \underline{n})$. The argument that they actually witness bi-interpretability as well is the same as before. □

Since bi-interpretability preserves solidity and since by Corollary 3.10 each extension of CT[PA] in the same language is solid, we obtain:

Corollary 4.19. *For each n, the theory* $\mathrm{ITD}(n)$ *is solid.*

One last missing piece is the retract-disjointness of the theories $\mathrm{ITD}(n)$.

Corollary 4.20. *The family* $\{\mathrm{ITD}(n)\}_{n\in\omega}$ *is retract-disjoint.*

Proof. By Corollary 4.13 and Lemma 4.18, it is enough to check that the family $\{V_n\}_{n\in\omega}$ defined by

$$V_n := \mathrm{CT}[\mathrm{PA}] + \forall x\,(\zeta(x) \leftrightarrow x = \underline{n})$$

is retract-disjoint. However, this easily follows from the solidity of CT[PA]. □

Finally, we define TD_n to be

$$\mathrm{I}\Sigma_n + \exp + \bigoplus_{k\geq n}(\mathrm{ITD}(k)|\mathrm{I}\Sigma_k \wedge \neg\,\mathrm{I}\Sigma_{k+1}).$$

We observe that the axioms of $\mathrm{ITD}(k)$ can be effectively generated given k, so TD_n really is an r.e. subtheory of PA. By repeating the argument from Lemma 4.17 (which requires using the obvious variant of Corollary 4.16 for $\mathrm{ITD}(k)$ instead of $\mathrm{IT}(k)$), we obtain:

Corollary 4.21. *The theory TD_n does not interpret* PA.

Proof of Theorem 4.15. We have already observed that TD_n is an r.e. subtheory of PA, and in Corollary 4.21 we have shown that it does not interpret PA. Clearly, TD_n contains $\mathrm{I}\Sigma_n$ by definition. So, it remains to show that TD_n is solid.

To this end, we want to invoke Proposition 4.10 with $U_k := \mathrm{ITD}(k)$, $V := \mathrm{I}\Sigma_n + \exp$, and $\varphi_k := \mathrm{I}\Sigma_k \wedge \neg\,\mathrm{I}\Sigma_{k+1}$ (where in φ_0 we use a finite fragment of $\mathrm{I}\Delta_0$ that is equivalent to $\mathrm{I}\Delta_0$ assuming exp). By Corollary 4.19, each theory U_k is solid. The sentences φ_k are obviously pairwise inconsistent. It is easy to see that $V \cup \{\neg\varphi_k : k \in \omega\}$ is simply PA, hence it is also a solid theory. The last thing we have to verify is that the family consisting of PA and of $\mathrm{ITD}(0), \mathrm{ITD}(1), \ldots$ is retract-disjoint. By Lemma 4.20, it is enough to check that each two-element family consisting of PA and a single theory $\mathrm{ITD}(k)$ is retract-disjoint. This follows by Lemma 4.18, Corollary 4.13, and Lemma 4.6. □

Remark. Note that each model of the theory TD_n actually interprets a model of PA. Intuitively, the reason why these interpretations cannot be merged into a single interpretation of PA in TD_n is that TD_n is defined by an "infinite case distinction", and a finite formula is unable to decide which of the cases applies. We return to this topic in the final section of the paper.

4.4 A solid theory infinitely below PA

This section improves on our main result from Section 4.1 by providing an example of a solid subtheory of PA which does not not imply PA even after being strengthened by an arbitrary true sentence.

Theorem 4.22. *For each $n \in \omega$, there is a solid r.e. subtheory TF_n of* PA *such that $TF_n \vdash \mathrm{I}\Sigma_n + \exp$, but for each k it holds that $TF_n + \mathrm{Th}_{\Pi_k}(\mathbb{N}) \nvdash \mathrm{PA}$.*

One of the motivations for this theorem is a result of Wilkie's that we will now explain. An $\mathcal{L}_{\mathrm{PA}}$ scheme template is an $\mathcal{L}_{\mathrm{PA}}$-formula $\theta(P)$ with a marked fresh predicate letter P. For a first-order formula $\psi(x)$, the notation $\theta[\psi/P]$ stands for the formula obtained by replacing each occurrence of a subformula $P(t)$ with $\psi(t)$ (preceded by renaming the bound variables if necessary). The scheme generated by the template $\theta(P)$, denoted by $\theta[\mathcal{L}_{\mathrm{PA}}]$, is the set $\{\theta[\psi/P] : \psi(x) \in \mathcal{L}_{\mathrm{PA}}\}$ A scheme template $\theta(P)$ is *restricted* if it is of the form

$$Q_0 x_0 \in P\, Q_1 x_1 \in P \ldots Q_n x_n \in P\, \varphi(x_0, \ldots, x_n),$$

where P does not occur in $\varphi(x_0, \ldots, x_n)$ and $Q_i x_i \in P\, \psi$ denotes either $\exists x_i\, (P(x_i) \wedge \psi)$ or $\forall x_i\, (P(x_i) \to \psi)$. We say that $\theta(P)$ is *second-order categorical* if $(\mathbb{N}, \mathcal{P}(\mathbb{N})) \models \forall P\, \theta(P)$ and for every $\mathcal{L}_{\mathrm{PA}}$-structure $\mathcal{M}$, if $(\mathcal{M}, \mathcal{P}(M)) \models \forall P\, \theta(P)$, then $\mathcal{M}$ is isomorphic to $\mathbb{N}$.

Theorem 4.23 (Wilkie [29])**.** *Let $\theta(P)$ be a restricted $\mathcal{L}_{\mathrm{PA}}$ scheme template which is second-order categorical. Then there is a true $\mathcal{L}_{\mathrm{PA}}$ sentence φ such that*

$$\theta[\mathcal{L}_{\mathrm{PA}}] + \varphi \vdash \mathrm{PA}\,.$$

Recall that a theory being solid is, in a loose sense, a categoricity property. Theorem 4.22 shows that a natural variant of Wilkie's result with "solid theory" replacing "restricted second-order categorical scheme" is false.

Our strategy for the proof of Theorem 4.15 is similar to the one used to prove Theorem 4.15 in the previous subsection. We will define a retract-disjoint family of theories $\mathrm{ITF}(n)$ with the following additional property: for every n, $\mathrm{ITF}(n)$ is consistent with $\mathrm{Th}_{\Pi_{n-1}}(\mathbb{N}) + \neg\, \mathrm{I}\Sigma_{n+1}$. We observe that the theories $\mathrm{ITD}(n)$ do not have this consistency property because they all imply the false Σ_1 statement $\neg\mathrm{Con(PA)}$ (additionally, each $\mathrm{ITD}(n)$ implies the Σ_1 statement $\exists x\, \xi_n(x)$ which is inconsistent with PA). Our way of making the family $\{\mathrm{ITF}(n)\}_{n\in\omega}$ retract-disjoint will also be different from the one in Section 4.3.

Below we describe the construction of $\mathrm{ITF}(n)$. As in the case of $\mathrm{IT}(n)$, we first describe the construction of the "intended model" of $\mathrm{ITF}(k)$ and then extract its relevant properties in the form of axioms.

The construction of $\mathrm{ITF}(n)$**.** Let ζ_n be a Σ_{n+1} formula that is (Σ_{n-1}, Σ_1)-flexible over $\mathrm{I}\Sigma_{n+1}$ and witnesses Theorem 3.6(b). We perform a procedure of finding a Henkin structure $\mathcal{H}$ and its pointwise Σ_{n+1}-definable substructure $\mathcal{K}$ similar to the one described by the interpretation K_n from Section 4.1, except that now instead of the theory from (4) we start with:

$$\mathrm{I}\Sigma_{n+1} \cup \{\zeta_n(\underline{k}) : P_n(\underline{k})\} \cup \{\neg\zeta_n(\underline{\ell}) : \neg P_n(\underline{\ell})\}, \tag{5}$$

and we also want to ensure that $\mathcal{H}$ satisfies all true Π_{n-1} sentences. Here, P_n is the highest-level truth predicate of $\mathrm{CT}^n[\mathrm{PA}]$. In particular, if this construction is carried out in the standard model of $\mathrm{CT}^n[\mathrm{PA}]$, i.e. in the n-th element of the sequence defined recursively by:

$$\mathcal{CT}_0 := \mathbb{N}$$
$$\mathcal{CT}_{j+1} := (\mathbb{N}, \mathrm{Th}(\mathcal{CT}_0), \ldots, \mathrm{Th}(\mathcal{CT}_j))$$

then ζ_n encodes the theory of $\mathcal{CT}_{n-1}$. We note that the argument that the theory in (5) is consistent formalizes in $\mathrm{CT}^n[\mathrm{PA}]$ thanks to Theorem 3.6.

