Interpretations in Presburger Arithmetic

Jetze Zoethout, 4014294

Bachelor Thesis

February 1, 2015

Supervisor: Albert Visser
Second corrector: Lev Beklemishev
Faculty: Humanities
Department: Philosophy and Religious Studies
Bachelor programme: Philosophy
ECTS: 15

Contents

Introduction

The general goal of this thesis is to obtain a better understanding of the relations between various weak systems of arithmetic. One specific relation we will investigate is interpretability, a notion we explain below. Both semantical and syntactical aspects of these arithmetics will play a role in this investigation. Thus, this thesis is a study of arithmetical theories, whose methods come both from model theory and from proof theory. The more specific goal of this thesis is to apply the concept of an *interpretation* to the theory *Presburger Arithmetic*. Let us briefly explain what these two terms entail.

Where Peano Arithmetic, or `PA` for short, is meant to be about the natural numbers with addition and multiplication, Presburger Arithmetic, or `PrA` for short, concerns itself only with the former of these operations. That is, `PrA` is a theory about the natural numbers with their addition structure. Since Mojżesz Presburger first studied `PrA` in 1929, the theory has been an object of interest, partly because of its striking analogies and disanalogies with `PA`. Leaving multiplication out makes `PrA` a lot less complicated than `PA`, but on the other hand, some characteristic properties of `PA` are preserved. Disanalogies of `PrA` with `PA` include completeness, decidability and the existence of recursive nonstandard models. Analogies include the order type of countable nonstandard models and the impossibility of giving a finite axiomatization. We will develop Presburger Arithmetic and its properties at length in chapter 1.

An *interpretation* is a way of acting as if some theory V were about the contents of some other theory U. More specifically, we translate U-statements into V-statements in such a way that V proves all translations of U-theorems. All details concerning interpretations can be found in section interpretations. During the rest of chapter 2, we will formulate two conjectures about interpretations of `PrA`. The first of these turns out to reduce to the second, which is why we dedicate chapter 3 to the investigation of this second conjecture. The first few sections offer some important preliminary considerations, and in the final section, we will outline two strategies that might lead to a proof of the conjecture under investigation.

While it is essential to explain what a thesis does, it is also important to note what it does *not* do. In this thesis, we will not go into decision procedures concerning `PrA` and not use automata theory, which has some applications in the study of `PrA`. These are beyond the scope of a bachelor thesis.

As the above discussion indicates, this thesis assumes a background in logic. Although decidability also plays a small role, no extensive knowledge in computability theory is required. Some familiarity with the intuition behind computability suffices. When proving decidability results, we use the Church-Turing thesis and for undecidability results, we will rely on other undecidability results that are well-known. Before we start, let us fix some basic logical concepts and notations.

- For sets A and B, we write $A \subset B$ to indicate that every element of A also belongs to B.
- For simplicity, constant symbols are considered 0-ary function symbols.
- We assume that our languages always contain the binary predicate symbol $=$. If M is a structure for a certain language L, then we do *not* demand $=$ to be interpreted in M as real identity on M. We *do* suppose the validity of the axioms for identity. That

is, the interpretation of $=$ should be some equivalence relation on M that respects the interpretations of all the other predicate and function symbols in L. Note that, from the point of view of predicate logic, this is the right way to handle identity. Indeed, the identity axioms express everything we can say about identity *in predicate logic*.

When we consider isomorphisms between structures, we should note that injectivity and well-definedness are relative to a notion of identity. It may seem that it doesn't really matter whether we allow the interpretation of $=$ to be something else than real identity. If $=$ isn't interpreted in M as real equality, we can construct an isomorphic structure in which it is. We may do this by taking as domain the $=$-equivalence classes in M and by inducing the other predicates and functions from M. However, when dealing with interpretations we have to contrast internal and external notions of identity, and then it is extremely convenient to employ the chosen conception of identity.

- We will use the symbols $\models$ and $\vdash$ for semantical and syntactical consequence respectively. As our proof system, we pick a system that (i) satisfies the completeness theorem for predicate logic and such that (ii) proofs are finite objects that can be recursively checked. We know such proof systems to exist, and the specific choice doesn't matter. Whenever we want to prove a statement of the form $\Gamma \vdash A$, we apply the completeness theorem. Such an application is indicated by a phrase like "Let M be a model of Γ", and then we continue to prove that $M \models A$.

- A theory T in a certain language L is a set of L-formulas that is closed under syntactical consequence. That is, if $T \vdash A$ for some L-formula A, then $A \in T$. Given two theories T_0 and T_1, we write $T_0 + T_1$ for the smallest theory containing both T_0 and T_1 (and this is *not* necessarily $T_0 \cup T_1$).

Finally, I would like to express a few thanks. To Lev Beklemishev, for suggesting the article [3], which eventually led to the strategy employed in section 3.3. To Clemens Grabmeyer, for supplying this thesis with a very convenient axiomatization of `PrA`. And of course to Albert Visser, who has taught me valuable lessons both about the subject-matter of this thesis, and about putting it on paper.

1 Presburger Arithmetic

This first chapter is devoted to developing Presburger Arithmetic. We start by giving its axioms, deriving some of its elementary properties and investigating its models. Subsequently, we will use quantifier elimination to prove the most striking properties of Presburger Arithmetic: completeness and decidability.

1.1 Definition and elementary properties

Throughout this thesis, our base language will be $\mathcal{L}^- = \langle 0, 1, + \rangle$. Here 0 and 1 are symbols for constants and + is a binary function symbol. For (possibly open) $\mathcal{L}^-$-terms t and s, we define

- $t \neq s$ as $\neg(t = s)$;
- $t < s$ as $\exists u\ (t + (u + 1) = s)$;
- $t \leq s$ as $t < s \vee t = s$;
- $t > s$ as $s < t$;
- $t \geq s$ as $s \leq t$;
- $\underline{n}t$ for $n \in \mathbb{N}$ by recursion: $\underline{0}t$ is 0 and $(\underline{k+1})t$ is $(\underline{k}t + t)$ for $k \in \mathbb{N}$;
- $\underline{n}$ as $\underline{n}1$ for $n \in \mathbb{N}$.
- $t \equiv_n s$ as $\exists u\ (t = \underline{n}u + s \vee s = \underline{n}u + t)$, for integers $n \geq 1$.

We also define the language $\mathcal{L}$ as $\langle 0, 1, +, <, \{\equiv_n \mid n \in \mathbb{Z}_{\geq 1}\}\rangle$. We will officially work in $\mathcal{L}^-$, but in practice we will work in the definitional extension of our theory in $\mathcal{L}$, given by the above definitions. The advantage of working with the language $\mathcal{L}$, besides prettier notation, may not yet be clear, but it will become clear in section 1.4.

We are now ready to define the theory that interests us. The following axiom set is due to Clemens Grabmayer [2], with a simplification due to Albert Visser [6].

Definition 1.1.1. The $\mathcal{L}$-theory *Presburger Arithmetic*, which we denote by `PrA`, is given by the following axioms:

`PrAx1` $x + 1 \neq 0$;

`PrAx2` $x + z = y + z \to x = y$;

`PrAx3` $x + 0 = x$;

`PrAx4` $x + (y + z) = (x + y) + z$;

`PrAx5` $x = 0 \vee \exists y\ x = y + 1$;

`PrAx6` $x + y = y + x$;

`PrAx7` $x < y \vee x = y \vee x > y$;

`PrAx8` for $n \geq 1$, $(x \equiv_n \underline{0}) \vee (x \equiv_n \underline{1}) \vee \ldots \vee (x \equiv_n \underline{n-1})$.

Remember that in the above axioms, we are officially supposed to have the definitions of $<$ and $\equiv_n$. Also, note that $\texttt{PrAx8}$ is not a proper axiom, but rather a set of axioms, indexed by the positive integers. We refer to the instances of $\texttt{PrAx8}$ as $\texttt{PrAx8}_n$. We write $\texttt{PrA}^-$ for the theory given by the first seven of the above axioms, and $\texttt{PrA}_n$ for the theory $\texttt{PrA}^- + \texttt{PrAx8}_n$. Note that $\texttt{PrAx8}_1$ is just an abbreviation for the formula $\exists u\ (x = u + 0 \vee 0 = u + x)$, which is provable in $\texttt{PrA}^-$ alone. So $\texttt{PrA}_1$ is really just $\texttt{PrA}^-$.

We will now introduce the standard model for Presburger Arithmetic.

Definition 1.1.2. The *standard model*, which we denote by $\mathbb{N}$, is an $\mathcal{L}^-$-structure, where the domain is the set of natural numbers, and where 0, 1 and + have their straightforward interpretations. As a result, $<$ is interpreted as the ordering on the natural numbers, while $\equiv_n$ is interpreted as congruence modulo n.

It is rather obvious that $\mathbb{N}$ is a model of $\texttt{PrA}$ and in particular, we see that $\texttt{PrA}$ is consistent. We now make some simple observations, but we will not include full proofs.

Proposition 1.1.1. *Let $n \geq 1$. The following statements are provable in $\texttt{PrA}_n$:*

1. *$<$ defines a discrete linear order with smallest element 0 and second element 1;*
2. *$x < y \leftrightarrow x + z < y + z$;*
3. *$\equiv_n$ defines an equivalence relation with exactly n equivalence classes;*
4. *$x \equiv_n y \leftrightarrow x + z \equiv_n y + z$;*
5. *The disjuncts in $\texttt{PrAx7}$ are mutually exclusive, and the same holds for $\texttt{PrAx8}_n$.*

Proof. This is an exercise in predicate logic. □

Proposition 1.1.2. *Let t be a closed $\mathcal{L}$-term that is interpreted in the standard model as $a \in \mathbb{N}$. Then $\texttt{PrA} \vdash t = \underline{a}$.*

Proof. Apply induction on complexity. □

Proposition 1.1.3. *Let ϕ be an atomic $\mathcal{L}$-sentence. Then $\texttt{PrA} \vdash \phi$ if $\mathbb{N} \models \phi$, while $\texttt{PrA} \vdash \neg\phi$ if $\mathbb{N} \not\models \phi$.*

Proof. Use proposition 1.1.2 and the last item of proposition 1.1.1. □

We mention the following important corollary:

Corollary 1.1.4. *$\texttt{PrA}$ decides atomic $\mathcal{L}$-sentences.*

We now shift our attention towards the relation between the various principles $\texttt{PrAx8}_n$.

Proposition 1.1.5. *For all integers $m, n \geq 1$, $\texttt{PrA}_m + \texttt{PrA}_n$ is equivalent to $\texttt{PrA}_{mn}$.*

Proof. Work in a model M of $\texttt{PrA}_{mn}$ and let $x \in M$ be arbitrary. Then there exists a $u \in M$ and an $r \in \mathbb{N}$ such that $M \models x = \underline{mn}u + \underline{r}$ and $r < mn$.[1] Use the Euclidean division algorithm to write $r = mq + s$ with $q, s \in \mathbb{N}$ and $s < m$, so that

$$M \models x = \underline{mn}u + \underline{r} = \underline{m}(\underline{n}u) + \underline{mq + s} = \underline{m}\left(\underline{n}u + \underline{q}\right) + \underline{s}.$$

[1] By the definition of $\equiv_n$, we could also have $M \models \underline{r} = \underline{mn}u + x$, but in that case we can easily show $M \models u = 0$ and we have the other case. This is because, for $k \geq 1$, the principle $\texttt{PrAx8}_k$ mentions the *smallest* representatives from each equivalence class. We will use this observation from now on.

So $\mathtt{PrA}_m$ holds in M; the result for n can be obtained analogously.
Now work in a model N of $\mathtt{PrA}_m + \mathtt{PrA}_n$ and let $x \in N$ be arbitrary. Then there exists a $u \in N$ and an $r \in \mathbb{N}$ such that $N \models x = \underline{m}u + \underline{r}$ and $r < m$. Furthermore, there are $v \in N$ and $s \in \mathbb{N}$ such that $N \models u = \underline{n}v + \underline{s}$ and $s < n$. Now we get

$$N \models x = \underline{m}u + \underline{r} = \underline{m}(\underline{n}v + \underline{s}) + \underline{r} = \underline{mn}v + \underline{ms + r}.$$

Since $ms + r \leq m(n-1) + (m-1) = mn - 1 < mn$, we are done. □

Corollary 1.1.6. *For integers $m, n \geq 1$, if every prime factor of m also divides n, we have $\mathtt{PrA}_n \vdash \mathtt{PrA}_m$.*

Proof. Immediate by proposition 1.1.5 and unique prime factorisation. □

1.2 Models of $\mathtt{PrA}$

In this section, we will investigate what a typical model of $\mathtt{PrA}$ or $\mathtt{PrA}_n$ looks like. This investigation will lead to the converse of corollary 1.1.6 and to the result that $\mathtt{PrA}$ is not finitely axiomatizable. We start with a theorem that reminds us of an insight concerning nonstandard models of $\mathtt{PA}$.

Theorem 1.2.1. *Let $n > 1$ be an integer. If M is some nonstandard model of $\mathtt{PrA}_n$, then it has the order type $\mathbb{N} + \mathbb{Z} \cdot A$, where $\langle A, <_A \rangle$ is a dense linear order without endpoints. In particular, every countable nonstandard model of $\mathtt{PrA}_n$ has order type $\mathbb{N} + \mathbb{Z} \cdot \mathbb{Q}$.*

Proof. We use the same technique as in the proof for the analogous fact concerning $\mathtt{PA}$: divide the nonstandard part of M into copies of $\mathbb{Z}$ and let A be the set of these copies. If $a \in M - \mathbb{N}$, we denote the copy that contains a by $[a]$ and we say $[a] <_A [b]$ iff $a < b$ and $[a] \neq [b]$. It is straightforward to check that this is a linear order, so it remains to check density and the absence of endpoints. That is, given $[a] <_A [b]$, we want to construct elements of A smaller than $[a]$, between $[a]$ and $[b]$, and larger than $[b]$.
The principle $\mathtt{PrAx8}_n$ asserts the existence of the quotient of an element $x \in M$ upon division by n. We can quite easily prove that is unique, so we may without ambiguity denote it by $\left\lfloor \frac{x}{n} \right\rfloor$. Now it isn't difficult to show that, since $n > 1$, we have

$$\left[\left\lfloor \frac{a}{n} \right\rfloor\right] <_A [a] <_A \left[\left\lfloor \frac{a + (n-1)b}{n} \right\rfloor\right] <_A [b] <_A [\underline{2}b]. \qquad \square$$

This shouldn't come as a surprise, because the only number theoretical fact used in the proof for $\mathtt{PA}$ is that we can divide by 2. The above proof shows that it is in fact sufficient to be able to divide by something larger than 1.