We will use the notation $\mathsf{K}'_n, \mathsf{N}'_n$ for analogues of the interpretations $\mathsf{K}_n, \mathsf{N}_n$ from Section 4.1 adapted to the current setting. So, K'_n describes the procedure of constructing the model $\mathcal{K}_n := \mathcal{K}^{n+1}(\mathcal{H})$, where $\mathcal{H}$ is the Henkin model a theory obtained by taking the theory in (5) extended by $\mathrm{Th}_{\Pi_{n-1}}(\mathbb{N})$, henkinizing it, and taking the leftmost completion. As in the case of $\mathrm{IT}(n)$, we have a canonical definition δ_n that isolates the smallest definable cut in $\mathcal{K}_n$ (which is the standard cut if we apply K'_n in $\mathcal{CT}_n$). The interpretation N'_n of $\mathrm{CT}^n[\mathrm{PA}]$ in $\mathcal{K}_n$ has universe defined by δ_n, arithmetical operations given by restricting $+$ and $\times$ to δ_n, and the n-th truth predicate P_n given by ζ_n; the predicates $P_1, \ldots, P_{n-1}$ are determined by P_n.

With the above definitions, $\mathrm{ITF}(n)$ is defined as $\mathrm{IT}(n)$ is Section 4.1, except that $\mathsf{K}_n, \mathsf{N}_n$ are changed to $\mathsf{K}'_n, \mathsf{N}'_n$, and "$\mathsf{N}_n \vDash \mathrm{CT}[\mathrm{PA}]$" in (iii) is replaced by

(iii)$'$ $\mathsf{N}'_n \vDash \mathrm{CT}^n[\mathrm{PA}]$.

Lemma 4.24. *For each $n \geq 1$, $\mathrm{ITF}(n)$ is consistent with $\mathrm{Th}_{\Pi_{n-1}}(\mathbb{N}) + \neg\, \mathrm{B}\Sigma_{n+1}$.*

Proof. This is essentially the same argument as the verification that $\mathrm{IT}(n)$ is consistent, adapted to the slightly more complicated setting.

Fix $n \geq 1$ and consider the standard model $\mathcal{CT}_n$ of $\mathrm{CT}^n[\mathrm{PA}]$. By compactness and the (Σ_{n-1}, Σ_1)-flexibility of ζ_n we conclude that the $\mathcal{L}_{P_n}$-definable theory

$$\mathrm{I}\Sigma_{n+1} \cup \{\zeta_n(\underline{k}), \neg\zeta_n(\underline{\ell}) : P_n(\underline{k}), \neg P_n(\underline{\ell}), k, \ell \in \omega\} \cup \{\varphi \in \Pi_{n-1} : \mathbb{N} \vDash \varphi\}$$

is consistent. Applying to this theory the henkinization and completion process described above and codified in K'_n, we obtain a $\mathcal{CT}_n$-definable Henkin model $\mathcal{H} \vDash \mathrm{I}\Sigma_{n+1}$ such that $\mathbb{N} \preceq_{\Sigma_{n-1}} \mathcal{H}$ and for each natural number k, we have that $\zeta_n(\underline{k})$ holds in $\mathcal{H}$ iff $P_n(\underline{k})$ holds in $\mathcal{CT}_n$.

By the usual arguments, $\mathcal{K}^{n+1}(\mathcal{H}) \preceq_{\Sigma_{n+1}} \mathcal{H}$, hence $\mathcal{K}^{n+1}(\mathcal{H}) \vDash \mathrm{I}\Sigma_n + \exp + \mathrm{Th}_{\Pi_{n-1}}(\mathbb{N})$ and, since ζ_n is a Σ_{n+1} formula of $\mathcal{L}_{\mathrm{PA}}$, for each $k \in \omega$ we have

$$\mathcal{K}^{n+1}(\mathcal{H}) \vDash \zeta_n(\underline{k}) \text{ iff } \mathcal{CT}_n \vDash P_n(\underline{k}).$$

With the above, it is now routine to verify that indeed $\mathcal{K}^{n+1}(\mathcal{H}) \vDash \mathrm{ITF}(n)$. □

Lemma 4.25. *For each $n \geq 1$, $\mathrm{ITF}(n)$ is bi-interpretable with $\mathrm{CT}^n[\mathrm{PA}]$.*

Proof. This can be shown essentially as in the proof of Proposition 4.7. The main change is that instead of a formalization of the arguments from Section 4.1 in $\mathrm{CT}[\mathrm{PA}]$, we use a formalization of the proof of Lemma 4.24 in $\mathrm{CT}^n[\mathrm{PA}]$. The verification that ζ_n has the required (Σ_{n-1}, Σ_1)-flexibility property over $\mathrm{I}\Sigma_{n+1}$ can be carried out in $\mathrm{CT}^n[\mathrm{PA}]$ thanks to Theorem 3.6(b) and the fact that $\mathrm{CT}^n[\mathrm{PA}] \vdash \Sigma_{n+2}\text{-}\mathrm{RFN}(\mathrm{I}\Sigma_{n+1})$. □

Corollary 4.26. *For each $n \geq 1$, the theory $\mathrm{ITF}(n)$ is solid.*

Proof. By Lemma 4.25, Corollary 3.10, and the fact that bi-interpretability preserves solidity. $\square$

Corollary 4.27. *The family* $\{\mathrm{PA}\} \cup \{\mathrm{ITF}(n)\}_{n\in\omega}$ *is retract-disjoint.*

Proof. By Lemma 4.25 and Corollary 4.13. $\square$

Proof of Theorem 4.22. Let TF_n be $\mathrm{I}\Sigma_n + \exp + \bigoplus_{k\geq n}(\mathrm{ITF}(k)|\mathrm{I}\Sigma_k \wedge \neg\,\mathrm{I}\Sigma_{k+1})$. Clearly, TF_n is an r.e. subtheory of PA, and it contains $\mathrm{I}\Sigma_n$ by definition.

The fact that $TF_n + \mathrm{Th}_{\Pi_k}(\mathbb{N})$ does not imply PA for any k follows immediately from Lemma 4.24.

To prove solidity of TF_n, we invoke Proposition 4.10 with $U_k := \mathrm{ITF}(k)$, $V := \mathrm{I}\Sigma_n + \exp$, and $\varphi_k := \mathrm{I}\Sigma_k \wedge \neg\mathrm{I}\Sigma_{k+1}$ (with $\mathrm{I}\Delta_0$ replaced its a sufficiently large finite fragment as before). By Corollary 4.26, each U_k is solid, and $V \cup \{\neg\varphi_k : k \in \omega\}$ is solid because it is equivalent to PA. The sentences φ_k are pairwise inconsistent. Finally, the family $\mathrm{PA} \cup \{U_k\}_{k\in\omega}$ is retract-disjoint by Corollary 4.27. $\square$

Remark. We can combine Theorems 4.15 and 4.22 in the following sense. Suppose that $\{\mathrm{ITD}(n)'\}_{n\in\omega}$ is a sequence of theories which is produced as $\{\mathrm{ITD}(n)\}_{n\in\omega}$, but with the change that instead of using the theory W and formula ζ from Section 4.3, we ensure retract-disjointness by making each $\mathrm{ITD}(n)'$ bi-interpretable with $\mathrm{CT}^n[\mathrm{PA}]$. That is, to define $\mathrm{ITD}(n)'$ we essentially repeat the construction of $\mathrm{IT}(n)$ from Section 4.1 but working with P_n instead of P and N'_n instead of N_n; unlike for $\mathrm{ITF}(n)$, we use the basic flexible formula ξ rather than ζ_n and do not include Π_{n-1}-truth in the theory of the Henkin model.

Then if we set $U_{2n} := \mathrm{ITF}(2n)$ and $U_{2n+1} := \mathrm{ITD}(2n+1)'$, one can easily check that for every $n \in \mathbb{N}$ the theory

$$\mathrm{I}\Sigma_n + \exp + \bigoplus_{k\geq n}(U_k|\mathrm{I}\Sigma_k \wedge \neg\,\mathrm{I}\Sigma_{k+1})$$

is a proper solid subtheory of PA that is both unable to interpret PA and "infinitely below" PA in the sense of Theorem 4.22.

5 Separation theorems

We now focus on separating the categoricity-like properties considered in this paper. In particular, we obtain a separation of tightness from neatness, and a nontrivial (e.g., not based on a complete theory) example separating neatness from semantical tighness and solidity.

In most of our constructions, we exploit the fact that an actually existing isomorphism that would be needed to witness one of the properties we study or the failure of another is somehow difficult to express. In some cases, the isomorphism cannot be defined internally in a structure, in others, its definition requires parameters from the structure.

Some of our nontrivial examples take the form $(T_1|\neg\psi) \bigoplus (T_2|\psi)$ for a sentence ψ. To avoid using such cumbersome notation, we will write $T_1 \oplus_\psi T_2$ in its stead.

5.1 Tame separators

We first discuss two simple examples of separations between the syntactic and the semantical notions, based on the fact that any complete theory has to be neat. The theories in these examples have the virtue of being r.e., but they do not interpret any arithmetic at all; in particular, they are not sequential.

Proposition 5.1. DLO *is neat and therefore tight, but it is not semantically tight.*

Proof. Note first that DLO is clearly neat because, it is complete. To show that it is not semantically tight, let $\mathcal{M} = \langle \mathbb{Q}, \leq \rangle + \langle \mathbb{R}, \leq \rangle$ and let $\mathcal{N} = \langle \mathbb{R}, \geq \rangle + \langle \mathbb{Q}, \geq \rangle$, where $+$ stands for the disjoint sum of linear orders. Since the order on $\mathcal{N}$ is just the inverse of the order on $\mathcal{M}$, the two structures are clearly bi-interpretable. However, $\mathcal{M}$ is not isomorphic to $\mathcal{N}$, so DLO fails to be semantically tight. $\square$

Remark. As another example in this spirit that additionally illustrates the role of multi-dimensional interpretations and the subtleties involved in defining semantical tightness, consider the theory of infinite sets, in the empty language. Just like DLO, this theory is complete and hence trivially neat.

We show that this theory is not semantically tight if one is willing to allow multi-dimensional interpretations. Let $\mathcal{M} = \mathbb{N}$, $\mathcal{N} = \mathbb{N} \setminus \{0\}$, $\mathcal{M}^* = \{\langle n, n \rangle : n \geq 1\} \cup \{\langle 1, 2 \rangle\}$, $\mathcal{N}^* = \{\langle n, n \rangle : n \geq 1\}$. Then we have three interpretations witnessing $\mathcal{M} \rhd \mathcal{N} \rhd \mathcal{M}^* \rhd \mathcal{N}^*$. The interpretations are defined in the obvious way, though the one in of $\mathcal{M}^*$ in $\mathcal{N}$ is two-dimensional, and they all use parameters, namely 0; $1, 2$; and $\langle 1, 2 \rangle$, respectively, in order to exclude the appropriate elements from the domain. There is an $\mathcal{M}$-definable isomorphism between $\mathcal{M}$ and $\mathcal{M}^*$ (map 0 to $\langle 1, 2 \rangle$, and any other n to $\langle n, n \rangle$), as well as an $\mathcal{N}$-definable isomorphism between $\mathcal{N}$ and $\mathcal{N}^*$ (map n to $\langle n, n \rangle$). Thus, we get a bi-interpretation between $\mathcal{M}$ and $\mathcal{N}$. But there is no $\mathcal{M}$-definable bijection between $\mathcal{M}$ and $\mathcal{N}$, because, by quantifier elimination, any definable injection from $\mathcal{M}$ to $\mathcal{M}$ has the following form: an arbitrary permutation of a finite set (the set of parameters involved in the definition), and the identity on all other elements.

The argument from the previous argument breaks down if in the definition of semantical tightness we only require isomorphism instead of $\mathcal{M}$-definable isomorphism: indeed, any two bi-interpretable infinite sets must have the same cardinality and thus be isomorphic. It also breaks down if we only allow one-dimensional interpretations as in the rest of this paper.

5.2 Tight but not neat

This subsection is devoted to the proof of a less trivial separation, between the two syntactic notions of tightness and neatness. In fact, we prove that there are arbitrarily strong subtheories of PA that are tight but not neat. We do not know whether the theories in question are semantically tight.

Theorem 5.2. *For every $n \geq 1$, there exists an r.e. subtheory of* PA *that contains* $\mathrm{B}\Sigma_n + \exp$ *but not* $\mathrm{I}\Sigma_n$ *and is tight but not neat.*

The overall structure of the argument is similar to that of Section 4.1, though the role of CT[PA] is played by PA, and pointwise definable models of $\mathrm{I}\Sigma_n + \neg\,\mathrm{B}\Sigma_{n+1}$ are replaced by models of $\mathrm{B}\Sigma_n + \neg\,\mathrm{I}\Sigma_n$.

Fix $n \geq 1$. In analogy to the theories T_n and $\mathrm{IT}(n)$ from the proof of Theorem 4.1 in Section 4.1, we will use the symbol S_n to denote the theory that will eventually witness Theorem 5.2, and we will define an auxiliary theory $\mathrm{IS}(n)$.

In Section 4.1, we had an interpretation K_n of a model of $\mathrm{I}\Sigma_n + \neg\,\mathrm{B}\Sigma_{n+1}$ in $(\mathbb{N}, \mathrm{Th}(\mathbb{N}))$, and an inverse interpretation N_n of $(\mathbb{N}, \mathrm{Th}(\mathbb{N}))$ in that model. The theory $\mathrm{IT}(n)$ axiomatized many features of our particular model of $\mathrm{I}\Sigma_n + \neg\,\mathrm{B}\Sigma_{n+1}$, and T_n said that we are either in a model of $\mathrm{IT}(n)$ or in one of PA. This time, K_n will be replaced by a parameter-free interpretation J_n of a model of $\mathrm{B}\Sigma_n + \neg\,\mathrm{I}\Sigma_n$ in $\mathbb{N}$, and again there will be an inverse

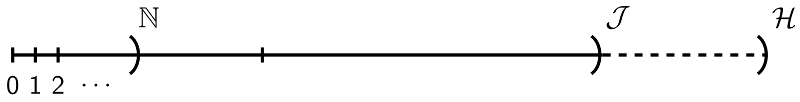


Figure 5.1: Construction of the model $\mathbb{N}^{\mathsf{J}_n}$ of $\mathrm{IS}(n)$. The solid horizontal lines represent $\mathbb{N}^{\mathsf{J}_n}$, which is a nonstandard Σ_{n+1}-elementary initial segment $\mathcal{J}$ of the Henkin structure $\mathcal{H}$. The dashed horizontal lines represent the rest of $\mathcal{H}$.

interpretation of $\mathbb{N}$ in our model. In fact, we will reuse the name N_n for that inverse interpretation, because it will essentially do the same job as before – pick out the smallest definable cut of our model – except that there will be no need to define the truth predicate. As before, $\mathrm{IS}(n)$ will axiomatize some properties of our model, and S_n will say that we are either in a model of $\mathrm{IS}(n)$ or in one of PA.

The interpretation J_n describes the following process, as carried out in $\mathbb{N}$:

- Consider a canonically defined binary tree whose paths correspond to complete consistent henkinized extensions of the theory $\mathrm{I}\Sigma_n + \neg\mathrm{Con}(\mathrm{I}\Sigma_n)$.
- Take the Henkin model $\mathcal{H}$ given by the leftmost path through that tree.
- Take the initial segment of $\mathcal{H}$ generated by the first $\mathbb{N}$ iterations of the witness-bounding function for the universal Σ_{n-1} formula (see discussion at the end of Section 2) on the smallest proof of inconsistency in $\mathrm{I}\Sigma_n$. For $n = 1$, instead of the witness-bounding function consider the first $\mathbb{N}$ iterations of exp.

This process is illustrated in Figure 5.1. As discussed at the end of Section 2, it produces a Σ_{n-1}-elementary cut $\mathcal{J}$ of $\mathcal{H}$ that is necessarily a proper cut, because $\mathcal{H}$ is nonstandard, and in a model of $\mathrm{I}\Sigma_n$ the witness-bounding function for a Σ_{n-1} formula can be iterated an arbitrary number of times (and so can exp in a model of $\mathrm{I}\Sigma_1$). Thus, $\mathcal{J}$ is a model of $\mathrm{B}\Sigma_n$. Moreover, $\mathcal{J}$ is a model of $\neg\,\mathrm{I}\Sigma_n$, because it is a nonstandard structure in which the standard cut $\mathbb{N}$ is Σ_n-definable, say by the formula $\delta_n(x)$ expressing "there exists an inconsistency proof for $\mathrm{I}\Sigma_n$, and on that proof the witness-bounding function for the universal Σ_{n-1} formula can be iterated x times".