Now let us try to construct a nonstandard model of $\mathtt{PrA}$. For simplicity, we index the copies of $\mathbb{Z}$ in the nonstandard part by $\mathbb{Q}_{>0}$, which is of course order-isomorphic to $\mathbb{Q}$, and we give the standard part $\mathbb{N}$ the index 0. We represent each element by a pair (q, n), where q is the index of the copy and n is its exact location in this copy. Thus, if $q > 0$, then n is some integer, and if $q = 0$, then $n \in \mathbb{N}$. Obviously, 0 is the element $(0, 0)$, and 1 is $(0, 1)$, so it remains to define a suitable addition. A natural choice would of course be coordinate-wise addition. We then get $(p, m) < (q, n)$ precisely if $p < q$, or both $p = q$ and $m < n$; congruence modulo

n is simply congruence modulo n for the second coordinate. This choice in fact satisfies all axioms of `PrA`.

The above model may not seem very interesting, but is shows that *Tennenbaum's theorem does not hold for `PrA`*. In other words, there are recursive nonstandard models for `PrA`. This result in itself isn't too surprising, because `PrA` is a rather less complicated object than `PA`. But the reason why Tennenbaum's theorem doesn't hold here, is quite interesting. In the case of `PA`, it is tempting to blame the presence of induction for the impossibility of defining a recursive addition. Indeed, the order type of a nonstandard model is not a well-order, so we do not expect induction to hold if the model looks too simple. However, as we will prove in section 1.5, we *do* have an induction scheme in `PrA`. In fact, `PrA` is as strong as it can be, given that it doesn't talk about multiplication. So the reason why we can define addition so easily in the case of `PrA` must be that it doesn't have to be compatible with a multiplication structure.

The model we have given above may seem a little artificial because we used ordered pairs, but in fact, it is quite a natural mathematical object. Consider linear polynomials in one variable X with coefficients in $\mathbb{Q}$. There are obvious candidates for 0, 1 and $+$. If we want to end up with a model of `PrA`, we need to define some order on this set.[2] A common way to do this is by making the variable X 'infinitely large'. That is, when comparing two polynomials, we first consider the coefficient of X and only when these are equal, we turn to the constant coefficient.[3] More to the point, we have $aX + b > cX + d$ precisely if $a > c$, or both $a = c$ and $b > d$. Now we immediately get another problem we have to solve, namely the existence of negative polynomials. Indeed, we obviously have $-1 < 0$, and also $-X < 0$. Fortunately, we can easily get around this by only considering the non-negative polynomials. Equivalently, we consider only the zero polynomial and those polynomials whose leading coefficient is positive. Thus, for example, $X - 10^{10}$ belongs to our model, but -1 does not. The final problem we have to solve is that the polynomial 1 should come immediately after the zero polynomial, which is not the case since our coefficients are from $\mathbb{Q}$. There is an ad hoc way around this: we demand the constant coefficient to be an integer. The resulting structure satisfies `PrA`, as one may check. In fact, it is isomorphic to the model we constructed earlier. This model arises so naturally that we as well call it the simple nonstandard model. We summarize:

Definition 1.2.1. The *simple nonstandard model*, that we denote by $\mathcal{S}$, is the $\mathcal{L}^-$-structure given by the set of linear polynomials p in one variable X with coefficients in $\mathbb{Q}$ such that:

- the constant coefficient of p is an integer;
- if p is not the zero polynomial, then its leading coefficient is positive.

The obvious interpretations are given to 0, 1 and $+$. As a result, the order is given by making X 'infinitely large', while $\equiv_n$ is just congruence modulo n for the constant coefficient.

Let us consider the matter of congruence modulo n more closely. Why does `PrAx8`$_n$ hold in $\mathcal{S}$? Well, given a polynomial $aX + b \in \mathcal{S}$, we can find $q, r \in \mathbb{Z}$ with $0 \le r < n$ and $b = qn + r$, so we can write $aX + b = aX + (qn + r) = n\left(\frac{a}{n}X + q\right) + r$, showing that we can indeed divide by n with remainder. Now suppose that we would like to construct models of `PrA`$_n$, but not necessarily of the whole of `PrA`. If we want a model in the style of $\mathcal{S}$, then we at least need

[2] In `PrA`, the order can be constructed from addition, but we do not yet have a structure satisfying `PrA`. Indeed, the set under consideration is a group under addition of polynomials, so this gives us no information.

[3] An equivalent formulation is: a polynomial is larger than another one iff the first becomes larger than the second for X large enough.

to be able to divide the coefficient of X by n, as the above shows. So let us take all those $aX + b \in \mathcal{S}$ such that the denominator of a is a power of n. This move could of course ruin closure under addition, but it doesn't. If $\frac{a}{n^k}X + b$ and $\frac{c}{n^l}X + d$ are in $\mathcal{S}$, then

$$\left(\frac{a}{n^k}X + b\right) + \left(\frac{c}{n^l}X + d\right) = \frac{an^l + cn^k}{n^{k+l}}X + (b + d)$$

is of the appropriate form. Here a, b, c and d are integers, and k and l are natural numbers. Checking the axioms of $\texttt{PrA}_n$ is as easy as it can be.

Definition 1.2.2. For an integer $n \geq 1$, we define the $\mathcal{L}^-$-structure $\mathcal{S}_n$ as the substructure of $\mathcal{S}$ given by those polynomials $aX + b \in \mathcal{S}$ such that the denominator of a is a power of n. It is even a substructure of $\mathcal{S}$ with respect to $<$.

The above models, due to Craig Smorynski [5], are constructed specifically for $\texttt{PrAx8}_n$ to hold. Which other instances of $\texttt{PrAx8}$ are valid in $\mathcal{S}_n$? The answer turns out to be: as few as corollary 1.1.6 permits.

Proposition 1.2.2. *For integers $m, n \geq 1$, we have $\mathcal{S}_n \models \texttt{PrAx8}_m$ iff every prime factor of m also divides n.*

Proof. Suppose $\texttt{PrAx8}_m$ holds in $\mathcal{S}_n$. Then in particular there must be a $u \in \mathcal{S}_n$ and an $r \in \mathbb{N}$ such that $\mathcal{S}_n \models X = \underline{m}u + \underline{r}$ and $r < m$. Write $u = \frac{a}{n^k}X + b$, then $X = m\left(\frac{a}{n^k}X + b\right) + r = \frac{am}{n^k}X + (bm + r)$, so we get $am = n^k$. In other words, m divides some power of n, which means exactly that every prime factor of m also divides n. The other direction follows from corollary 1.1.6. □

In particular, not all of $\texttt{PrA}$ holds in $\mathcal{S}$. This also means that $\mathcal{S}_n$ cannot be a substructure with respect to all the congruence symbols. Indeed, that would imply that $\mathcal{S}_n$ models all instances of $\texttt{PrAx8}$, since these axioms, *when formulated in* $\mathcal{L}$, are universal.

Corollary 1.2.3. *For integers $m, n \geq 1$, we have $\texttt{PrA}_n \vdash \texttt{PrA}_m$ if and only if every prime factor of m also divides n.*

Proof. Immediate by corollary 1.1.6 and proposition 1.2.2. □

This corollary has two important implications. The first is that, when we consider some axiom $\texttt{PrAx8}_n$, it only matters what prime factors occur in n. So it is in fact no restriction to consider only axioms $\texttt{PrAx8}_p$, where p is prime. We also expect these axioms to be rather independent. This is partly expressed in the second implication, which is an analogy with $\texttt{PA}$:

Theorem 1.2.4. *$\texttt{PrA}$ is not finitely axiomatizable.*

Proof. Suppose the contrary. Then $\texttt{PrA}$ must be axiomatizable by some finite subset of the axioms stated in definition 1.1.1, so in particular it is axiomatizable by the theory $\texttt{PrA}^- + \{\texttt{PrAx8}_{n_i} \mid 1 \leq i \leq k\}$ for some $n_1, \ldots, n_k \geq 1$ and $k \geq 1$. By proposition 1.1.5, this theory is equivalent to $\texttt{PrA}_N$ with $N = \Pi_{i=1}^k n_i$. Now by our assumption, $\mathcal{S}_N \models \texttt{PrA}$, but by proposition 1.2.2, $\mathcal{S}_N \not\models \texttt{PrAx8}_{N+1}$, since $\gcd(N, N+1) = 1$, contradiction. □

It is this result that we will try to strengthen in various ways.

1.3 Classification of terms and atomic formulas

Despite the extensive syntactic and semantic explorations concerning PrA we have conducted so far, we have yet to expose the true nature of PrA. We will show PrA to be decidable, complete, and in fact equal to the true theory of the standard model. This situation is radically different from the one where multiplication is present, because theories like PA are not only weaker than the true theory of $\mathbb{N}$, but this shortcoming is also fundamental: there *is* no recursive axiomatization for the latter. The technique we will use to show all these facts is *quantifier elimination.* The fact that PrA admits quantifier elimination was first discovered by Mojżesz Presburger in 1929. The version we will use can be found in Herbert Enderton's [1]. In this section we will discuss some preliminary results, while the actual quantifier elimination will be carried out in section 1.4. In section 1.5 we will discuss the implications mentioned above.

Lemma 1.3.1. *Let $k \geq 0$ and let $t(x_1, \ldots, x_k)$ be an $\mathcal{L}$-term. Then PrA proves that t is equal to a term of the form*

$$\underline{n_0} + \underline{n_1}x_1 + \cdots + \underline{n_k}x_k,$$

where $n_0, n_1, \ldots, n_k \in \mathbb{N}$.

Proof. An easy induction on complexity. □

Suppose ϕ is an atomic formula and x is a variable. Then ϕ is of the form $t_0 R s_0$, where t_0 and s_0 are $\mathcal{L}$-terms and R is $=$, $<$, or $\equiv_n$ for some $n \geq 1$. By the above lemma, PrA proves ϕ to be equivalent to $(t + \underline{a}x)R(s + \underline{b}x)$ for some $a, b \in \mathbb{N}$ and $\mathcal{L}$-terms t and s not containing x. Furthermore, we can cancel out all x's on at least one side of R. Since $=$ and $\equiv_n$ are symmetric, we get:

Proposition 1.3.2. *Suppose ϕ is an atomic $\mathcal{L}$-formula in one variable x. Then ϕ is equivalent in PrA to one of the following formulas:*

$$\begin{aligned} &\underline{a}x + t = s; \\ &\underline{a}x + t < s; \\ &\underline{a}x + t > s; \\ &\underline{a}x + t \equiv_n s \textit{ for some } n \geq 1, \end{aligned}$$

where $a \in \mathbb{N}$ and s and t are $\mathcal{L}$-terms not containing x.

More down to earth, we can extract x on both sides and cancel it out on one side. Finally, we need the following well-known lemma:

Lemma 1.3.3. *Let T be some theory. Suppose that for every formula ϕ of the form*

$$\exists x\ (\alpha_1 \wedge \ldots \wedge \alpha_k), \tag{1}$$

where $k > 0$ and the α_i are atomic or the negation of an atomic formula, we can find a quantifier-free ψ such that $T \vdash \phi \leftrightarrow \psi$. Then T admits quantifier elimination.

1.4 Quantifier elimination for `PrA`

Let's get straight to work.

Theorem 1.4.1. *`PrA` admits quantifier elimination in the language $\mathcal{L}$.*

Proof. Suppose we have an $\mathcal{L}$-formula ϕ of the form (1). We will show how to find a quantifier-free equivalent for ϕ, which suffices by lemma 1.3.3.

1. First of all, we eliminate the occurrence of negated atoms. Suppose t and s are $\mathcal{L}$-terms. The last item of proposition 1.1.1 tells us that we can prove in `PrA` that:

$$\begin{aligned} t \neq s &\leftrightarrow t < s \vee s < t; \\ \neg(t < s) &\leftrightarrow t = s \vee s < t; \\ \neg(t \equiv_n s) &\leftrightarrow (t \equiv_n s + \underline{1}) \vee \ldots \vee (t \equiv_n s + \underline{n-1}). \end{aligned}$$

   In this way, all negated atoms are replaced by disjunctions of atoms, which means that the formula in the scope of $\exists x$ is in conjunctive normal form. Put this in disjunctive normal form by distributing the $\wedge$'s over the $\vee$'s. Now we can distribute $\exists$ over the $\vee$'s to obtain a disjunction of formulas of the form (1), where all α_i are *atomic*. It is now sufficient to find quantifier-free equivalents for ϕ of this form.

2. We can assume that x occurs in all α_i, because otherwise we can just bring such an α_i out of the existential quantifier. So all α_i are without loss of generality of one of the forms mentioned in proposition 1.3.2, with $a > 0$. We call this a the *coefficient* of x in the atom α_i. The next step is to *uniformize* these coefficients. That is, let A be the least common multiple of these coefficients and use the following easily provable equivalences to make sure that x only occurs as $\underline{A}x$ in our formula ϕ:

$$\begin{aligned} t = s &\leftrightarrow \underline{c}t = \underline{c}s; \\ t < s &\leftrightarrow \underline{c}t < \underline{c}s; \\ t \equiv_n s &\leftrightarrow \underline{c}t \equiv_{cn} \underline{c}s, \end{aligned}$$

   where $c > 0$ is an integer, and t and s are $\mathcal{L}$-terms.

3. Next we eliminate the coefficient of x altogether by replacing $\underline{A}x$ everywhere by x and adding as a new conjunct: $x \equiv_A \underline{0}$. This obviously gives an equivalent formula, and now x only occurs in ϕ with coefficient 1.

4. We distinguish two cases.
   (a) Some α_j is an equality, say $x + t = s$. Then there is only one option for x, so we can eliminate x from all the other α_i's by replacing x by s and adding t on the other side. Now we can bring all these α_i out of the existential quantifier. We are left with $\exists x\ x + t = s$, but this is equivalent to $t < s \vee t = s$. We have arrived at a quantifier-free formula.
   (b) Equality does not occur. Now ϕ is $\exists x\ \theta$, where θ is of the form

$$\bigwedge_{0 \leq i < k} x + t_i > s_i \wedge \bigwedge_{0 \leq i < l} x + r_i < u_i \wedge \bigwedge_{0 \leq i < m} x + v_i \equiv_{n_i} w_i.$$

   Here k, l, m and the n_i are natural numbers, the n_i are positive, and the t_i, s_i, r_i, u_i, v_i and w_i are $\mathcal{L}$-terms not containing x. We say that θ consists of *lower bounds*, *upper bounds*, and *congruences* respectively.

Let N be the least common multiple of all the n_i, which is well-defined because there is at least one congruence, namely the one introduced in step 3. For simplicity, we assume without loss of generality that one of the lower bounds is $x+1>0$. We can do this because if this lower bound isn't present, then we may simply add it, since $x+1>0$ is provable in $\mathtt{PrA}$. We can furthermore suppose that it is the first one: t_0 is 1, s_0 is 0. We claim that $\mathtt{PrA}$ proves: if there is an x such that θ, then there is such an x satisfying the following disjunction of equalities:

$$\bigvee_{0\leq i<k} \Big[(x+t_i=s_i+\underline{1}) \vee (x+t_i=s_i+\underline{2}) \vee \ldots \vee (x+t_i=s_i+\underline{N})\Big]. \qquad (2)$$

We will prove this in lemma 1.4.2. Now we can simply add (2) as a conjunct to θ, then distribute the $\wedge$'s over the $\vee$'s, and finally distribute the $\exists x$ over the $\vee$'s. We then end up with a disjunction of formulas of the form we had at the beginning of case 4. The difference is that all these formulas now contain an equality, and the problem is reduced to case 4a. □

It remains to prove the missing lemma. In the proof we will freely use facts about the ordering; all of these follow quite easily from proposition 1.1.1.

Lemma 1.4.2. *Let θ be the formula*

$$\bigwedge_{0\leq i<k} x+t_i>s_i \wedge \bigwedge_{0\leq i<l} x+r_i<u_i \wedge \bigwedge_{0\leq i<m} x+v_i \equiv_{n_i} w_i,$$

where k, l, m and the n_i are natural numbers, the n_i are positive, and the t_i, s_i, r_i, u_i, v_i and w_i are $\mathcal{L}$-terms not containing x. Then $\mathtt{PrA}$ proves: if there is an x such that θ, then there is such an x for which (2)*, i.e.*

$$\bigvee_{0\leq i<k} \Big[(x+t_i=s_i+\underline{1}) \vee (x+t_i=s_i+\underline{2}) \vee \ldots \vee (x+t_i=s_i+\underline{N})\Big],$$

holds.

Proof. Work in some model M of $\mathtt{PrA}$ and let x be such that $M \models \theta$. For $0\leq i<k$, we have $M \models x+t_i>s_i$, so we can select some $y_i \in M$ such that $M \models x+t_i = s_i+(y_i+1)$. Let y_j be the smallest of these y_i. Since $M \models \mathtt{PrAx8}_N$, we can select $u \in M$ and $r \in \mathbb{N}$ such that $M \models y_j = \underline{N}u+\underline{r}$ and $r<N$. Because we assumed that $x+1>0$ was the first lower bound, we get $M \models x+1 = 0+(y_0+1)$, so $M \models x = y_0 \geq y_j = \underline{N}u+\underline{r} \geq \underline{N}u$. This means we can select some $x_0 \in M$ such that $x = x_0+\underline{N}u$. We will show that x_0 satisfies θ and (2).
For the lower bounds: for $0\leq i<k$ we have

$$M \models (x_0+t_i)+y_j \geq (x_0+t_i)+\underline{N}u = (x_0+\underline{N}u)+t_i = x+t_i = s_i+(y_i+1) = (s_i+1)+y_i.$$

Because also $M \models y_j \leq y_i$, we can now easily see that $M \models x_0+t_i \geq s_i+1$, or equivalently $M \models x_0+t_i>s_i$.
For the upper bounds: just note that $M \models x_0 \leq x$.
For the congruences: for $0\leq i<m$ we have $N = n_i k_i$ for some $k_i \in \mathbb{N}$, and now we easily see $M \models x = x_0+\underline{N}u = x_0+\underline{n_i}\left(\underline{k_i}u\right)$, so $M \models x \equiv_{n_i} x_0$ and it is clear that x_0 satisfies the congruences.
For (2): we have

$$M \models (x_0+\underline{N}u)+t_j = x+t_j = s_j+(y_j+1) = s_j+(\underline{N}u+\underline{r}+1) = (s_j+(\underline{r+1}))+\underline{N}u,$$

and substracting $\underline{N}u$ gives $M \models x_0+t_j = s_j+(\underline{r+1})$, where $r+1$ is some number v such that $1\leq v\leq N$. □

Note that the x_0 we constructed is simply the smallest x such that θ holds. We couldn't just *define* x_0 to be this smallest solution, because we don't have a least number principle for `PrA` (yet). The above proof makes it clear how this smallest solution can be found anyway, using the division axioms.

1.5 Decidability and completeness

In the above procedure, it was crucial that we were working in the language $\mathcal{L}$ and not in $\mathcal{L}^-$. Indeed, while `PrA` can easily be formulated in $\mathcal{L}^-$, it does *not* allow quantifier elimination in this language. Without $<$, we cannot get rid of the last remaining conjunct in case 4a, and without congruences, we don't have any tools to eliminate the coefficient of x in case 3. On the other hand, the presence of bounds and congruences presented us with the difficult case 4b, but it turned out to be quite feasible. Now let us turn to the implications of the quantifier elimination.

Theorem 1.5.1. *`PrA` is complete, in the sense that it decides all sentences.*

Proof. Given some sentence ϕ, apply quantifier elimination and obtain some equivalent ψ that is a truth-functional combination of atomic $\mathcal{L}$-sentences. Now apply corollary 1.1.4 and the fact that propositional logic is complete. □

Corollary 1.5.2. *`PrA` is the true theory of the standard model.*

Proof. By the preceding theorem, any extension of `PrA` must be equal to `PrA`. But the true theory of the standard model clearly extends `PrA`, so the claim is obvious. □

By inspecting the proofs in sections 1.3 and 1.4 carefully, one can see that finding quantifier-free equivalents is an automatic, or recursive, or effective, process. We get:

Theorem 1.5.3. *`PrA` is decidable.*

Proof. Any recursively axiomatized complete theory is decidable, but the quantifier elimination gives a much more insightful decision procedure. Again, given some sentence ϕ, apply quantifier elimination and *effectively* obtain some equivalent ψ that is a truth-functional combination of atomic $\mathcal{L}$-sentences. Now all terms are closed, and therefore denote something in $\mathbb{N}$. It is not hard to see that we can effectively find these denotations. Since the function $+$ and the relations $=$, $<$ and $\equiv_n$ are recursive, we can effectively find the truth-values of these atomic sentences in $\mathbb{N}$. Since propositional logic is decidable, we can find the truth value of ψ, and hence of ϕ, in the standard model. By corollary 1.5.2, this determines its provability in `PrA`. □

The above described decision procedure isn't too fast, however. If quantifiers occur nested, then the complicated procedure from the proof of theorem 1.4.1 must be carried out repeatedly. This causes the running time to be multi-exponential.

As we mentioned at the beginning of section 1.3, the above results put the theory of $\mathbb{N}$ with addition in a rather different situation than the theory of $\mathbb{N}$ with both addition and multiplication. Exactly how close is `PrA` to the latter theory? Given only addition, we can already define many things in $\mathbb{N}$: zero, one, the ordering and congruence modulo n for every

$n \geq 1$. A natural next step would be the divisibility relation, which essentially uniformizes the infinitely many congruence relations. We end this section by showing that, in this sense, `PrA` is very close to the theory of $\mathbb{N}$ with addition and multiplication. The following result is due to Julia Robinson, who proves a somewhat stronger result in [4]. We are not concerned with this stronger result here, so we can present a simpler proof, due to the author.

Theorem 1.5.4. *Let $\mathcal{L}^+$ be the language $\mathcal{L}$ expanded by a binary relation symbol $|$, which is to be interpreted in $\mathbb{N}$ as the divisibility relation.*[4] *Then the multiplication relation, i.e. $x \cdot y = z$, is definable in the structure $\langle \mathbb{N}, \mathcal{L}^+ \rangle$. As a result, the divisibility relation is not definable in the standard model $\langle \mathbb{N}, \mathcal{L} \rangle$.*

Proof. First of all, we define the ternary relation $C(x, y, z)$,[5] which holds iff $(x+y)(x+y+1) = z$. The definition is as follows:

$$\begin{aligned} C(x,y,z) :\Leftrightarrow\ & (x = 0 \wedge y = 0 \wedge z = 0) \vee \\ & \Big[(x \neq 0 \vee y \neq 0) \wedge z \neq 0 \wedge (x + y \mid z) \wedge ((x+y)+1 \mid z) \\ & \wedge \forall w\ ((w \neq 0 \wedge (x+y \mid w) \wedge ((x+y)+1 \mid w)) \to z \leq w)\Big]. \end{aligned}$$

In other words, if $x + y$ is nonzero, then z is the smallest positive w satisfying $x + y \mid w$ and $x + y + 1 \mid w$. This is indeed $(x + y)(x + y + 1)$, since $\gcd(x + y, x + y + 1) = 1$.[6] Before defining multiplication, we first need the square relation $y = x^2$, definable as:

$$y = x^2 :\Leftrightarrow \exists z\ (C(x, 0, z) \wedge x + y = z).$$

Now multiplication can be defined as:

$$x \cdot y = z :\Leftrightarrow \exists w\ (C(x, y, w) \wedge x^2 + y^2 + x + y + z + z = w).$$

Both definitions can be verified by expanding both sides of the equality.
Thus, the decision problem for the theory of $\langle \mathbb{N}, +, \cdot \rangle$ reduces to that for the theory of $\langle \mathbb{N}, \mathcal{L}^+ \rangle$. Since the former is known to be undecidable, the latter must be undecidable as well. Now if the divisibility relation were definable in $\langle \mathbb{N}, \mathcal{L} \rangle$, then the decision problem for the theory of $\langle \mathbb{N}, \mathcal{L}^+ \rangle$ would reduce to that for the theory of the standard model, i.e. to that of `PrA`. But the former is undecidable, while the latter is decidable, which is impossible. □

[4] By convention, $0 \mid n$ iff $n = 0$.

[5] This approach was inspired by the Cantor pairing function.

[6] More generally, we can define $\operatorname{lcm}(x, y) = z$ in this way. We can also define $\gcd(x, y) = z$ in a similar fashion.

2 Interpreting PrA

We begin this chapter by introducing the second ingredient of this thesis: interpretations. Next, we prove a result that was already known and continue the investigation from there. This will lead to the formulation of two conjectures.

2.1 Interpretations

In this section, we introduce the concept of an interpretation, which we will apply to Presburger Arithmetic later. More specifically, we will deal with what are called *one-dimensional parameter-free* interpretations. Throughout this section, U and V will be theories in the languages K and L respectively, where K and L *only contain predicate symbols as non-logical symbols.* In other words, function symbols are not allowed. This may seem a restriction at first, but it will turn out that this restriction isn't essential.

Definition 2.1.1. A *translation* τ from K to L is a quadruple $\langle K, \mathcal{D}, \mathcal{F}, L \rangle$, where $\mathcal{D}$ is some L-formula in one free variable, and where $\mathcal{F}$ assigns to every predicate symbol P in K some L-formula $\mathcal{F}(P)(x_0, \ldots, x_{n-1})$, where n is the arity of P. These $\mathcal{F}(P)$ should have the property that

$$\mathcal{F}(P)(x_0, \ldots, x_{n-1}) \to \bigwedge_{i=0}^{n-1} \mathcal{D}(x_i) \tag{3}$$

is provable in predicate logic. We call $\mathcal{D}$ the *domain formula* and we can write $\tau : K \to L$ to indicate that τ is from K to L.

As stated in the introduction, we consider identity to be an always present predicate. So note that, in particular, it is not necessary that $\mathcal{F}$ sends identity in K to identity in L.

A translation is in fact nothing more than a piece of information that allows us to translate K-formulas into L-formulas. Let us make this precise.

Definition 2.1.2. Given a translation $\tau : K \to L$, we define the function $(\cdot)^\tau$ from the set of K-formulas to the set of L-formulas by recursion.

- For every predicate symbol P of K, we define $(P(x_0, \ldots, x_{n-1}))^\tau$ as $\mathcal{F}(P)(x_0, \ldots, x_{n-1})$, where $x_0, \ldots, x_{n-1}$ are variables.[7]
- $(\cdot)^\tau$ commutes with the propositional connectives.
- For K-formulas A, we define $(\forall x\ A)^\tau$ as $\forall x\ \ (\mathcal{D}(x) \to A^\tau)$.
- For K-formulas A, we define $(\exists x\ A)^\tau$ as $\exists x\ \ (\mathcal{D}(x) \wedge A^\tau)$.

One can notice something odd about this translation procedure, namely that everything seems to be restricted to things satisfying $\mathcal{D}$. Indeed, by (3), the translated predicates can only hold for things satisfying $\mathcal{D}$, and moreover, the quantifiers are relativized to $\mathcal{D}$. We will soon see what this restriction means, when we consider the matter from a semantical point of view.

[7] Note that variables are the only K-terms, since function symbols are not present.

Until this point, we have only considered the languages K and L. Now let us bring the theories U and V into the discussion.

Definition 2.1.3. An *interpretation* ι of U in V is a triple $\langle U, \tau, V\rangle$, where $\tau : K \to L$ is a translation satisfying $U \vdash A \Rightarrow V \vdash A^\tau$ for all K-sentences A. We say that ι is based on τ, and we can write $\iota : U \to V$ to indicate that ι is an interpretation of U in V.

What exactly is happening here? Suppose that we are living in V, and that we are confronted with the theory U, perhaps written in an entirely different language. Now we interpret the concepts that U mentions in some way (this is the translation), such that they become provable in our world, i.e. in V (and this makes it an interpretation). One should note that one of the concepts of U that we interpret is 'being an element'; we take that to be 'satisfying $\mathcal{D}$'. In this way, although U may be something entirely different, we can act as if our world V were about U.

Interpreting is not an unnatural concept. In fact, set theory interprets other theories all the time. While set theory is, quite obviously, meant to be about sets, we can act as if it is, for example, also about the natural numbers. A more trivial example is given by the theory of linear orders in the language $\langle < \rangle$, and the theory of linear orders in the language $\langle \leq \rangle$. These interpret each other simply because $<$ and $\leq$ are interdefinable. We can even look at theories that interpret themselves. Consider, for example, duality in projective geometry. Given some provable sentence, we can switch the concepts 'point' and 'line', and obtain another one. This is a very informative self-interpretation of projective geometry. In fact, all theories have at least one self-interpretation, namely the following trivial one.

Definition 2.1.4. The *identity translation* $\mathrm{id}_K : K \to K$ is the translation we obtain by taking $\mathcal{D}$ to be some tautology and by taking $\mathcal{F}(P)$ to be just P. The *identity interpretation* $\mathrm{id}_U : U \to U$ is the interpretation based on id_K. We can drop the subscripts of id if the language or theory is clear from the context.