Thus, $\mathbb{N}$ is the smallest definable cut of $\mathcal{J}$, and it can be interpreted in $\mathcal{J}$ by the interpretation N_n in which the domain is defined by δ_n and the arithmetical operations are unchanged. As in the proof of Theorem 4.1, there is an $\mathbb{N}$-definable isomorphism j_n between $\mathsf{J}_n\mathsf{N}_n$ and the identity interpretation of $\mathbb{N}$ in itself, namely the map from Lemma 3.2: take $x \in \mathbb{N}$ to the x-th smallest element of $\mathcal{J}$. Each of $\mathsf{N}_n, \mathsf{J}_n, j_n$ is definable without parameters.

However, we no longer have a $\mathcal{J}$-definable isomorphism between $\mathsf{N}_n\mathsf{J}_n$ and the identity interpretation of $\mathcal{K}$ in itself. The reason is that $\mathcal{J}$ is a model of $\mathrm{B}\Sigma_n + \exp + \neg\,\mathrm{I}\Sigma_n$, and the domain of N_n is a proper cut in it. It is known that models of $\mathrm{B}\Sigma_n + \exp + \neg\,\mathrm{I}\Sigma_n$ cannot have a definable injective multifunction into a proper initial segment:

Theorem 5.3. *[17] Let the cardinality scheme CARD say that no formula defines an injective multifunction from the universe into $[0, x]$ for any number x. Then, for each $n \geq 1$, it holds that $\mathrm{B}\Sigma_n + \exp + \neg\,\mathrm{I}\Sigma_n \vdash CARD$.*

On the other hand, the structure produced by $\mathsf{N}_n\mathsf{J}_n$ is in fact isomorphic to $\mathcal{J}$, even though $\mathcal{J}$ does not see the isomorphism. In particular, the two structures are elementarily equivalent.

We let $\mathrm{IS}(n)$ axiomatize the properties of $\mathcal{J}$ discussed above. The axioms of $\mathrm{IS}(n)$ are:

(i) $\mathrm{B}\Sigma_n + \exp + \neg\,\mathrm{I}\Sigma_n$,

(ii) "δ_n defines a cut which is the smallest definable cut",

(iii) $\psi^{\mathsf{N}_n \mathsf{J}_n} \leftrightarrow \psi$, for each sentence ψ,

(iv) $\mathsf{N}_n \vDash$ "$j_n \colon \mathsf{id} \to \mathsf{J}_n\mathsf{N}_n$ is an isomorphism".

We let S_n be $\mathrm{IS}(n) \oplus_{\mathrm{I}\Sigma_n} \mathrm{PA}$. Note that it follows from axioms (ii) of $\mathrm{IS}(n)$ that N_n is an interpretation of PA in $\mathrm{IS}(n)$.

Lemma 5.4. *The theory S_n contains $\mathrm{B}\Sigma_n + \exp$ but not $\mathrm{I}\Sigma_n$. Thus, it is a proper subtheory of* PA*.*

Proof. The argument is analogous to the one for Lemma 4.4: by the construction of the model $\mathcal{J}$ described above, $\mathrm{IS}(n)$ is consistent and contains $\mathrm{B}\Sigma_n + \exp$ but contradicts $\mathrm{I}\Sigma_n$. □

Lemma 5.5. *The theory S_n is not neat.*

Proof. Consider $U = \mathrm{PA}$ and $V = \mathrm{IS}(n)$. Note that both these theories extend S_n. Moreover, J_n is an interpretation of $\mathrm{IS}(n)$ in PA, and $\mathsf{J}_n\mathsf{N}_n$ is an interpretation of PA in PA. By design, the latter interpretation is PA-provably isomorphic to the identity interpretation: J_n is the shortest initial segment of $\mathcal{H}$ which contains all the finite (in the sense of the ground model) iterations of the witness-bounding function for the universal Σ_{n-1} formula on the smallest witness to $\neg\mathrm{Con}(\mathrm{I}\Sigma_n)$, and N_n isolates precisely the set of those numbers a for which that function can be iterated a-times. Thus, PA is a retract of $\mathrm{IS}(n)$. Clearly, however, $\mathrm{IS}(n)$ is not a subtheory of PA. □

Lemma 5.6. *The theory $\mathrm{IS}(n)$ is neat.*

Proof. It is enough to prove that if $\mathcal{M}_1 \rhd \mathcal{M}_2 \rhd \mathcal{M}_3$ are models of $\mathrm{IS}(n)$ and there is an $\mathcal{M}_1$-definable isomorphism from $\mathcal{M}_1$ onto $\mathcal{M}_3$, then $\mathcal{M}_1 \equiv \mathcal{M}_2$. So, let $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ be as described.

For each $i \in \{1,2,3\}$, let $\mathcal{N}_i$ be the model of PA obtained by applying the interpretation N_n in $\mathcal{M}_i$, and let $\mathcal{M}_i'$ be the model of $\mathrm{IS}(n)$ obtained by applying J_n in $\mathcal{N}_i$. See Figure 5.2. Note that the domain of each $\mathcal{N}_i$ is the smallest definable cut of $\mathcal{M}_i$ and that there is an $\mathcal{M}_1$-definable isomorphism from $\mathcal{N}_1$ onto $\mathcal{N}_3$ (induced by the $\mathcal{M}_1$-definable isomorphism between $\mathcal{M}_1$ and $\mathcal{M}_3$).

By axioms (ii) of $\mathrm{IS}(n)$ and Lemma 3.9 we have an ($\mathcal{M}_1$-definable) isomorphism between $\mathcal{N}_1$ and $\mathcal{N}_2$. This clearly gives rise to an ($\mathcal{M}_1$-definable) isomorphism between $\mathcal{M}_1'$ and $\mathcal{M}_2'$.

By axioms (iii) of $\mathrm{IS}(n)$, we know that $\mathcal{M}_1 \equiv \mathcal{M}_1'$ and $\mathcal{M}_2 \equiv \mathcal{M}_2'$. Since $\mathcal{M}_1'$ and $\mathcal{M}_2'$ are isomorphic, this implies that $\mathcal{M}_1 \equiv \mathcal{M}_2$. □

Proof of Theorem 5.2. We have already shown in Lemmas 5.4 and 5.5 that S_n is a subtheory of PA containing $\mathrm{B}\Sigma_n + \exp$ but not $\mathrm{I}\Sigma_n$, and that S_n is not neat. Clearly, S_n is an r.e. theory, so it remains to prove that it is tight.

It is enough to show that if $\mathcal{M}_1$ is a model of S_n and $(\mathsf{M}_2, \mathsf{M}_1)$ is a bi-interpretation in $\mathcal{M}_1$, then $\mathcal{M}_1 \equiv \mathcal{M}_1^{\mathsf{M}_2}$. Put $\mathcal{M}_2 = \mathcal{M}_1^{\mathsf{M}_2}$.

By the definition of S_n, each $\mathcal{M}_i$ satisfies either PA or $\mathrm{IS}(n)$. If $\mathcal{M}_1$ and $\mathcal{M}_2$ both satisfy PA or both satisfy $\mathrm{IS}(n)$, then $\mathcal{M}_1 \equiv \mathcal{M}_2$ follows from the solidity of PA or the proof of Lemma 5.6, respectively.

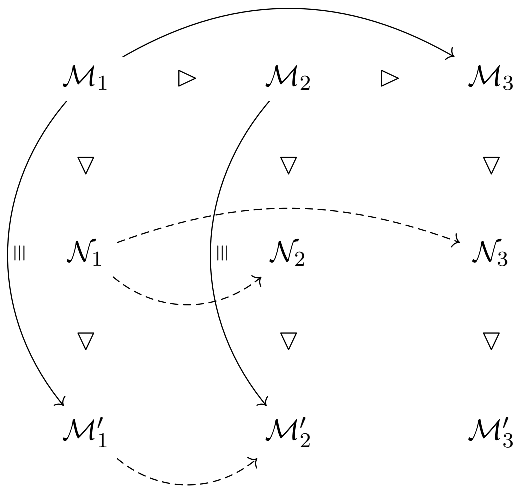


Figure 5.2: The proof of Lemma 5.6. The horizontal solid arrow represents an isomorphism given directly by the assumptions about $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$, and the dashed arrows represent isomorphisms shown to exist during the argument. The vertical solid arrows stand for elementary equivalences, which together with the isomorphisms let us conclude that $\mathcal{M}_1$ is elementarily equivalent to $\mathcal{M}_2$.

The remaining case is that exactly one of $\mathcal{M}_1, \mathcal{M}_2$ satisfies PA. Assume w.l.o.g. that $\mathcal{M}_1 \models \mathrm{IS}(n)$ and $\mathcal{M}_2 \models \mathrm{PA}$. We will show that this leads to a contradiction, which will complete the proof of the theorem.