We also have a notion of composition of two interpretations. Let W be a theory in the language J.

Definition 2.1.5. Let $\tau : K \to L$ and $\sigma : L \to J$ be translations. We define their *composition* $\sigma \circ \tau : K \to J$ as follows:

- the domain formula $\mathcal{D}_{\sigma\circ\tau}(x)$ is given by $\mathcal{D}_\sigma(x) \wedge (\mathcal{D}_\tau(x))^\sigma$;
- for a predicate symbol P in K, its translation $\mathcal{F}_{\sigma\circ\tau}(P)$ is given by

$$\bigwedge_{i=0}^{n-1} \mathcal{D}_{\sigma\circ\tau}(x_i) \ \wedge (\mathcal{F}_\tau(P)(x_0, \ldots, x_{n-1}))^\sigma .$$

One can quite easily prove that, for all K-formulas, $(A^\tau)^\sigma$ is equivalent to $A^{\sigma\circ\tau}$ in W. So if $\iota : U \to V$ and $\kappa : V \to W$ are interpretations, then we have

$$U \vdash A \Rightarrow V \vdash A^\tau \Rightarrow W \vdash (A^\tau)^\sigma \Rightarrow W \vdash A^{\sigma\circ\tau}$$

for all K-sentences A, so $\langle U, \sigma \circ \tau, W\rangle$ is an interpretation as well. We shall call this interpretation the *composition* of ι and κ, and we denote it by $\kappa \circ \iota$.

Until now our investigations have been purely syntactical, and it is quite useful to consider interpretations from a semantical point of view as well. Suppose we have an interpretation $\iota : U \to V$ based on a translation $\tau : K \to L$, and a model M of V. We can now construct a model N of U in the following way. As domain we take the set $\{x \in M \mid M \models \mathcal{D}(x)\}$; by abuse of notation, we will denote this set also by $\mathcal{D}$. For a predicate symbol P in K, we take the extension of P in N to be the extension of $\mathcal{F}(P)$ in M. Note that this makes sense because of (3). One can easily prove by induction on complexity that for a K-sentence A, we have $N \models A$ iff $M \models A^\tau$. This means that for every K-sentence A, we have

$$U \vdash A \Rightarrow V \vdash A^\tau \Rightarrow M \models A^\tau \Rightarrow N \models A,$$

so the model N indeed satisfies U. In particular, $\mathcal{F}(=_K)$ is an equivalence relation that respects all the $\mathcal{F}(P)$. So although identity is in N not necessarily interpreted as the 'real' identity, i.e. the one inherited from M, it is a permissible notion of identity. We call N the *inner model* in M given by ι.

An important question concerning interpretations is when we consider two interpretations to be the same. A natural first attempt would be: two interpretations $\iota, \kappa : U \to V$ with underlying translations resp. $\tau, \sigma : K \to L$ are the same if $\mathcal{D}_\tau$ and $\mathcal{D}_\sigma$ are equivalent in V, and $\mathcal{F}_\tau(P)$ and $\mathcal{F}_\sigma(P)$ are also equivalent in V for all predicate symbols P in K. This is, however, a very strict notion of sameness, and there are more useful ones.[8] The semantical discussion above provides a hint: given two interpretations from U two V and a model M of V, we can consider the inner models in M. As far as M is concerned, the two interpretations are the same if these inner models are isomorphic. We also demand the isomorphism to be representable in the language of M, so that the notion can be formulated completely in terms of M.

Definition 2.1.6. Let M be a model of V. Two interpretations $\iota, \kappa : U \to V$ based on respectively $\tau, \sigma : K \to L$ are *representably isomorphic in M* if there is some L-formula $F(x, y)$ such that the following formulas are valid in M:

- $(\exists y\ F(x, y)) \leftrightarrow \mathcal{D}_\tau(x)$;
- $(\exists x\ F(x, y)) \leftrightarrow \mathcal{D}_\sigma(y)$;
- $\bigwedge_{i=0}^{n-1} F(x_i, y_i) \to (\mathcal{F}_\tau(P)(x_0, \ldots, x_{n-1}) \leftrightarrow \mathcal{F}_\sigma(P)(y_0, \ldots, y_{n-1}))$ for all predicate symbols P in K.

We write $M \models F : \iota \cong \kappa$ ("M models F to be an isomorphism between ι and κ").[9]

It takes some effort to see that the above constraints indeed express the fact that F represents an isomorphism of models. The first constraint says that we have the right domain, and that every element had at least one image. The second constraint says that we have the right codomain and that our function is surjective. By the third constraint, we have in particular $(F(x_0, y_0) \wedge F(x_1, y_1)) \to (\mathcal{F}_\tau(=_K)(x_0, x_1) \leftrightarrow \mathcal{F}_\sigma(=_K)(y_0, y_1))$, which means exactly that our function is well-defined and injective; it is important to realize that these two concepts are relative to a notion of identity. The remainder of the third constraint of course says that structure is preserved.

[8] This notion *is* interesting, however, when we consider the category of interpretations. This is the category with theories as objects and interpretations *modulo this strict notion of sameness* as arrows. We need to consider interpretations modulo this notion, because otherwise composition isn't associative and identity doesn't behave like identity. Even worse, identity wasn't even unique as we defined it, because we took the domain formula to be 'some tautology'.

[9] We will use $F : \iota \cong \kappa$ as an abbreviation for the conjunction of the listed formulas.

It is now easy to give a syntactical notion of sameness of interpretations as well.

Definition 2.1.7. Two interpretations $\iota, \kappa : U \to V$ based on respectively $\tau, \sigma : K \to L$ are *provably isomorphic* if there is some L-formula $F(x, y)$ such that the formulas mentioned in definition 2.1.6 are provable in V. We write $V \vdash F : \iota \cong \kappa$ ("V proves F to be an isomorphism between ι and κ").

This syntactical notion actually uniformizes the semantical model-relative notion of sameness, in the sense that two interpretations are provably isomorphic in V if and only if they are representably isomorphic in every model of V. We will not prove this result here, but we can mention that it involves a compactness argument. If two interpretations are isomorphic, representably in a model or provably, then the model respectively theory knows that these two interpretations behave in similar fashions. Let us make this precise.

Lemma 2.1.1. *Suppose $\iota, \kappa : U \to V$ are interpretations based on translations $\tau, \sigma : K \to L$ respectively. Then for every K-formula A in n free variables, we have*

$$\vdash \left((F : \iota \cong \kappa) \wedge \bigwedge_{i=0}^{n-1} F(x_i, y_i) \right) \to \Big((A(x_0, \ldots, x_{n-1}))^\tau \leftrightarrow (A(y_0, \ldots, y_{n-1}))^\sigma \Big).$$

In particular, we have $\vdash (F : \iota \cong \kappa) \to (A^\tau \leftrightarrow A^\sigma)$ for all K-sentences A.

Proof. We prove this by induction on the complexity of A. The atomic case is just the third clause of definition 2.1.6, and the propositional clauses are trivial because $(\cdot)^\tau$ and $(\cdot)^\sigma$ commute with the propositional connectives. Since $(\exists x\ B)^\tau$ is equivalent (in predicate logic) to $(\neg \forall x\ \neg B)^\tau$ for all K-sentences B, and similarly for σ, it suffices to prove the theorem for $\forall x_0 A(x_0, \ldots, x_{n-1})$, given that it holds for $A(x_0, \ldots, x_{n-1})$, where $n \geq 1$.
Let M be some L-structure such that $M \models (F : \iota \cong \kappa) \wedge \bigwedge_{i=1}^{n-1} F(x_i, y_i)$ and such that $M \models (\forall x_0\ A(x_0, \ldots, x_{n-1}))^\tau$, i.e. $M \models \forall x_0\ (\mathcal{D}_\tau(x_0) \to (A(x_0, \ldots, x_{n-1}))^\tau)$. Furthermore, let $y_0 \in M$ be an arbitrary element satisfying $M \models \mathcal{D}_\sigma(y_0)$. Now there must be an $x_0 \in M$ such that $M \models F(x_0, y_0)$, because F represents an isomorphism. By the induction hypothesis, we have

$$M \models (A(x_0, \ldots, x_{n-1}))^\tau \leftrightarrow (A(y_0, \ldots, y_{n-1}))^\sigma.$$

Moreover, since F represents an isomorphism, we see that $M \models \mathcal{D}_\tau(x_0)$ and this gives us $M \models (A(x_0, \ldots, x_{n-1}))^\tau$. So we get $M \models (A(y_0, \ldots, y_{n-1}))^\sigma$. Since y_0 was arbitrary such that $M \models \mathcal{D}_\sigma(y_0)$, we may conclude that $M \models \forall y_0\ (\mathcal{D}_\sigma(y_0) \to (A(y_0, \ldots, y_{n-1}))^\sigma)$, which is just $M \models (\forall y_0\ A(y_0, \ldots, y_{n-1}))^\sigma$. By the completeness theorem,

$$\vdash \left((F : \iota \cong \kappa) \wedge \bigwedge_{i=1}^{n-1} F(x_i, y_i) \right) \to ((\forall x_0\ A(x_0, \ldots, x_{n-1}))^\tau \to (\forall y_0\ A(y_0, \ldots, y_{n-1}))^\sigma).$$

The other direction is proven analogously, and this completes the induction. □

Note that the theorem makes a statement about provability in predicate logic; it doesn't really matter what U and V are here. In particular, we didn't even use that ι and κ are interpretations. We end this section by making a few remarks about more general interpretations. We only consider one-dimensional parameter-free interpretations. Our interpretations are one-dimensional because our domain formula has exactly one free variable; as a result, an inner model N consists of elements of the original M. We can also take the domain formula to be a formula in m free variables for some $m > 1$, and translate n-ary predicates to formulas in mn free variables. Now inner models are composed of m-tuples of elements of the original model. There is a slight technicality with more-dimensional interpretations, because one needs a lot of new variables. But of course, such problems are solvable in predicate logic.

In an interpretation with parameters, we may first pick some elements satisfying certain constraints at random (these are the parameters), and then specify the domain formula and translations of predicate symbols. An elegant example is the interpretation of hyperbolic geometry in Euclidian geometry via the Poincaré disk model; we first need to pick two distinct points, before we can specify the actual domain and what our points, lines, distances, etc are.

2.2 Local interpretability for PrA in PrA$^-$

Before we can apply the concepts from the previous section to our study of PrA, we need to solve the problem that the languages from the previous section were not allowed to contain functions, while $\mathcal{L}^-$ *does* contain them. Fortunately, there is a way around this; we can also formulate PrA and its related theories in a language containing as non-logical symbols only the predicates $+(x, y, z)$, $0(x)$ and $1(x)$, which hold iff $x+y = z$, $x = 0$ and $x = 1$ respectively. It is quite obvious that we can translate statements in one language into a statement in the other language. For example, the $\mathcal{L}^-$-formula $x + 0 = y + z$ can be translated as

$$\exists u \exists v\ (+(x, u, v) \wedge +(y, z, v) \wedge 0(u)).$$

Translations the other way are even more obvious.[10] So, when we talk about interpretations from or to a theory in the language $\mathcal{L}^-$, we assume that we first formulate our starting theory in a function-free language, then carry out the actual translation on which the interpretation is based, and finally translate back to $\mathcal{L}^-$. In practice, this will provide no difficulties.

We now get to the central result of this section. It may happen that a theory is able to interpret some stronger theory in the same language. In fact, we have a perfect example in stock, as the following result from [6] shows.

Theorem 2.2.1. *For all positive integers n, PrA$^-$ interprets PrA$_n$.*

Proof. The domain $\mathcal{D}_n$ of our translation τ_n is given by

$$\forall z \leq x\ ((z \equiv_n \underline{0}) \vee (z \equiv_n \underline{1}) \vee \ldots \vee (z \equiv_n \underline{n-1})).$$

For $\mathcal{F}_n$, we take the identity translation restricted to $\mathcal{D}_n$, i.e.

- $\mathcal{F}_n(0)$ is $\mathcal{D}_n(x) \wedge 0(x)$;
- $\mathcal{F}_n(1)$ is $\mathcal{D}_n(x) \wedge 1(x)$;
- $\mathcal{F}_n(+)$ is $\mathcal{D}_n(x) \wedge \mathcal{D}_n(y) \wedge \mathcal{D}_n(z) \wedge +(x, y, z)$;
- $\mathcal{F}_n(=)$ is $\mathcal{D}_n(x) \wedge \mathcal{D}_n(y) \wedge x = y$.

We need PrA$^-$ to prove $(\exists x\ 0(x))^{\tau_n}$, $(\exists x\ 1(x))^{\tau_n}$ and $(\forall x \forall y \exists z\ +(x, y, z))^{\tau_n}$.[11] Writing this out, we see that these three statements just claim that 0 and 1 belong to $\mathcal{D}_n$ and that $\mathcal{D}_n$ is

[10] This is just the introduction of Skolem functions.

[11] We need to check this because PrA$_n$ proves the functionality of 0, 1 and +. Of course, we also need uniqueness of 0, 1 and $x + y$, but this follows because they were already unique in PrA$^-$.

closed under addition. The first two claims are rather obvious in PrA$^-$, so let us check that $\mathcal{D}_n$ is provably closed under addition. Work in a model M of PrA$^-$ and consider $x_0, x_1, z \in M$ such that $M \models \mathcal{D}_n(x_0)$, $M \models \mathcal{D}_n(x_1)$ and $M \models z \leq x_0 + x_1$. We have to prove that there are a $y \in M$ and an $r \in \mathbb{N}$ such that $M \models z = \underline{n}y + \underline{r}$ and $r < n$.
If $M \models z \leq x_0$, then this is immediately clear. Therefore, we can suppose $M \models x_0 + u = z$ for some $u \in M$. We have $M \models x_0 + u = z \leq x_0 + x_1$, whence $M \models u \leq x_1$. Now there exist $y_0, y_1 \in M$ and $r_0, r_1 \in \mathbb{N}$ such that $M \models x_0 = \underline{n}y_0 + \underline{r_0}$, $M \models u = \underline{n}y_1 + \underline{r_1}$ and $r_0, r_1 < n$. We get:

$$M \models z = x_0 + u = (\underline{n}y_0 + \underline{r_0}) + (\underline{n}y_1 + \underline{r_1}) = \underline{n}(y_0 + y_1) + \underline{r_0 + r_1}.$$

If $0 \leq r_0 + r_1 < n$, then we can take $y = y_0 + y_1$ and $r = r_0 + r_1$. Otherwise, we have $n \leq r_0 + r_1 < 2n$ and we can take $y = y_0 + y_1 + 1$ and $r = r_0 + r_1 - n$.
It remains to check that the axioms of PrA$_n$ hold inside $\mathcal{D}_n$. But this is easy; we can bound every existential quantifier among the axioms of PrA$^-$ from above by some open $\mathcal{L}^-$-term, and the the validity of PrA$^-$ then follows because $\mathcal{D}_n$ is downwards closed under $\leq$.[12] For PrAx8$_n$: by the definition of $\mathcal{D}_n$ there are for every $x \in \mathcal{D}_n$ some $y \in M$ and $r \in \mathbb{N}$ such that $M \models x = \underline{n}y + \underline{r}$ and $r < n$. It is easy to see that $M \models y \leq x$, so we also have $M \models \mathcal{D}_n(y)$, again because $\mathcal{D}_n$ is downwards closed, and we are done. □

In particular, every finitely axiomatizable subtheory of PrA can be interpreted in PrA$^-$. We also say that PrA is *locally interpretable* in PrA$^-$. Since interpretability is *not* a compact concept and PrA is *not* finitely axiomatizable, it remains open whether the whole of PrA can be interpreted in PrA$^-$. We will take up this question in the next section.