Let $\mathcal{M}_3 = \mathcal{M}_2^{\mathsf{M}_1}$ be the structure interpreted in $\mathcal{M}_2$ which is $\mathcal{M}_1$-definably isomorphic to $\mathcal{M}_1$. That isomorphism between $\mathcal{M}_1$ and $\mathcal{M}_3$, say f_1, may in particular be viewed as an $\mathcal{M}_1$-definable injective multifunction from $\mathcal{M}_1$ into $\mathcal{M}_2$.

By Lemma 3.2, there is an $\mathcal{M}_2$-definable (hence $\mathcal{M}_1$-definable) embedding f_2 from $\mathcal{M}_2$ into the initial segment $(\delta_n)^{\mathcal{M}_3}$ of $\mathcal{M}_3$. However, $(f_1)^{-1}$ restricted to $(\delta_n)^{\mathcal{M}_3}$ is an $\mathcal{M}_1$-definable isomorphism between $(\delta_n)^{\mathcal{M}_3}$ and $(\delta_n)^{\mathcal{M}_1}$. So, $(f_1)^{-1} \circ f_2 \circ f_1$ is an $\mathcal{M}_1$-definable injective multifunction from $\mathcal{M}_1$ into $(\delta_n)^{\mathcal{M}_1}$, where $(\delta_n)^{\mathcal{M}_1}$ is a proper initial segment of $\mathcal{M}_1$. This contradicts Theorem 5.3. $\square$

5.3 Tight but neither neat nor semantically tight

In this subsection, we aim to define a theory that separates tightness from neatness and for which we also know that it is not semantically tight. We are able to find a sequential theory of this kind, but we do not know whether such theories can have arbitrary arithmetical strength.

Below, $\mathbb{Z}[X]$ denotes the ring of polynomials over $\mathbb{Z}$, which we see as a model for $\mathcal{L}_{\mathrm{PA}}$, with the ordering determined by making X greater than all the integers. We write $(\mathbb{Z}[X])_{\geq 0}$ for the nonnegative part of $\mathbb{Z}[X]$, which is a model of PA^-.

Lemma 5.7 ([5])**.** *The structure $((\mathbb{Z}[X])_{\geq 0}, X)$ is parameter-free bi-interpretable with $\mathbb{N}$. As a consequence, $(\mathbb{Z}[X])_{\geq 0}$ is bi-interpretable with $\mathbb{N}$; but it is not parameter-free bi-interpretable with $\mathbb{N}$.*

Proof. $(\mathbb{Z}[X])_{\geq 0}$ is clearly a computable structure, hence it is arithmetically definable, and we can fix an interpretation Z of a copy of $((\mathbb{Z}[X])_{\geq 0}, X)$ in the standard model $\mathbb{N}$. To be more specific, we represent polynomials from $((\mathbb{Z}[X])_{\geq 0}, X)$ as (natural numbers coding) finite sequences of integers.

This provides us with one interpretation needed for the bi-interpretability. To define the other one, observe that $\mathbb{N}$, the standard cut, is parameter-free definable in $(\mathbb{Z}[X])_{\geq 0}$. Namely, let $\delta(x)$ say that all numbers smaller or equal to x are either even or odd. Since elements of the form $X + k$, where $k \in \mathbb{Z}$, are downwards cofinal over $\mathbb{N}$ in $(\mathbb{Z}[X])_{\geq 0}$, and no such element is divisible by 2, only the standard integers satisfy $\delta(x)$ in $(\mathbb{Z}[X])_{\geq 0}$. This gives rise to an interpretation of $\mathbb{N}$ in $((\mathbb{Z}[X])_{\geq 0}, X)$ (in fact, in $(\mathbb{Z}[X])_{\geq 0}$), which we will denote by N.

Now we show that there is a parameter-free definable isomorphism between the identity interpretation on $((\mathbb{Z}[X])_{\geq 0}, X)$ and NZ. This crucially depends on the fact that PA^- is sequential, so internally in PA^- we have a notion of finite sequence that is well-behaved for sequences of standard length. Consequently, by mimicking the usual recursive definitions, we can define such notions as (standard) finite sums and finite products. In particular there is a $((\mathbb{Z}[X])_{\geq 0}, X)$-definable function with domain NZ which, given a coded finite sequence $a = (a_0, \ldots, a_n) \in \mathsf{NZ}$, returns $a_n X^n + a_{n-1} X^{n-1} + \ldots + a_1 X + a_0$; the definition of the function needs no parameters beyond X itself, which is named by a constant in $((\mathbb{Z}[X])_{\geq 0}, X)$. The inverse of this function is our $((\mathbb{Z}[X])_{\geq 0}, X)$-definable isomorphism between id and NZ. The $\mathbb{N}$-definable isomorphism j between id and ZN is the usual map from Section 3.1.

Thus, $\mathbb{N}$ is parameter-free bi-interpretable with $((\mathbb{Z}[X])_{\geq 0}, X)$, which means that it is also bi-interpretable with $(\mathbb{Z}[X])_{\geq 0}$. To prove that the bi-interpretability with $(\mathbb{Z}[X])_{\geq 0}$ requires parameters, it is enough to observe that $(\mathbb{Z}[X])_{\geq 0}$ carries a non-trivial automorphism: namely, the semiring homomorphism generated by $X \mapsto X + 1$, whose inverse is given by $X \mapsto X - 1$. On the other hand, $\mathbb{N}$ has no automorphisms other than the identity, and as mentioned in one of the remarks following the definition of bi-interpretation in Section 2, structures that are parameter-free bi-interpretable have isomorphic automorphism groups. □

In the remainder of this subsection, we will continue to use the notation $\delta(x)$, j, N, Z for the formulas resp. interpretations thus denoted in the proof of Lemma 5.7. We let ι_z stand for the $(\mathbb{Z}[X])_{\geq 0}$-definable map that, given a parameter z and a sequence $(a_0, \ldots, a_n)$ in NZ, outputs $a_n z^n + a_{n-1} z^{n-1} + \ldots + a_1 z + a_0$. Thus, $(\iota_X)^{-1}$ is an isomorphism between $((\mathbb{Z}[X])_{\geq 0}, X)$ and $((\mathbb{Z}[X])_{\geq 0}, X)^{\mathsf{NZ}}$

We also let h_z be the $(\mathbb{Z}[X])_{\geq 0}$-definable operation that maps the parameter z to $z + 1$ and extends to values of polynomials in z in the obvious way. More precisely: h_z takes $p \in (\mathbb{Z}[X])_{\geq 0}$ and searches for $a \in \mathsf{NZ}$ such that $p = \iota_z(a)$. If such an a does not exist, the function is undefined. If it does, then $h_z(p) := \iota_{z+1}(a)$. Note that in general, h_z is only a partial function – for instance, the domain of h_{X^2} is $(\mathbb{Z}[X^2])_{\geq 0}$ rather than all of $(\mathbb{Z}[X])_{\geq 0}$ – but h_X is an automorphism of $(\mathbb{Z}[X])_{\geq 0}$.

Let U be the following theory, axiomatizing some properties of $(\mathbb{Z}[X])_{\geq 0}$ in the spirit of the theories $\mathrm{IT}(n)$ and $\mathrm{IS}(n)$ of Sections 4.1 and 5.2, respectively:

(i) PA^-,

(ii) "δ defines a cut which is the smallest definable cut",

(iii) "there exists x such that h_x is a nontrivial automorphism of id",

(iv) "there exists x such that $(\iota_x)^{-1} \colon \mathsf{id} \to \mathsf{NZ}$ is an isomorphism",

(v) $\mathsf{N} \vDash$ "$j\colon \mathsf{id} \to \mathsf{ZN}$ is an isomorphism".

We observe that (ii) is an axiom scheme, while the other axioms of U are single statements.

Our goal is to prove the following theorem:

Theorem 5.8. *There is a sequential r.e. subtheory of* PA *which is tight but is neither neat nor semantically tight.*

The theory in question is $U \oplus_{\mathrm{I}\Sigma_1} \mathrm{PA}$. As usual, to prove a tightness-like property of the theory (here, specifically tightness only), we need to show the corresponding property for U and to obtain a result that rules out some "mixed cases" of interpretations.

Lemma 5.9. *The theory U is solid.*

Proof. Let $\mathcal{M}_1 \rhd \mathcal{M}_2 \rhd \mathcal{M}_3$ be models of U such that there is an $\mathcal{M}_1$-definable isomorphism from $\mathcal{M}_1$ onto $\mathcal{M}_3$. For each $i \in \{1,2,3\}$, consider $\mathcal{N}_i := \mathcal{M}_i^{\mathsf{N}}$ and $\mathcal{M}_i^* := \mathcal{N}_i^{\mathsf{Z}}$. Since each $\mathcal{M}_i$ is a model of U, it follows that $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ and $\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3$ satisfy the assumptions of Lemma 3.9 for $m = 0$, so $\mathcal{N}_1$ is $\mathcal{M}_1$-definably isomorphic to $\mathcal{N}_2$. Hence, $\mathcal{M}_1^*$ is $\mathcal{M}_1$-definably isomorphic to $\mathcal{M}_2^*$.

By axiom (iv) of U, each $\mathcal{M}_i$ is $\mathcal{M}_i$-definably and thus also $\mathcal{M}_1$-definably isomorphic to $\mathcal{M}_i^*$. Composing these isomorphisms with the one between $\mathcal{M}_1^*$ and $\mathcal{M}_2^*$, we obtain an $\mathcal{M}_1$-definable isomorphism between $\mathcal{M}_1$ and $\mathcal{M}_2$. $\square$

Remark. Note that in the proof of Lemma 5.9, the isomorphism between $\mathcal{M}_1$ and $\mathcal{M}_2$ needed to witness solidity is defined using parameters from $\mathcal{M}_1$ that might not be involved in defining the interpretations between $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$ and the isomorphism between $\mathcal{M}_1$ and $\mathcal{M}_3$. In any case, we will only need the tightness of U in the remainder of our argument.

Lemma 5.10. *If $\mathcal{M}_1 \vDash \mathrm{PA}$ and $\mathcal{M}_2 \vDash U$, then $\mathcal{M}_1$ and $\mathcal{M}_2$ are not parameter-free bi-interpretable.*

Proof. Suppose the contrary and let $(\mathsf{M}_2, \mathsf{M}_1)$ be a bi-interpretation in $\mathcal{M}_1 \models \mathrm{PA}$ such that $\mathcal{M}_1^{\mathsf{M}_2} \models U$. Let $\mathcal{K}$ be the prime substructure of $\mathcal{M}_1$. Then $\mathcal{K} \preccurlyeq \mathcal{M}_1$, so M_1 and M_2 witness that $\mathcal{K}$ is parameter-free bi-interpretable with a model of U. This cannot be the case, because the automorphism group of $\mathcal{K}$ is trivial, while every model of U, specifically of axiom (iii), carries a nontrivial automorphism. $\square$

Proof of Theorem 5.8. Clearly, $U \oplus_{\mathrm{I}\Sigma_1} \mathrm{PA}$ is an r.e. subtheory of PA, and it is sequential because it implies PA^-.

We now show that it is tight. Let V_1, V_2 be bi-interpretable extensions of $U \oplus_{\mathrm{I}\Sigma_1} \mathrm{PA}$, and let V_1, V_2 be interpretations witnessing the bi-interpretability. We claim that applying V_{3-i} in a model of $V_i + \mathrm{PA}$ gives rise to a model of $V_{3-i} + \mathrm{PA}$, and analogously for models of U instead of PA. To prove the claim, note that if $\mathcal{M} \vDash V_i + \mathrm{PA}$, then by the choice of V_1, V_2 the structures $\mathcal{M}$ and $\mathcal{M}^{\mathsf{V}_{3-i}}$ are in fact parameter-free bi-interpretable. So, by Lemma 5.10, it must be the case that $\mathcal{M}^{\mathsf{V}_{3-i}} \vDash \mathrm{PA}$. The proof for models of U is similar.

By the claim, $V_1 + \mathrm{PA}$ is bi-interpretable with $V_2 + \mathrm{PA}$, and $V_1 + U$ is bi-interpretable with $V_2 + U$. Thus, the solidity of PA implies that $V_1 + \mathrm{PA} \equiv V_2 + \mathrm{PA}$, and Lemma 5.9 implies that $V_1 + U \equiv V_2 + U$. So, $V_1 \equiv V_2$, proving tightness of $U \oplus_{\mathrm{I}\Sigma_1} \mathrm{PA}$.

The lack of semantical tightness is witnessed by the structures $\mathbb{N}$ and $(\mathbb{Z}[X])_{\geq 0}$, which are both models of $U \oplus_{\mathrm{I}\Sigma_1} \mathrm{PA}$ and are bi-interpretable by Lemma 5.7 but are not isomorphic. The lack of neatness is witnessed by the theories $\mathrm{Th}(\mathbb{N})$ and $\mathrm{Th}((\mathbb{Z}[X])_{\geq 0})$. The former is a retract of the latter, because not only the interpretations N and Z, but also the isomorphism j between $\mathbb{N}$ and $(\mathbb{N})^{\mathsf{ZN}}$ are defined without parameters. $\square$

5.4 Neat but not semantically tight

Our final separation result takes the following form.

Theorem 5.11. *There is a sequential r.e. theory which is neat but not semantically tight.*

This time, the theory in question will be a strengthening of PA formulated in the language extending $\mathcal{L}_{\mathrm{PA}}$ by a fresh constant symbol c. It will take the form $\mathrm{PA} + p(c)$, where $p(x)$ is a partial type with some particular properties.

Lemma 5.12. *There exists an r.e. partial type $p(x)$ over* PA *such that:*

(i) every element realizing p is undefinable,

(ii) if $\mathcal{M} \equiv \mathcal{N} \models \mathrm{PA}$ and a, b realize p in $\mathcal{M}$, $\mathcal{N}$ respectively, then $\mathrm{tp}^{\mathcal{M}}(a) = \mathrm{tp}^{\mathcal{N}}(b)$.

The Lemma can be obtained from known constructions of indiscernible types (see the proof of Theorem 3.1.2 in [18] and the Remark following it), but in order to make the paper more self-contained, we give a relatively simple proof based on flexible formulas.

Proof of Lemma 5.12. Let $\varphi_0(x), \varphi_1(x), \ldots$ be a computable enumeration of $\mathcal{L}_{\mathrm{PA}}$-formulas with one free variable in prenex normal form. Let $k_0, k_1, \ldots$ be the sequence of natural numbers defined inductively by $k_0 = \Sigma(\varphi_0)$, $k_{n+1} = k_n + \Sigma(\varphi_{n+1}) + 2$, where $\Sigma(\psi)$ stands for the smallest ℓ such that the formula ψ is Σ_ℓ.

Let $\neg\mathrm{Def}(x)$ be the partial type $\{\exists! y\, \varphi(y) \to \neg\varphi(x) : \varphi(x) \in \mathcal{L}_{\mathrm{PA}}\}$. Define the sets of $\mathcal{L}_{\mathrm{PA}}$-formulas $p_0(x), p_1(x), \ldots$ as follows:

$$\begin{aligned} p_0(x) &:= \neg\mathrm{Def}(x) \\ p_{n+1}(x) &:= p_n(x) \cup \{\forall y\, \xi_n(y) \leftrightarrow \varphi_n(x)\}, \end{aligned}$$

where ξ_n is a Σ_{k_n}-flexible formula over the $\mathcal{L}_{\mathrm{PA}}$-consequences of $\mathrm{PA} + p_n(c)$ (this assumes the satisfiability of each $p_n(x)$, which we will verify below). By Theorem 3.6 and its proof, we can require ξ_n to be a Σ_{k_n} formula of $\mathcal{L}_{\mathrm{PA}}$, and we can assume that the construction is computable in n, so that $p(x) := \bigcup_n p_n(x)$ is a computable set of $\mathcal{L}_{\mathrm{PA}}$-formulas.

We still have to check that $p(x)$ is in fact a type, i.e. that each $p_n(x)$ is satisfiable. We do this by induction on n. Clearly, $p_0(x)$ is satisfiable. Now assume that $p_n(x)$ is satisfiable and consider $p_{n+1}(x)$. Let $\alpha_n(x)$ denote the $\mathcal{L}_{\mathrm{PA}}$-sentence

$$\bigwedge_{k<n} \forall y\, \xi_k(y) \leftrightarrow \varphi_k(x)$$

(note that $p_n(x)$ is logically equivalent to $p_0(x) \cup \alpha_n(x)$). We claim that one of the following cases holds:

1° There is no $\ell \in \omega$ such that $\mathrm{PA} + \neg\mathrm{Def}(c) + \alpha_n(c) + \forall y\, \xi_n(y) \vdash \exists^{<\ell} x\, (\alpha_n(x) \land \varphi_n(x))$.

2° There is no $m \in \omega$ such that $\mathrm{PA} + \neg\mathrm{Def}(c) + \alpha_n(c) + \exists y\, \neg\xi_n(y) \vdash \exists^{<m} x\, (\alpha_n(x) \land \neg\varphi_n(x))$.