2.3 Interpreting PrA in PrA$^-$

In this section we consider the question whether PrA is interpretable in PrA$^-$. As we noted at the end of the previous sections, the fact that PrA is not finitely axiomatizable has something to do with this. Indeed, *were* PrA finitely axiomatizable, then we would have PrA = PrA$_N$ for some integer $N \geq 1$, and the affirmative answer to our question is given by theorem 2.2.1. But since this is not the case, we conjecture the answer to be *no*.

Conjecture 2.3.1. *The theory PrA is not interpretable in PrA$^-$.*

Note that this conjecture expresses a strengthening of theorem 1.2.4.

We will attack our conjecture by considering interpretations of PrA in PrA itself, i.e. self-interpretations of PrA. The reason for doing so is that, unlike interpretations of PrA in PrA$^-$, we actually expect self-interpretations of PrA to exist, which is convenient when we wish to consider examples. In fact, we know of a self-interpretation of PrA, namely the identity interpretation. Are there others? Consider the following example.

Example 2.3.1. Let τ be the translation from the function-free variant of $\mathcal{L}^-$ to itself given by:

- The domain formula $\mathcal{D}$ is $x \equiv_2 0$, which is of course an abbreviation of $\exists y \ +(y, y, x)$;
- We translate 0, + and = in the trivial manner, i.e. $0(x)$ is translated as $\mathcal{D}(x) \wedge 0(x)$, $+(x, y, z)$ as $\mathcal{D}(x) \wedge \mathcal{D}(y) \wedge \mathcal{D}(z) \wedge +(x, y, z)$ and $x = y$ as $\mathcal{D}(x) \wedge \mathcal{D}(y) \wedge x = y$.

[12] Note that this is not just in PrAx5; there is also a 'hidden' existential quantifier in PrAx7. When dealing with PrAx8$_n$, we shall also have to consider the fact that such a hidden quantifier is present.

- We translate $1(x)$ as $\mathcal{D}(x) \wedge x = \underline{2}$, where $x = \underline{2}$ of course abbreviates the formula $\exists y\ (1(y) \wedge +(y, y, x))$.

Now $\iota := \langle \mathtt{PrA}, \tau, \mathtt{PrA} \rangle$ is an interpretation.

When we consider inner models of this interpretation, it becomes clear what ι does. Given a model M of $\mathtt{PrA}$, our ι picks out the even elements and preserves addition. As a result, order is preserved, as we can easily check. The first element, i.e. 0, is the interpretation of 0, while the second element, i.e. 2, is the interpretation of 1.

This inner model certainly looks a lot like the original M, and we can indeed give a *provable* isomorphism $\iota \to \mathrm{id}_{\mathtt{PrA}}$: take $F(x, y)$ to be $x = \underline{2}y$. Checking that $\mathtt{PrA} \vdash F : \iota \cong \mathrm{id}$ isn't too difficult. We thus see that ι is provably isomorphic to the identity interpretation; we also say that ι is *provably trivial.* When dealing with a certain model, we can say that ι is *representably trivial* if in M, it is representably isomorphic to the identity interpretation. $\Diamond$

The above example is a bit simple, but we can give more complicated ones.

Example 2.3.2. Take as domain all numbers congruent to 0 or 1 modulo 3, send 0, 1 and $=$ just to 0, 1 and $=$, and translate $x + y$ as $\begin{cases} x + y \text{ if } x \equiv_3 0 \text{ or } y \equiv_3 0; \\ x + y + 1 \text{ if } x \equiv_3 1 \text{ and } y \equiv_3 1. \end{cases}$

Something like this can clearly be made into a translation, and one may prove that it yields an interpretation. What happens here is that we delete all elements congruent to 2 modulo 3, then jam the remaining elements together, and finally define 0, 1 and + as if nothing had happened. In other words, the elements congruent to 0 modulo 3 play the role of the even numbers, while the elements congruent to 1 modulo 3 play that of the odd numbers. In the standard model we may picture this by the following array, with on the first line the standard model, and with on the second line the standard model with elements congruent to 2 modulo 3 removed.

$$\begin{array}{ccccccccccccccccccc}
0 & 1 & & 2 & 3 & & 4 & 5 & & 6 & 7 & & 8 & 9 & & 10 & 11 & & \dots \\
0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 & 17 & \dots
\end{array}$$

Our translations of 0, 1, = and + are now the ones induced on the second line by the first line.

So is this interpretation provably trivial? Well, if we view the first line as representing the trivial translation, then we can the above array tells us immediately what the isomorphism from our interpretation to the identity should be. We should send elements of the form $3a$ to $2a$ and elements of the forms $3a + 1$ to $2a + 1$. But such a function is expressible in $\mathcal{L}$ as

$$(x \equiv_3 0 \wedge \underline{2}x = \underline{3}y) \vee (x \equiv_3 1 \wedge \underline{2}x + \underline{1} = \underline{3}y).$$

It isn't hard to show that this is a provable isomorphism, so this interpretation is provably trivial as well. $\Diamond$

We will give one more example, which shows, unlike the previous two examples, that we can distort the ordering in interesting ways.

Example 2.3.3. As domain, we take some tautology and we send identity to identity. We again picture our interpretation by an array.

$$\begin{array}{cccccccccccccccccccc}
0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 & 17 & \dots \\
0 & 2 & 1 & 4 & 6 & 3 & 8 & 10 & 5 & 12 & 14 & 7 & 16 & 18 & 9 & 20 & 22 & 11 & \dots
\end{array}$$

Again 0, 1 and + are induced from the first line. It takes some effort to see that addition is actually definable in $\mathcal{L}^-$. We will not give the definition here, because it is a bit of a mess, but it is a useful exercise.[13] Under this translation, all axioms of `PrA` hold in $\mathbb{N}$, and this means that we indeed have an interpretation, since `PrA` is the theory of the standard model. This interpretation is interesting because of the peculiar ordering; the odd numbers are running away twice as fast as the even numbers.

Again, the array suggests that this interpretation is trivial. One may check that in $\mathbb{N}$, an isomorphism from our interpretation to the identity is given by

$$(x \equiv_4 0 \wedge \underline{3}x = \underline{4}y) \vee (x \equiv_4 2 \wedge \underline{3}x = \underline{4}y + \underline{2}) \vee (x \equiv_2 1 \wedge \underline{3}x + \underline{1} = \underline{2}y),$$

so this interpretation is provably trivial, again because `PrA` is the theory of $\mathbb{N}$. ◊

One may search for further examples, but all of these will prove to be provably trivial. We thus make another conjecture:

Conjecture 2.3.2. *Every self-interpretation of `PrA` is provably trivial.*

Our conjecture 2.3.1 turns out to follow from this one, as the following general result shows.

Theorem 2.3.3. *Let U be some theory such that all its self-interpretations are provably trivial. If U is interpretable in one of its finitely axiomatizable subtheories, then U itself is finitely axiomatizable as well.*

Proof. Let $\iota : U \to U_0$ be an interpretation based on the translation $\tau : K \to K$, where $U_0 \subset U$ is finitely axiomatizable and K is the language of U. For all K-sentences A, we have

$$U \vdash A \Rightarrow U_0 \vdash A^\tau \Rightarrow U \vdash A^\tau,$$

so $\kappa := \langle U, \tau, U \rangle$ is a self interpretation of U. By our assumption, κ is provably trivial, i.e. there is some K-formula $F(x, y)$ such that $U \vdash F : \kappa \cong \mathrm{id}$. Since $F : \kappa \cong \mathrm{id}$ can be expressed as a finite K-statement, it must have a finite proof in U, and therefore it must be provable in some finitely axiomatizable $U_1 \subset U$. By lemma 2.1.1, we have $\vdash (F : \kappa \cong \mathrm{id}) \to \big(A^\tau \leftrightarrow A^{\mathrm{id}}\big)$ for all K-sentences A. Because A^{id} is obviously equivalent (in predicate logic) to A itself, we have $U_1 \vdash A^\tau \leftrightarrow A$ for all K-sentences A.
Now let $U_2 = U_0 + U_1$; then U_2 is a finitely axiomatizable subtheory of U as well. Since $U_1 \subset U_2$, we have $U_2 \vdash A^\tau \leftrightarrow A$ for all K-sentences A. We use this to get

$$U \vdash A \Rightarrow U_0 \vdash A^\tau \Rightarrow U_2 \vdash A^\tau \Rightarrow U_2 \vdash A,$$

for all K-sentences A. But this is to say that $U_2 \vdash U$, so U must be finitely axiomatizable. □

Corollary 2.3.4. *Suppose that every self-interpretation of `PrA` is provably trivial. Then for all integers $n \geq 1$, `PrA` is not interpretable in `PrA`$_n$.*

[13] For the reader who wishes to attempt it: when defining $x + y$, distinguish cases modulo 4.

Proof. Suppose the contrary. Applying theorem 2.3.3 to $\mathtt{PrA}$ and its finitely axiomatized subtheory $\mathtt{PrA}_n$ gives that $\mathtt{PrA}$ itself is finitely axiomatizable, which contradicts theorem 1.2.4. □

This result provides the motivation for an investigation into self-interpretations of $\mathtt{PrA}$, which we will conduct in chapter 3.

2.4 Approximating $\mathtt{PrA}$ in $\mathtt{PrA}^-$

Before we set out to attack conjecture 2.3.2, we will investigate what we *can* do, should it turn out that $\mathtt{PrA}$ isn't interpretable in $\mathtt{PrA}^-$. Let us consider the interpretations given in the proof of theorem 2.2.1; we will use the notation introduced there. These interpretations turn out to have some nice properties. We can prove that, for integers $m, n \geq 1$, the domain formula of $\iota_n \circ \iota_m$, i.e. $\mathcal{D}_n \wedge \mathcal{D}_m$, is equivalent in $\mathtt{PrA}^-$ to $\mathcal{D}_{mn}$.[14] The proof is quite similar to the proof of proposition 1.1.5, so we omit it.

What does this mean? Suppose we have some model M of $\mathtt{PrA}^-$. If we take the inner model given by ι_n, and take *in this model* the inner model given by ι_m, we end up with the same structure we would have ended up with if we had just taken the inner model in M given by ι_{mn}. So in particular, this inner inner model still satisfies $\mathtt{PrA}_n$, and not just $\mathtt{PrA}_m$. This means we can interpret the division axioms one by one, and obtain a sequence of models with decreasing domains, which approximate the whole of $\mathtt{PrA}$ better as we get up the sequence.[15] Let us make this precise by the following definition.

Definition 2.4.1. Let M be a model of $\mathtt{PrA}^-$. We define the sequence $(\mathcal{J}_n(M))_{n\in\mathbb{N}}$ of $\mathcal{L}^-$-substructures of M by

- $\mathcal{J}_0(M) = M$;
- $\mathcal{J}_{n+1}(M)$ is the inner model in $\mathcal{J}_n(M)$ given by ι_{p_n}, where p_n is the n^{th} prime number. Here $p_0 = 2$.

We set $\mathcal{J}(M) = \bigcap_{n\in\mathbb{N}} \mathcal{J}_n(M)$.

All the $\mathcal{J}_n(M)$ are $\mathcal{L}^-$-substructures of M because all predicates are translated in the trivial manner by the ι's. They are even substructures with respect to $<$ as well, because their domains are downwards closed under $\leq$. Consequently, $\mathcal{J}(M)$ is a $(\mathcal{L}^- + \{<\})$-substructure of M as well. We only interpret for the division axioms for the prime numbers because corollary 1.2.3 tells us that this suffices. By the above remarks, we have $\mathcal{J}(M) \models \mathtt{PrA}_{P_n}$, where $P_n = \prod_{i=0}^{n-1} p_n$. One can now quite easily prove that $\mathcal{J}(M)$ is a model of the whole of $\mathtt{PrA}$. However, and fortunately for our conjectures, this procedure does *not* necessarily yield an *interpretation* of $\mathtt{PrA}$ in $\mathtt{PrA}^-$. Indeed, taking an infinite intersection prevents the domain of $\mathcal{J}(M)$ from being definable by a first order sentence.

[14] The notation $\iota_n \circ \iota_m$ may seem a bit odd, because the interpreting theory of ι_m, i.e. $\mathtt{PrA}^-$, is not equal to the interpreted theory of ι_n, i.e. $\mathtt{PrA}_n$. But can naturally view ι_m as an interpretation of $\mathtt{PrA}_m$ in $\mathtt{PrA}_n$ as well, and when we write $\iota_n \circ \iota_m$, it is understood that we do so.

[15] The equivalence of domains we described above also has a consequence for the category of interpretations from footnote 8, namely that $\iota_n \circ \iota_m = \iota_{mn} = \iota_m \circ \iota_n$. We can express this as a commutative diagram. As a result, the 'route' we choose to approximate $\mathtt{PrA}$ doesn't really matter.

We may wonder whether this approximating procedure always gives us $\texttt{PrA}$ after a finite number of steps, no matter what M we start with. In other words, does there always exist an $n \in \mathbb{N}$ such that $\mathcal{J}_n(M) \models \texttt{PrA}$ (and consequently $\mathcal{J}_k(M) = \mathcal{J}(M)$ for every $k \geq n$)? It is obvious that for a lot of models, this is the case. Indeed, let M be any model of $\texttt{PrA}$, then we have $\mathcal{J}_n(M) = M$ for all $n \in \mathbb{N}$ and in particular $\mathcal{J}(M) = M = \mathcal{J}_0(M)$. Also note that an affirmative answer to our question would again follow if $\texttt{PrA}$ were finitely axiomatizable, because in that case we would have $\texttt{PrA} = \texttt{PrA}_N$ for some integer $N \geq 1$. However, since $\texttt{PrA}$ can*not* be finitely axiomatized, we conjecture the answer to be no. That is, we do not expect that our procedure *in general* only takes a finite number of steps. This will be another strengthening of theorem 1.2.4.