Assume the contrary. Then, since $\mathrm{PA} + \neg\mathrm{Def}(c) + \alpha_n(c) \vdash \exists^{\infty} x\, \alpha_n(x)$, for some $r \in \omega$ we have

$$\mathrm{PA} + \neg\mathrm{Def}(c) + \alpha_n(c) \vdash \forall y\, \xi_n(y) \leftrightarrow \exists^{<r} x\, (\alpha_n(x) \land \varphi_n(x)).$$

Since $\exists^{<r} x\, (\alpha_n(x) \land \varphi_n(x))$ is an $\mathcal{L}_{\mathrm{PA}}$-sentence of complexity at most Π_{k_n} (the formula $\alpha_n(x)$ is at most $\Sigma_{k_{n-1}+2}$), this contradicts the flexibility of ξ_n. Hence at least one of 1° and 2° holds.

Finally, we show that if 1° holds, then $p_{n+1}(x)$ is satisfiable (the argument for the case when 2° holds is analogous). From 1° it follows that for every ℓ, the theory

$$\mathrm{PA} + \neg\mathrm{Def}(c) + \alpha_n(c) + \forall y\, \xi_n(y) + \exists^{\geq \ell} x\, (\alpha_n(x) \wedge \varphi_n(x))$$

is consistent. In particular there is a model of $\mathrm{PA}+\neg\mathrm{Def}(c)+\alpha_n(c)+\forall y\, \xi_n(y)$ in which there are uncountably many elements satisfying the formula $\alpha_n(x) \wedge \varphi_n(x)$. So, the following theory is consistent as well

$$\mathrm{PA} + \neg\mathrm{Def}(c) + \alpha_n(c) + \forall y\, \xi_n(y) + (\alpha_n(d) \wedge \varphi_n(d)) + \neg\mathrm{Def}(d).$$

Any element interpreting the constant d in a model of this theory realizes $p_{n+1}(x)$. □

Proof of Theorem 5.11. Let T be $\mathrm{PA} + p(c)$, where $p(x)$ is the type provided by Lemma 5.12. Clearly, T is an r.e. theory.

We claim that T is not semantically tight, even in the weak sense of [8] (see one of the remarks following Definition 2.1) . Indeed, fix a model $(\mathcal{M}, c) \models T$ in which there is $d \neq c$ with the same complete $\mathcal{L}_{\mathrm{PA}}$-type as c (such a model exists since any element realizing $p(x)$ is undefinable). Consider $\mathcal{N} := \mathcal{K}(\mathcal{M}, c, d)$ - the submodel of $\mathcal{M}$ with the universe consisting of the elements which are definable with parameters from the set $\{c, d\}$ (or, equivalently from the pair $\langle c, d\rangle$). Since $\mathcal{N}$ is an elementary submodel of $\mathcal{M}$, it follows that both $(\mathcal{N}, c)$, $(\mathcal{N}, d)$ are models of T. Moreover both $(\mathcal{N}, c)$ and $(\mathcal{N}, d)$ are interpretable in $\mathcal{N}$ and it follows that they are bi-interpretable (using c, d as parameters). However, it follows from Ehrenfeucht's Lemma (see Section 2) that $\mathcal{N}$ admits no non-trivial automorphisms. Hence $(\mathcal{N}, c)$ and $(\mathcal{N}, d)$ witness that T is not semantically tight.

We now argue that T is neat. Take any two extensions U and V of T and assume that $\mathsf{V} : U \rhd V$ and $\mathsf{U} : V \rhd U$ witness that U is a retract of V. In particular VU is U-provably isomorphic to id_U. Take any $(\mathcal{M}, c) \models U$. We claim that $(\mathcal{M}, c) \simeq (\mathcal{M}, c)^{\mathsf{V}}$, so in particular $(\mathcal{M}, c) \models V$, which will suffice to prove neatness.

Consider $(\mathcal{N}, d) := (\mathcal{M}, c)^{\mathsf{V}}$ and $(\mathcal{M}^*, c^*) := (\mathcal{N}, d)^{\mathsf{U}}$. Since V and U witness the retraction between U and V, it follows that $(\mathcal{M}^*, c^*)$ is $(\mathcal{M}, c)$-definably isomorphic to $(\mathcal{M}, c)$. By the solidity of PA, there is also an $\mathcal{M}$-definable isomorphism ι between $\mathcal{M}$ and $\mathcal{N}$. Moreover, since ι is in fact the map from Lemma 3.2, and the interpretation V of $(\mathcal{N}, d)$ uses no parameters from $\mathcal{M}$ other than c, the definition of ι also has c as its unique parameter. This means that the element $\iota^{-1}(d)$ of $\mathcal{M}$ is definable in $\mathcal{M}$ from c.

Both $(\mathcal{M}, c)$ and $(\mathcal{N}, d)$ are models of T, so the properties of the type $p(x)$ imply that the $\mathcal{L}_{\mathrm{PA}}$-types of c and d are determined by the theories of $\mathcal{M}$ and $\mathcal{N}$, which are the same because $\mathcal{M}$ and $\mathcal{N}$ are isomorphic. Hence, $\iota^{-1}(d)$ is not only definable from c in $\mathcal{M}$, but also has the same arithmetical type as c. By Ehrenfeucht's Lemma (see Section 2), we obtain $\iota^{-1}(d) = c$, which means that ι is an isomorphism between $(\mathcal{M}, c)$ and $(\mathcal{N}, d)$; hence $(\mathcal{M}, c) \simeq (\mathcal{N}, d)$ as claimed. □

6 A weak subtheory of Z_2

The focus of this paper is on subtheories of first-order arithmetic. However, the question on the existence of solid proper subtheories asked in [4] concerned not only PA, but also other foundationally relevant axiom schemes, like second-order arithmetic Z_2 and ZF set theory.

As in the case of PA, we feel that (in order to avoid trivial examples) in the case of more powerful systems the question should also be not about proper subtheories as such, but proper subtheories containing a sufficiently strong characteristic fragment of a given axiom scheme, or even better about arbitrarily strong subtheories. In the present paper, we do not take up that problem. Nevertheless, to illustrate what can be done using our methods in a rather straightforward manner, we provide a simple example of a proper solid subtheory of Z_2 containing arithmetical comprehension.

Recall that ACA_0' is the theory that extends ACA_0 by the axiom "for every set X and every number k, the set $X^{(k)}$ exists", and ACA' is ACA_0' plus the full induction scheme for the language of second-order arithmetic.

Proposition 6.1. *There exists an r.e. solid proper subtheory of* Z_2 *extending* ACA'.

Proof. Let U be the theory of those models of ACA' that consider themselves to consist of the arithmetical sets. In other words, U is ACA' plus the axiom

$$\forall X \, \exists k \, (X \text{ is Turing-reducible to } 0^{(k)}).$$

We claim that $U \oplus_{\Pi^1_1\text{-}\mathrm{CA}_0} \mathrm{Z}_2$ is solid.

By [4], Z_2 is solid. It is also easy to show that U is solid: if $\mathcal{M}_1 \rhd \mathcal{M}_2 \rhd \mathcal{M}_3$ are models of U and we are given an $\mathcal{M}_1$-definable isomorphism between $\mathcal{M}_1$ and $\mathcal{M}_3$, then an argument just like the one for PA shows that there is also an $\mathcal{M}_1$-definable isomorphism between the first-order parts of $\mathcal{M}_1$ and $\mathcal{M}_2$ (the argument makes use of the full induction scheme available in U, because the interpretations between the models might be second-order definable). This extends to the second-order parts in the natural way: if $e_1, k_1 \in \mathcal{M}_1$ are mapped by the first-order isomorphism to $e_2, k_2 \in \mathcal{M}_2$, respectively, then map the set $\{e_1\}^{0^{(k_1)}}$ to $\{e_2\}^{0^{(k_2)}}$. We can prove by induction on k_1 that this is an embedding of the second-order universes, and it is surjective because $\mathcal{M}_2$ satisfies U.

By (an obvious variant of) Proposition 4.10, it remains to show that the family $\{U, \mathrm{Z}_2\}$ is retract-disjoint. This is similar to the proof of Lemma 4.6. If say $\mathcal{M} \models U$ is a retract of $\mathcal{N} \models \mathrm{Z}_2$, then again by the usual argument the first-order universes of $\mathcal{M}$ and $\mathcal{N}$ are definably isomorphic. But the second-order universe of $\mathcal{N}$ contains a set that is a definition of satisfaction for second-order formulas in $\mathcal{M}$, because all sets in $\mathcal{M}$ are internally arithmetical. We could use the isomorphism between the first-order universes to transfer this definition to $\mathcal{M}$, contradicting Tarski's theorem. The argument for the case when a model of Z_2 is a retract of a model of U is analogous. $\square$

By a somewhat more involved argument in a similar spirit, we can prove the solidity of a proper fragment of Z_2 containing Π^1_1-comprehension (and full induction). Solid proper subtheories of Z_2 containing fragments at the level of Π^1_2-comprehension and beyond are left as a possible topic for future work.