We will now have to look for a countermodel that shows the answer to our question to indeed be no. It seems plausible that the models $\mathcal{S}_n$ from definition 1.2.2, which are specifically constructed for some principles $\texttt{PrA}_m$ to hold and some others not, are candidates. However, the fact that the domain in our interpretations ι_n are downwards closed under $\leq$ makes matters more complicated. Consider, for example, the model $\mathcal{S}_1$ and suppose we wish to execute ι_2. In $\mathcal{S}_1$, there is a smallest nonstandard element, namely X. It is not hard to see that X is neither even nor odd. But now the domain of $\mathcal{J}_1(M_1)$ cannot contain any nonstandard element, since such an element must at least be X, for which $\texttt{PrAx8}_2$ doesn't hold. So $\mathcal{J}_1(\mathcal{S}_1)$ contains only standard elements of $\mathcal{S}_1$ and, being a model of $\texttt{PrA}^-$, it should contain all of them. That is, $\mathcal{J}_1(\mathcal{S}_1) \cong \mathbb{N}$, so our sequence immediately collapses.

In $\mathcal{S}_2$, matters aren't any better. When we execute ι_2 our sequence does not collapse, but this is for the rather trivial reason that $\texttt{PrAx8}_2$ already held in $\mathcal{S}_2$. When we go to the next step, our sequence does collapse, again onto $\mathbb{N}$. Although $\mathcal{S}_2$ does not have a smallest nonstandard element, there are arbitrarily small nonstandard elements for which we can guarantee that they are not divisible by 3 with remainder. Indeed, consider elements of the form $\frac{1}{2^k}X$ with $k \in \mathbb{N}$. So $\mathcal{J}_0(\mathcal{S}_2) = \mathcal{J}_1(\mathcal{S}_2) = \mathcal{S}_2$, whereas $\mathcal{J}_n(\mathcal{S}_2) \cong \mathbb{N}$ for $n \geq 2$. In general, we can show: as soon as we try to carry out ι_p in $\mathcal{S}_n$ for some prime p not dividing n, our sequence collapses onto $\mathbb{N}$.

These considerations show that we have to be more clever to find a countermodel. Thus far, we haven't even found a model of $\texttt{PrA}^-$ in which $\texttt{PrAx8}_2$ doesn't already hold and in which our sequence doesn't immediately collapse when we execute ι_2. So let us first try to find such a model. When carrying out ι_2, we want to end up with more than just $\mathbb{N}$. That is, we want to end up with some *nonstandard* model of $\texttt{PrA}_2$. However, we cannot *start* with such a model. These observations lead to the following idea: let us take some nonstandard model of $\texttt{PrA}_2$ and paste another piece at the end of it (more precisely, construct an end-extension) in which division by 2 with remainder is *not* always possible.

Let $\mathcal{R}_1$ denote the set of all $a + bX + cX^2 \in \mathbb{Q}[X]$ such that a and c are in $\mathbb{Z}$, the denominator of b is a power of 2, and the leading coefficient is nonexistent or positive. Note that the last constraint just means that we consider X to be infinitely large. One easily checks that this is a model of $\texttt{PrA}^-$, with 0, 1 and + induced by $\mathbb{Q}[X]$. What happens if we carry out ι_2 in this $\mathcal{R}_1$? First of all, consider an element $a + bX$ without quadratic coefficient. We can divide this element by 2. Indeed, $a + bX = 2\left(q + \frac{b}{2}X\right) + r$ for some standard q and r, namely the quotient resp. remainder of a upon division by 2. We conclude that $\mathcal{J}_1(\mathcal{R}_1)$ contains all elements without quadratic coefficient. To see which elements *cannot* be divided by two, we use an argument similar to the one we used for $\mathcal{S}_1$: there is a smallest element with nonzero quadratic coefficient, namely X^2, and one easily checks that division by 2 is impossible. Summarizing, $\mathcal{J}_1(\mathcal{R}_1)$ contains exactly the elements with $c = 0$.

We take matters a step further. We now want a model in which the executions of both ι_2 and ι_3 are nontrivial and do not cause our sequence to collapse. But it is quite obvious how to accomplish this. Let $\mathcal{R}_2$ be the set of all $a + bX + cX^2 + dX^3$ such that a and d are in $\mathbb{Z}$, the denominator of c is a power of 2, the denominator of b contains only prime factors 2 and 3,[16] and it is nonnegative when we consider X to be infinitely large. We can prove: $\mathcal{J}_1(\mathcal{R}_2)$ contains exactly the elements without cubic coefficient, and $\mathcal{J}_2(\mathcal{R}_2)$ loses the elements with nonzero quadratic coefficient as well. The ideas for the proof have been encountered before, so we will not give it here.

It is possible generalize these examples in an infinitary way and obtain the desired countermodel.[17] We now have to leave the world of polynomials, since they would become infinitely large. It is, however, instructive to see what this our model amounts to in terms of infinite "polynomials". What we want to construct is $\mathbb{N}$, followed by segments in which we can divide by p_i, but not by p_{i+1}, but in reverse order. That is, i runs $\ldots, 2, 1, 0$. This gives a certain ungroundedness that makes it impossible to work with a single variable any longer. We therefore introduce a countable set of variables $X_0, X_1, \ldots$ and we consider linear, possibly infinite, "polynomials" P in these variables, that satisfy the following constraints:

- For all $i \geq 0$, X_i in infinitely large and X_{i+1} is infinitely larger than X_i;
- P is nonnegative in the above sense;
- The prime divisors of the denominator of the coefficient of X_i are among $p_0, \ldots, p_{i-1}$;
- The constant coefficient is an integer.

We now introduce the model in terms that do not use infinite polynomials, but rather functions from a certain ordinal to the set of rational numbers.

Definition 2.4.2. Let $\mathcal{R}_\infty$ be the set of functions $f : \omega + 1 \to \mathbb{Q}$ such that

- If f is not the zero function and α is the smallest element of $\omega + 1$ such that $f(\alpha)$ is nonzero, then $f(\alpha) > 0$;
- For $i \in \omega$, the denominator of $f(i)$ is not divisible by p_j for all $j \geq i$;
- $f(\omega) \in \mathbb{Z}$.

We make this set into a $\mathcal{L}^-$-structure by interpreting 0 as the zero function, 1 as the function given by $f(i) = 0$ for all $i \in \omega$ and $f(\omega) = 1$, and $+$ as addition of functions.

One now has to do a lot of checking: that this is indeed a $\mathcal{L}^-$-structure; that it is a model of `PrA`; and what happens when we start interpreting the division axioms. There are no new ideas involved, and we give the answer to the last question at once. For $n \in \mathbb{N}$, we have $\mathcal{J}_n(\mathcal{R}_\infty) = \{f \in \mathcal{R}_\infty \mid f(0) = \ldots = f(n-1) = 0\}$. Letting n tend to infinity, we see that $\mathcal{J}(\mathcal{R}_\infty) = \{f \in \mathcal{R}_\infty \mid \forall i \in \omega\ f(i) = 0\} \cong \mathbb{N}$. We have our desired model.

[16] An equivalent formulation is: some denominator of b is a power of 6, a constraint we have encountered before. Since we are primarily working with prime numbers now, the now chosen formulation is more convenient.

[17] We can also generalize the above examples in a finite manner, and use a compactness argument to prove the existence of a desired counterexample. However, our procedure gives such a counterexample *explicitly*, which is more informational.

This model does not only show that the conjecture we made at the beginning of this section is indeed true, it also provides an insight into the nature of the division axioms $\texttt{PrAx8}_p$, for p prime. We already expressed a feeling that these are rather independent, and $\mathcal{R}_\infty$ confirms this. For every prime p, there is a region in $\mathcal{R}_\infty$ where we can divide by p but by no means by everything.

3 Self-interpretations of `PrA`

Motivated by conjecture 2.3.2, we dedicate this chapter to studying self-interpretations of `PrA`. The first sections present results about $\mathcal{L}$-definability, inner models in $\mathbb{N}$, definable order types and definable functions respectively. In the final section, we use these ideas to give two possible proof strategies for conjecture 2.3.2.

3.1 Definable predicates in $\mathbb{N}$

In this section, we will discuss a rather important preliminary result about unary predicates that are $\mathcal{L}$-definable in the standard model, and several of its implications. This result will provide us with a fundamental insight in the nature of definability in the standard model.

Theorem 3.1.1. *Let $\phi(x)$ be some $\mathcal{L}$-formula in one free variable. Then the set A that ϕ defines in the standard model is eventually periodic. That is, there exist $M, N \in \mathbb{N}$ such that for all $x \geq N$, we have $x \in A$ iff $x + M \in A$. Furthermore, if ϕ is quantifier free, we can choose $M = \operatorname{lcm}(m_1, \cdots, m_r)$, where the m_i are the moduli of all congruence symbols occurring in ϕ.*

Proof. We first consider the quantifier free case and proceed by induction on the $\mathcal{L}$-complexity of ϕ.
Every closed term must denote some standard number, so if ϕ is atomic, proposition 1.3.2 tells us we may assume ϕ to be of one of the following forms:

$$\underline{a}x + \underline{b} = \underline{c}; \tag{4}$$
$$\underline{a}x + \underline{b} < \underline{c}; \tag{5}$$
$$\underline{a}x + \underline{b} > \underline{c}; \tag{6}$$
$$\underline{a}x + \underline{b} \equiv_n \underline{c} \text{ for some } n \geq 1, \tag{7}$$

where $a, b, c \in \mathbb{N}$. If $a = 0$, then the above formulas define either $\emptyset$ or $\mathbb{N}$, and both are purely periodic with period 1. So suppose that $a > 0$. If ϕ is of form (4), then A has at most one element; if ϕ is of the form (5), then it defines a finite set; and if ϕ is of the form (6), then it defines a cofinite set. In all cases, A is eventually periodic with period 1. Finally, if ϕ is of the form (7), then A is purely periodic with period n.
Suppose ϕ is $\neg\psi$ for some ψ with corresponding set B that is eventually periodic with period M. Now for some $N \in \mathbb{N}$, we have: if $x \geq N$, then

$$x \in A \Leftrightarrow x \notin B \Leftrightarrow x + M \notin B \Leftrightarrow x + M \in A.$$

So A is also eventually periodic with period M. We can pick M to be the least common multiple of all moduli occurring in ψ; exactly the same moduli occur in ϕ.
Finally, suppose ϕ is $\psi_0 \wedge \psi_1$ for some ψ_0 and ψ_1 with corresponding sets B_0 and B_1 that are eventually periodic with periods M_0 and M_1 respectively. Furthermore, suppose the periodicity holds for $x \geq N_0$ and $x \geq N_1$ respectively. Now define $N = \max(N_0, N_1)$ and $M = \operatorname{lcm}(M_0, M_1)$. Since M is a multiple of both M_0 and M_1, we have for $x \geq N$:

$$x \in A \Leftrightarrow x \in B_0 \text{ and } x \in B_1 \Leftrightarrow x + M \in B_0 \text{ and } x + M \in B_1 \Leftrightarrow x + M \in A.$$

So A is eventually periodic with period M. We can pick M_0 and M_1 to be the least common multiple of all moduli occurring in ψ_0 resp. ψ_1, so that M is the least common mulitple of all moduli occurring in both ψ_0 and ψ_1, that is, in ϕ.

For the general case: the theory of the standard model in the language $\mathcal{L}$ admits quantifier elimination, so this reduces to the quantifier free case. □

Conversely, every eventually periodic subset of $\mathbb{N}$ is definable in the standard model, as one can easily show. Although $\mathcal{L}$-definable unary predicates behave rather well in the standard model, things are not so tidy for more-place predicates. A general result can be found in [3], but we will not be able to use those results here. We *will* give some corollaries of the preceding theorem concerning $\mathcal{L}$-definable *binary* predicates, one of which concerns $\mathcal{L}$-definable functions. A full attack on definable functions will be placed in section 3.4, where we show that they have a decent normal form.

Corollary 3.1.2. *Let $\phi(x, y)$ be a $\mathcal{L}$-formula in two free variables. For $n \in \mathbb{N}$, let A_n be the set $\{y \in \mathbb{N} \mid \mathbb{N} \models \phi(n, y)\}$. Then there is an $M \in \mathbb{N}$ such that: there are no $n_0, \ldots, n_M \in \mathbb{N}$ such that all A_{n_i} are infinite and pairwise disjoint.*

Proof. Let M be the least common multiple of all moduli occurring in a quantifier free version of ϕ and consider $n_0, \ldots, n_M \in \mathbb{N}$. Now for $0 \leq i \leq M$, we can define A_{n_i} by $\phi\left(\underline{n_i}, y\right)$, so by theorem 3.1.1, the A_{n_i} are eventually periodic with period M. Let $N \in \mathbb{N}$ be sufficiently large such that all A_{n_i} are periodic for $x \geq N$.
Consider the set $V = \{N, N+1, \ldots, N+M-1\}$. If all A_{n_i} are infinite, then they must all contain at least one element from V. Indeed, if some A_{n_i} does not contain an element of V, then by the periodicity of A_{n_i}, it can contain only elements smaller than N, and therefore it is finite. But now by the pigeon hole principle, the A_{n_i} cannot be disjoint. □

Corollary 3.1.3. *With the notation as in corollary 3.1.2: if all A_n are pairwise disjoint, then there are only finitely many n such that A_n is infinite.*

Proof. Immediate. □

Corollary 3.1.4. *Let $F(x, y)$ be an $\mathcal{L}$-formula defining a function f in the standard model. That is, we have $\mathbb{N} \models \forall x \exists! y\, F(x, y)$ and we denote this unique y by $f(x)$. If f has finite range, then it must be eventually periodic, i.e. there exist $M, N \in \mathbb{N}$ such that $f(x) = f(x + M)$ for all $x \geq N$.*

Proof. All values of f must lie in the set $\{0, 1, \ldots, S\}$ for some $S \in \mathbb{N}$. For $0 \leq i \leq S$, we can define the pre-image $f^{-1}(i)$ in $\mathcal{L}$ as $F(x, \underline{i})$. So by theorem 3.1.1, all these pre-images must be eventually periodic; let us say that $f^{-1}(i)$ has period M_i beyond a certain N_i. Now we can take $M = \operatorname{lcm}\{M_i \mid 0 \leq i \leq S\}$ and $N = \max\{N_i \mid 0 \leq i \leq S\}$. □

3.2 Interpretations in $\mathbb{N}$

When considering interpretation from some theory to PrA, it is quite interesting to study the behaviour of such an interpretations in $\mathbb{N}$. First of all, $\mathbb{N}$ is an object we already know a lot about, and which we can embed in larger number systems, such as $\mathbb{Z}$ and $\mathbb{Q}$. Secondly, it isn't really a restriction.

Lemma 3.2.1. *Suppose U is some theory and $\iota, \kappa : U \to$ PrA are interpretations. Then ι and κ are provably isomorphic if and only if they are representably isomorphic in $\mathbb{N}$.*

Proof. The lemma claims that there is an $\mathcal{L}^-$-formula $F(x, y)$ such that PrA $\vdash F : \iota \cong \kappa$ if and only if there is an $\mathcal{L}^-$-formula $F(x, y)$ such that $\mathbb{N} \models F : \iota \cong \kappa$. But this is obvious, since PrA is the true theory of $\mathbb{N}$. □

So if we want to prove our conjecture that every self-interpretation of `PrA` is provably trivial, then it suffices to prove that every such interpretation is representably trivial in the standard model. It is for this reason that we now dedicate a section to investigating the general behaviour of interpretations from some theory to `PrA` in the standard model. In particular, we will show that we can get rid of the subtleties concerning identity and the domain formula.

As a first step, we use the least number principle in $\mathbb{N}$ to show that identity may always be translated to identity on the appropriate domain.

Lemma 3.2.2. *Suppose U is some theory in the language K and $\iota : U \to \texttt{PrA}$ is some interpretation based on $\tau : K \to \mathcal{L}^-$. Then there is an interpretation $\kappa : U \to \texttt{PrA}$ based on a $\sigma : K \to \mathcal{L}^-$ such that*

- *$\mathcal{F}_\sigma\left(=_U\right)(x, y)$ is equivalent in $\mathbb{N}$ to $\mathcal{D}_\sigma(x) \wedge \mathcal{D}_\sigma(y) \wedge x = y$;*
- *ι and κ are representably isomorphic in $\mathbb{N}$.*

Proof. For $\mathcal{D}_\sigma(x)$, we take

$$\mathcal{D}_\tau(x) \wedge \Big[\forall y \ \left(\left(\mathcal{D}_\tau(y) \wedge \mathcal{F}_\tau\left(=_U\right)(x, y)\right) \to x \leq y\right)\Big].$$

For a predicate symbol P in U, we take $\mathcal{F}_\sigma(P)$ to be

$$\bigwedge_{i=0}^{n-1} \mathcal{D}_\sigma(x_i) \ \wedge \mathcal{F}_\tau(P)(x_0, \ldots, x_{n-1}).$$

In other words, we take as domain the *smallest* representatives from the $\mathcal{F}\left(=_U\right)$-equivalence classes, and let the predicates be induced from τ. Note that such a smallest representative always exists in $\mathbb{N}$. Since each $\mathcal{F}_\tau\left(=_U\right)$-equivalence class has at most one representative in $\mathcal{D}_\sigma$, we have $(\mathcal{D}_\sigma(x) \wedge \mathcal{D}_\sigma(x)) \to (\mathcal{F}_\tau\left(=_U\right)(x, y) \leftrightarrow x = y)$ in $\mathbb{N}$, so the first constraint is satisfied. It is very easy to see that κ is an interpretation: because $\mathcal{F}_\tau\left(=_U\right)$ respects all the $\mathcal{F}_\tau(P)$, really nothing has changed about the inner model apart from the fact that the $\mathcal{F}_\sigma\left(=_U\right)$-equivalence classes are all singletons now. This insight also shows us that the isomorphism $F(x, y)$ from ι to κ should be

$$\mathcal{D}_\tau(x) \wedge \mathcal{D}_\sigma(y) \wedge \mathcal{F}_\tau\left(=_U\right)(x, y),$$

which sends every element of $\mathcal{D}_\tau$ to the unique representative of its $\mathcal{F}_\tau\left(=_U\right)$-equivalence class in $\mathcal{D}_\sigma$. □

The advantage of having the above κ is that, when one considers some K-definable function f, the translation of f behaves well. More precisely, we do not only have functionality in the $\mathcal{F}_\sigma\left(=_U\right)$-sense, but even in the $=$-sense. That is, for every x there is a (really) unique y such that y is the f of x, and not only a $\mathcal{F}_\sigma\left(=_U\right)$-unique y. Next, we eliminate concerns about the domain formula.

Lemma 3.2.3. *Suppose U is some theory in the language K and $\iota : U \to \texttt{PrA}$ is some interpretation based on $\tau : K \to \mathcal{L}^-$. Then there is an interpretation $\kappa : U \to \texttt{PrA}$ based on a $\sigma : K \to \mathcal{L}^-$ such that*

- *$\mathcal{D}_\sigma$ is some tautology.*
- *$\mathcal{F}_\sigma\left(=_U\right)(x, y)$ is equivalent in $\mathbb{N}$ to $x = y$;*
- *ι and κ are representably isomorphic in $\mathbb{N}$.*

Proof. By lemma 3.2.2, we may assume without loss of generality that $\mathcal{F}_\tau\left(=_U\right)(x,y)$ is equivalent in $\mathbb{N}$ to $\mathcal{D}_\tau(x) \wedge \mathcal{D}_\tau(y) \wedge x = y$.
We work in $\mathbb{N}$. Our goal is to define a bijection between the domain $\mathcal{D}_\tau$ and $\mathbb{N}$. By theorem 3.1.1, the domain $\mathcal{D}_\tau$ is eventually periodic, so there exist $M, N \in \mathbb{N}$ such that for $x \geq N$, we have $x \in \mathcal{D}_\tau$ if and only if $x + M \in \mathcal{D}_\tau$. Suppose that

$$\mathcal{D}\mathcal{T} \cap \{0, 1, \ldots, N-1\} = \{b_0, b_1, \ldots, b_{r-1}\},$$

where $b_0 < b_1 < \cdots < b_{r-1}$, and that

$$\mathcal{D}\mathcal{T} \cap \{N, N+1, \ldots, N+M-1\} = \{N+k_0, N+k_1, \ldots, N+k_{s-1}\},$$

where $k_0 < k_1 < \cdots < k_{s-1}$. Possibly, some of r and s are zero. Note that beyond N, precisely the elements from some equivalence class $N + k_j$ modulo M belong to $\mathcal{D}_\tau$. Now we can build our bijection by sending b_i to i and beyond N, by sending the equivalence class $N + k_j$ modulo M to the equivalence class $r + j$ modulo s. It is not hard to check that this indeed gives a bijection, and that it is $\mathcal{L}^-$-definable as

$$\bigvee_{i=0}^{r-1} \left(x = \underline{b_i} \wedge y = \underline{i}\right) \vee \bigvee_{j=0}^{s-1} \exists u \left(x = \underline{N} + \underline{M}u + \underline{k_j} \wedge y = \underline{r} + \underline{s}u + \underline{j}\right).$$

Call this formula $F(x,y)$ and take $\mathcal{D}_\sigma$ to be some tautology. Now we let the translations of the predicate symbols P in U be induced by τ and this bijection, i.e. $\mathcal{F}_\sigma(P)$ is

$$\exists y_0 \exists y_1 \cdots \exists y_{n-1} \left[\bigwedge_{i=0}^{n-1} F(y_i, x_i) \ \wedge \mathcal{F}_\tau(y_0, \ldots, y_{n-1})\right].$$

Because of the way we constructed σ, it is immediately clear that κ is an interpretation as well, and that ι are κ are representably isomorphic in $\mathbb{N}$. Indeed, the isomorphism is F itself. To finish the proof, we show that the second constraint is satisfied. By definition, we have $\mathcal{F}_\sigma\left(=_U\right)(x,y)$ if and only if there are $u, v \in \mathcal{D}_\tau$ such that $F(u,x)$, $F(v,y)$ and $u = v$. But since F represents a bijection from $\mathcal{D}_\tau$ to $\mathbb{N}$, such u and v exist if and only if $x = y$. □

The above lemma shows that, when we are concerned with the behaviour in $\mathbb{N}$ of interpretations from some theory U to PrA, we may assume without loss of generality that the domain is trivial and that U-identity is translated to identity in $\mathbb{N}$.

3.3 Definable order types in $\mathbb{N}$

Although addition possesses a much simpler structure than multiplication, its simplicity does not stretch far enough. Indeed, we can view addition as a *ternary* predicate, and we haven't gathered results about such predicates yet. This is why we focus on the ordering in this section. The ordering still gives an interesting structure, but is easier to handle because it is a binary predicate.

What can we say about the ordering, when we consider self-interpretations of PrA carried out in the standard model? The inner models we get are countable, so theorem 1.2.1 tells us the ordering must be the standard one, or isomorphic to $\mathbb{N} + \mathbb{Z} \cdot \mathbb{Q}$. The goal of this section is to show that the expressive power of $\mathcal{L}$ isn't enough to define the latter in $\mathbb{N}$.

Theorem 3.3.1. *Let* `LO` *be the theory of linear orders written in the language $\langle < \rangle$ and suppose we have an interpretation $\iota : \texttt{LO} \to \texttt{PrA}$. Then the ordering in the inner model in $\mathbb{N}$ given by ι cannot be isomorphic to $\mathbb{N} + \mathbb{Z} \cdot \mathbb{Q}$.*

Proof. Let τ be the translation that ι is based on. By theorem 3.2.3, we may assume without loss of generality that the $\mathcal{D}_\tau$ is trivial and that $\mathcal{F}_\tau(=)(x,y)$ is equivalent in $\mathbb{N}$ to $x = y$. For the sake of readability, we write $x <_* y$ for $\mathcal{F}_\tau(<)(x,y)$.
We suppose the contrary of the theorem. Let $X \subset \mathbb{N}$ be the set defined by the $\mathcal{L}$-formula

$$\forall y \ (y < x \to y <_* x). \tag{8}$$

Informally, an $x \in \mathbb{N}$ is in X precisely if it is $<_*$-larger than all the natural numbers smaller than x *in the usual ordering.* We now make the following crucial observation.

Lemma 3.3.2. *The set X is $<_*$-cofinal. That is, there is no $b \in \mathbb{N}$ such that $x <_* b$ for all $x \in X$. In particular, X is infinite.*

Proof. Suppose there is such a b. There exist $n \in \mathbb{N}$ such that $n \geq_* b$.[18] Indeed, b is such an n. Now let x the $<$-smallest such n. Note that this x must exist, since $<$ is a well-order. Now, by the minimality of x, for every $y < x$, we have $y <_* b$. We also had $b \leq_* x$, so by transitivity, $y <_* x$. But this means that x is in X, contradiction with our assumption.
Now if X were finite, then we could pick some $b \in \mathbb{N}$ that is $<_*$-larger than all $x \in X$. Such a b must exist, because $\mathbb{N} + \mathbb{Z} \cdot \mathbb{Q}$ does not allow a largest element. □

Let $\phi(x,y)$ be the formula

$$x \in X \wedge \exists z \ [z \in X \wedge z > x \wedge \forall w \ ((w \in X \wedge w > x) \to w \geq z) \wedge (x <_* y <_* z)]. \tag{9}$$

Here $x \in X$ is of course shorthand for the formula (8), with a suitable different choice for the bound variable y. Furthermore, $x <_* y <_* z$ is shorthand for $x <_* y \wedge y <_* z$. So (9) is an $\mathcal{L}$-formula. Informally, (9) says that y is $<_*$-between x and the $<$-successor of x *in the set X*. We can speak of *the* $<$-successor of x in X because the set X with the induced order from $\mathbb{N}$ is isomorphic to $\langle \mathbb{N}, < \rangle$, in which successors always uniquely exist.
We now use the notations from corollary 3.1.2 for ϕ given as in (9), and for $x \in X$, we denote the $<$-successor of x in the set X by $s^X(x)$. We first claim that all A_n are pairwise disjoint. So consider different $m, n \in \mathbb{N}$. If $m \notin X$, then A_m is empty, so A_m and A_n are certainly disjoint. We handle the case $n \notin X$ similarly. So we can assume $m, n \in X$; w.l.o.g. $m < n$. Suppose there is a $y \in (A_m \cap A_n)$. That is, $m <_* y <_* s^X(m)$ and $n <_* y <_* s^X(n)$. Since $m < n$, we can quite easily see that $s^X(m) \leq n$. Now, because $n \in X$, we also have $s^X(m) \leq_* n$. But now we get $y <_* s^X(m) \leq_* n <_* y$, contradiction. So $A_m \cap A_n = \emptyset$.
We can now apply corollary 3.1.3: only a finite number of A_n is infinite. So there is an $N \in \mathbb{N}$ such that for all $n \geq N$, we have $|A_n| < \infty$. We call $a, b \in \mathbb{N}$ *segment-equivalent* if we can get from a to b or from b to a by applying the $<_*$-successor function a finite number (possibly zero) of times. It is rather well-known that this is an equivalence relation. Informally, a and b are segment-equivalent iff they are both in the standard part $\mathbb{N}$, or in the same copy of $\mathbb{Z}$. Note that, if a and b are *not* segment-equivalent, then there are infinitely many elements $<_*$-between them.
Since X is infinite, there is some $x \in X$ such that $x \geq N$. Let v the $<$-smallest such x. We now prove by induction *on the order in X* that every $x \in X$ such that $x \geq N$ is segment-equivalent to v.

[18] Of course, this notation means $b <_* n \vee b = n$, or equivalently, $\neg(n <_* b)$. We will use similar abbreviations as well.