7 Conclusion and open problems

The work presented in this paper provides significant new insight into the behaviour of solidity and similar properties for subtheories of first-order arithmetic, as well as into the precise relations between the properties.

When it comes to solid subtheories of PA, we were able to not only show the existence of relatively strong solid proper subtheories of PA, but also to provide examples that are strictly below PA in terms of interpretability rather than just provability. Still, it seems that a piece of the picture remains missing.

Recall the Remark at the end of Section 4.3, pointing out that the reason why the theories TD_n defined in that section fail to interpret is that they are unable to make an

	sequential	arbitrarily strong below PA
not tight	[4]	[4]
tight only	Sec. 5.3	?
neat but not sem. tight	Sec. 5.4	?
sem. tight but not neat	?	?
sem. tight and neat only	?	?
solid	[4]	Sec. 4.1

Table 1: Possible combinations of categoricity-like properties discovered up to and including the present paper.

infinite case distinction, but each model of each TD_n actually interprets a model of PA. In fact, among the solid subtheories of PA that we are able to come up with, all the ones containing a reasonable amount of arithmetic (we leave aside trivial counterexamples like "either PA holds or the universe has one element") have the property that each of their models interprets a model of PA. Thus, the following question seems to be of interest.

Question 1. Can a solid subtheory of PA containing $\mathrm{I}\Delta_0 + \exp$ have a model that does not interpret any model of PA?

A potential negative answer to Question 1 could be interpreted as meaning that, in an appropriately weakened sense, PA is a minimal solid theory after all.

We also mention two questions in a similar spirit originally raised by other authors. One of the questions was already asked by Enayat in [4]:

Question 2. Is there a consistent finitely axiomatizable solid sequential theory?

The other was suggested by Fedor Pakhomov (private communication):

Question 3. Is there a solid sequential theory that is interpretable in $\mathrm{I}\Sigma_n$, for some n?

Note that a positive answer to Question 3 implies a positive answer to Question 1, because (for Gödel-style reasons) $\mathrm{I}\Sigma_n$ has models that do not interpret any model of PA.

Turning now to the topic of relations between tightness, solidity and the other notions, since solidity is the strongest of the four categoricity-like properties considered and tightness the weakest, *a priori* there could be up to six combinations of the properties. These combinations correspond to rows of Table 1 below and are listed in the first column.

Before our work, all theories that had been classified were either solid or not even tight. We were able to come up with examples witnessing some separations between the properties, including a theory that is tight but has none of the stronger properties and a theory that is neat but has neither of the semantical properties. All the separations we obtained can be witnessed by theories that are at least sequential theories, although separating examples that we managed to classify exactly do not have arbitrary arithmetical strength. Table 1 lists the combinations of properties known to occur based on results up to and including the present paper, with references to sections of this paper or to earlier work, as appropriate.

Recently, the first author developed a new method that makes it possible to show that in fact all six combinations corresponding to rows of Table 1 are possible, as witnessed by variants of finite set theory. These advances will be reported in a separate work [11].

One example considered in the present paper that is not represented in Table 1 is the family of tight but not neat theories S_n studied in Section 5.2. These theories could fill either the second or the fourth row of Table 1, depending on whether they are semantically tight. Thus, we ask:

Question 4. Are the theories S_n from Section 5.2 semantically tight?

Acknowledgements

The work of the first and third authors was supported by grant no. 2022/46/E/HS1/00452 of the National Science Centre, Poland. The work of the second author was supported by grant no. 2023/49/B/ST1/02627 of the National Science Centre, Poland.

The authors would like to thank Ali Enayat for fruitful discussions providing comments on the earlier versions of this paper.

References

[1] Gisela Ahlbrandt and Martin Ziegler. Quasi-finitely axiomatizable totally categorical theories. *Ann. Pure Appl. Logic*, 30(1):63–82, 1986.

[2] Lev Beklemishev. Reflection principles and provability algebras in formal arithmetic. *Russian Math. Surveys*, 60(2):197–268, 2005.

[3] Andrzej Ehrenfeucht. Discernible elements in models for Peano arithmetic. *J. Symb. Log.*, 38:291–292, 1973.

[4] Ali Enayat. Variations on a Visserian theme. In *A tribute to Albert Visser*, volume 30 of *Tributes*, pages 99–110. College Publications, London, 2016.

[5] Ali Enayat, Mateusz Łełyk, and Albert Visser. Completions of restricted complexity I, weak arithmetical theories. Preprint, available at arXiv:2508.14758, 2025.

[6] Ali Enayat and Mateusz Łełyk. Categoricity-like properties in the first order realm. *J. Phil. Math.*, 1:63–98, 2024.

[7] Solomon Feferman. Arithmetization of metamathematics in a general setting. *Fund. Math.*, 49:35–92, 1960/61.

[8] Alfredo Roque Freire and Joel David Hamkins. Bi-interpretation in weak set theories. *J. Symb. Log.*, 86(2):609–634, 2021.

[9] Alfredo Roque Freire and Kameryn J. Williams. Non-tightness in class theory and second-order arithmetic. *J. Symb. Log.*, 90(2):627–654, 2025.

[10] Harvey Friedman and Albert Visser. When bi-interpretability implies synonymy. *Review of Symbolic Logic*, pages 1–20, forthcoming.

[11] Piotr Gruza. Separations between definiteness properties for sequential theories [working title]. in preparation.

[12] Petr Hájek and Pavel Pudlák. *Metamathematics of first-order arithmetic.* Perspectives in Mathematical Logic. Springer-Verlag, Berlin, 1998. Second printing.

[13] Volker Halbach. *Axiomatic theories of truth.* Cambridge University Press, Cambridge, 2011.

[14] Wilfrid Hodges. *Model theory*, volume 42 of *Encyclopedia of Mathematics and its Applications.* Cambridge University Press, Cambridge, 1993.

[15] Emil Jeřábek. Sequence encoding without induction. *Math. Log. Q.*, 58(3):244–248, 2012.

[16] Richard Kaye. *Models of Peano Arithmetic*, volume 15 of *Oxford Logic Guides*. The Clarendon Press, Oxford University Press, New York, 1991. Oxford Science Publications.

[17] Richard Kaye. The theory of κ-like models of arithmetic. *Notre Dame J. Formal Logic*, 36(4):547–559, 1995.

[18] R. Kossak and J. Schmerl. *The Structure of Models of Peano Arithmetic*. Oxford University Press, 2006.

[19] Saul A. Kripke. "Flexible" predicates of formal number theory. *Proc. Amer. Math. Soc.*, 13:647–650, 1962.

[20] Mateusz Łełyk and Bartosz Wcisło. Universal properties of truth. *J. Math. Log.*, 2025. Online Ready.

[21] Per Lindström. *Aspects of incompleteness*, volume 10 of *Lecture Notes in Logic*. Association for Symbolic Logic, Urbana, IL; A K Peters, Ltd., Natick, MA, second edition, 2003.

[22] Franco Montagna. Relatively precomplete numerations and arithmetic. *J. Philos. Logic*, 11(4):419–430, 1982.

[23] Andrzej Mostowski. A generalization of the incompleteness theorem. *Fund. Math.*, 49:205–232, 1960/61.

[24] Fedor Pakhomov. How to escape Tennenbaum's theorem. Preprint, available at arXiv:2209.00967, 2022.

[25] Pavel Pudlák. Some prime elements in the lattice of interpretability types. *Trans. Amer. Math. Soc.*, 280(1):255–275, 1983.

[26] Albert Visser. An inside view of EXP; or, The closed fragment of the provability logic of $\mathrm{I}\Delta_0 + \Omega_1$ with a propositional constant for EXP. *J. Symb. Log.*, 57(1):131–165, 1992.

[27] Albert Visser. Categories of theories and interpretations. In *Logic in Tehran*, volume 26 of *Lect. Notes Log.*, pages 284–341. Assoc. Symbol. Logic, La Jolla, CA, 2006.

[28] Albert Visser. The small-is-very-small principle. *MLQ Math. Log. Q.*, 65(4):453–478, 2019.

[29] A. J. Wilkie. On schemes axiomatizing arithmetic. In *Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Berkeley, Calif., 1986)*, pages 331–337. Amer. Math. Soc., Providence, RI, 1987.