Basis. The smallest $x \in X$ such that $x \geq N$ is v itself, so this is obvious.
Step. Consider an $x \in X$ such that $x \geq N$ and x is segment-equivalent to v. Now if $s^X(x)$ were *not* segment-equivalent to x, then there would be an infinite number of elements $<_*$-between x and $s^X(x)$. This means exactly that A_x is infinite. But $x \geq N$, so this cannot be the case. We conclude that $s^X(x)$ *is* segment-equivalent to x, and therefore to v. This completes the induction.
Now for any $x \in X$ such that $x < N$, we have $x < v$, and therefore also $x <_* v$, because $v \in X$. So every $x \in X$ that is $<_*$-larger than v is still segment-equivalent to v. But now X can be $<_*$-bounded from above by some $b \in \mathbb{N}$ that is larger than, but not segment-equivalent to, v. Such a b must exist, since the ordering of the segments in $\mathbb{N} + \mathbb{Z} \cdot \mathbb{Q}$ has order type $1 + \mathbb{Q}$, which does now allow a largest element. We have arrived at a contradiction with lemma 3.3.2. □

The above proof works so well because the set X allows us to create an interplay between the two orderings $<$ and $<_*$. To end this section, let us indicate the relevance of the above result for our project and in a more general setting. Theorem 3.3.1 provides us with an important insight. Of course, since $\texttt{PrA} \vdash \texttt{LO}$, a self-interpretation of $\texttt{PrA}$ also yields an interpretation $\texttt{LO} \to \texttt{PrA}$. So the inner model in $\mathbb{N}$ given by a self-interpretation of $\texttt{PrA}$ cannot have order-type $\mathbb{N} + \mathbb{Z} \cdot \mathbb{Q}$ and therefore it must be isomorphic to the standard model. So not only are we working in a familiar object, namely $\mathbb{N}$, but the inner model is also isomorphic to this familiar object. This also means that this inner model is completely determined by its successor function, so it might be feasible to attack conjecture 2.3.2 by studying this function.

In a more general setting, we can use a similar proof to show that order types like $\mathbb{Q}$ and ω^2 are not definable in $\mathbb{N}$. Let us consider the last result a bit more closely. We say that two theories U and V are *bi-interpretable* if there are interpretations $\iota : U \to V$ and $\kappa : V \to U$ such that $\kappa \circ \iota$ and $\iota \circ \kappa$ are provably isomorphic to id_U and id_V respectively. Define T as the true theory of the structure $\langle \omega^2; 0, 1, +, f \rangle$, where f is a function such that $f(\omega \cdot a + b) = a$. Now we can give an interpretation ι of $\texttt{PrA}$ in T by taking as domain all x such that $f(x) = 0$ and by translating 0, 1 and + trivially. There is, however, no interpretation the other way around, since that would cause the order type ω^2 to be definable $\mathbb{N}$, which is not the case. It turns out that we *can* give a *two-dimensional* interpretation κ of T in $\texttt{PrA}$ such that the two compositions of ι and κ are provably isomorphic. So $\texttt{PrA}$ and T are bi-interpretable if we may use more-dimensional interpretations, but not if we are restricted to one-dimensional ones.

3.4 A normal form for definable functions

As indicated near the end of the previous section, we may approach conjecture 2.3.2 by considering a certain definable function. In order to do this, we first need to know something about $\mathcal{L}$-definable functions. Let us consider some intuitions concerning definability in $\texttt{PrA}$ we have developed so far. First of all, the examples in section 2.3 we can distinguish cases modulo a certain fixed number. Secondly, as theorem 3.1.1 shows, $\texttt{PrA}$ may make a mess for small numbers, but if we make our numbers sufficiently large, things behave nicely. And finally, the predicate symbol in $\mathcal{L}$ we typically expect to behave functionally, is identity. All these intuitions are expressed in the following result. Basically, it says that for x sufficiently large, we can express $F(x, y)$ as a normal form involving a case-distinction according to x's value modulo a certain number and in each separate case, an equality.

Theorem 3.4.1. *Let $F(x,y)$ be an $\mathcal{L}$-formula defining a function f in the standard model. Then there exist $M, N \in \mathbb{N}$ and for $0 \le i < M$, coefficients $a_i, b_i, c_i, d_i \in \mathbb{N}$ such that*

$$x \geq \underline{N} \to \left[F(x,y) \leftrightarrow \bigvee_{i=0}^{M-1} \left(x \equiv_M \underline{i} \wedge \underline{a_i} y + \underline{b_i} = \underline{c_i} x + \underline{d_i} \right) \right]$$

holds in $\mathbb{N}$, and $a_i > 0$ for $0 \le i < M$.

Proof. Since the proof is somewhat involved, we will show how to find the appropriate M, N, a_i, b_i, c_i and d_i by transforming the formula F in a number of stages.
Stage 1. Apply quantifier elimination to obtain a quantifier free equivalent $F_0(x,y)$ of $F(x,y)$. Since F_0 is nothing but a truth-functional combination of atomic $\mathcal{L}$-formulas, we may write it in disjunctive normal form. Next, we eliminate negation the same way as in the quantifier elimination from section 1.4, using the equivalences

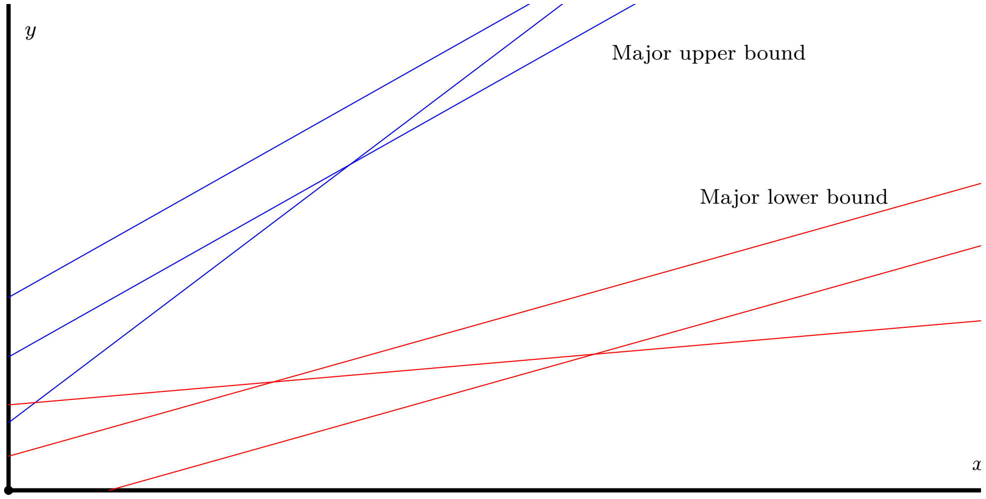
$$\begin{aligned} t \neq s &\leftrightarrow t < s \vee s < t; \\ \neg(t < s) &\leftrightarrow t = s \vee s < t; \\ \neg(t \equiv_n s) &\leftrightarrow (t \equiv_n s + \underline{1}) \vee \ldots \vee (t \equiv_n s + \underline{n-1}). \end{aligned}$$

and distributing the $\wedge$'s over the $\vee's$. We obtain a new disjunctive normal form.

Stage 2. This is the most involved part of the proof. At all points where we say that we can pick x sufficiently large for a certain purpose, we assume x to be at least this large for the rest of this stage. To ensure that the procedure is easy to follow, we have indicated these points by boldfaced text.
The previous step showed F to be equivalent to a formula F_1 that is a disjunction of conjunctions of atomic $\mathcal{L}$-formulas. We claim that, for x sufficiently large, F is equivalent to such a form *where each conjunction contains a nontrivial equality.* By the nontriviality we mean that the equality doesn't hold for all $(x,y) \in \mathbb{N}^2$, or for no $(x,y) \in \mathbb{N}^2$. So consider one of the conjunctions of F_1, say $\alpha_0 \wedge \cdots \wedge \alpha_{k-1}$, where $k \in \mathbb{N}$ and the α_j are atomic $\mathcal{L}$-formulas. Furthermore, we suppose that there are infinitely many pairs $(x,y) \in \mathbb{N}^2$ satisfying the conjunction. We assume w.l.o.g. that one of the α_j is $y + 1 > 0$.
If one of the α_j already is a nontrivial equality, we can leave the conjunction unchanged. We can forget about the trivial qualities: an always true equality can simply be omitted, and an always false equality causes the whole conjunction to be always false, which cannot be the case. So suppose all the α_j are inequalities or congruences. By lemma 1.3.1, we can write every inequality in the form $\underline{r_0}x + \underline{s_0}y + \underline{t_0} < \underline{r_1}x + \underline{s_1}y + \underline{t_1}$, where the coefficients are natural numbers. So this inequality holds for exactly those pairs of natural numbers (x,y) satisfying $r_0 x + s_0 y + t_0 < r_1 x + s_1 y + t_1$. Bringing all the y's to the left and everything else to the right, we see that the inequality holds for exactly those (x,y) satisfying $sy < rx + t$ for certain *integers* r, s and t.[19]
First suppose $s = 0$, then the inequality only puts either a lower or an upper bound on x. But it cannot put an upper bound, because that would contradict our supposition that there are infinitely many pairs (x,y) satisfying our conjunction. So they all put a lower bound on x, and we may choose x **sufficiently large** such that it satisfies all of them. Now suppose $s \neq 0$, then we can divide by s and write the inequality in one of the forms $y > px + q$ or $y < px + q$, where $p, q \in \mathbb{Q}$. The sign may change because s may be negative. We call the first form a *lower bound on y*, and the second an *upper bound on y*.

[19] Note that this is *not* an abuse of notation in $\mathcal{L}$. We are working in $\mathbb{N}$ now, and the natural numbers simply are embedded in $\mathbb{Z}$. We even have the freedom of using $\mathbb{Q}$, which we will do shortly.

Since we added $y + 1 > 0$, there is at least one lower bound. Consider all the lower bounds and pick the one with largest coefficient of x. If these are equal for some lower bounds, pick the one with largest constant coefficient. If even these are equal, then the lower bounds say the same, so it doesn't matter which one we pick. We call this lower bound, say $y > \beta x + \gamma$, the *major* lower bound. One can quite easily prove that, for x **sufficiently large**, a pair (x, y) satisfies all the lower bounds if and only if it satisfies the major one. Indeed, the major lower bound is just the one with steepest slope. Suppose that there are upper bounds as well. Then similarly, we may pick the upper bound with smallest slope, say $\delta x + \epsilon$, and call it the major upper bound. Again, for x **sufficiently large**, a pair (x, y) satisfies all the upper bounds precisely if it satisfies the major one.



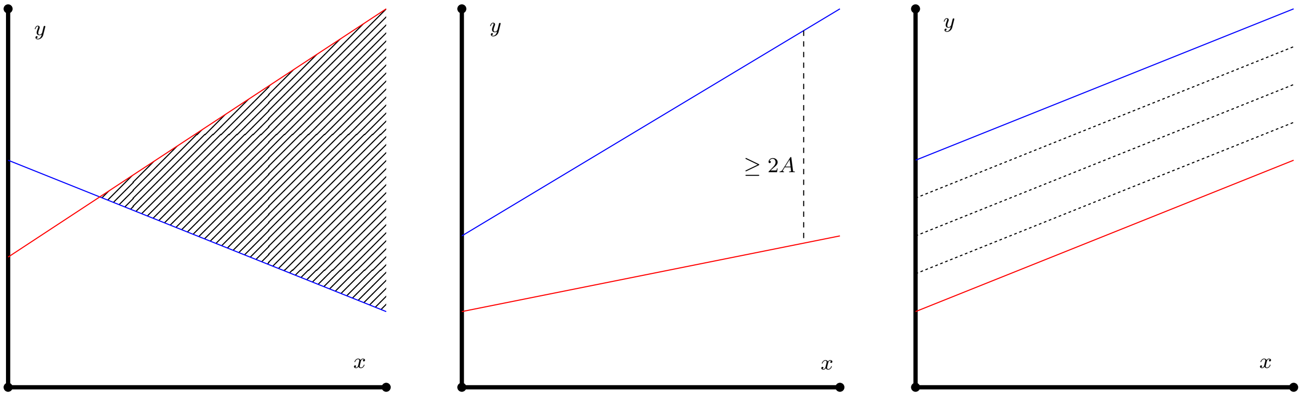
We now consider three cases.

1. $\beta > \delta$. Then the lines $y = \beta x + \gamma$ will eventually, i.e. for x large enough, lie above the line $y = \delta x + \epsilon$. But then for x **sufficiently large**, no pair (x, y) can satisfy both the upper and the lower bounds, which contradicts our supposition that there are infinitely many pairs (x, y) satisfying our conjunction. So this case cannot occur.

2. $\beta < \delta$. Define A as the least common multiple of the moduli of all the congruence symbols occurring in our disjunction. Note that, since $y + 1 > 0$ was a conjunct, we have at least one lower bound on y where the coefficient of x is zero, so β must be at least zero. Now the lines $y = \beta x + \gamma$ and $y = \delta x + \epsilon$, both with nonnegative slopes, will diverge. Thus for x **sufficiently large**, we have: for a fixed x, the values of $y \in \mathbb{N}$ that satisfy both the major upper and the major lower bound, i.e. all the inequalities, form a consecutive sequence of at least length $2A$. Furthermore, we can pick x such there is a z such that (x, z) satisfies our conjunction. We can do this because the number of such x is infinite. Now $(x, z + A)$ and $(x, z - A)$ both satisfy all the congruences as well. But also, one of those pairs satisfies all the inequalities. Therefore the function f has two values for this x, contradiction. So this case cannot occur.

3. $\beta = \delta$. Then the lines $y = \beta x + \gamma$ and $y = \delta x + \epsilon$ are parallel. So all (x, y) satisfying the inequalities must be on some parallel line between the upper and lower bound. But the number of such lines with rational coefficients that pass through lattice points of $\mathbb{N}^2$ is *finite*, as one may show with some number theory. Every equation of a line with rational coefficients can be expressed in $\mathcal{L}$; just multiply by all the denominators and bring some terms to the other side to get rid of negative numbers. So we may replace the conjunction of all our inequalities with a finite *disjunction* of equalities. Now distribute

the remaining $\wedge$'s over the $\vee$'s and obtain a disjunction of conjunctions, each of which contains an equality.



$\beta > \delta$ $\beta < \delta$ $\beta = \delta$

The case in which there are no upper bounds can be handled in a way similar to the above case $\beta < \delta$. But it is even easier, because we don't have to wait for our lower and upper bounds to diverge far enough.
We can carry out this procedure for every conjunction $\alpha_0 \wedge \cdots \wedge \alpha_{k-1}$ having an infinitude of pairs (x, y) satisfying it. Finally, we pick x **sufficiently large** such that there are no pairs (x, y) satisfying the other conjuncts any more. This proves the claim of this stage.

Stage 3. The previous stage showed that there are an $N_0 \in \mathbb{N}$ and an $\mathcal{L}$-formula $F_2(x, y)$ such that $x \geq N_0 \to (F(x, y) \leftrightarrow F_2(x, y))$, where F_2 is a disjunction of conjunctions of atomic formulas, such that each conjunction contains an equality. We may assume that each conjunction is satisfied by infinitely many pairs (x, y); if this is not already the case, we can pick N_0 a bit larger.
Now let $C(x, y)$ be one of F_2's conjunctions. We can define the set $P_x(C)$ of x's for which there is a y such that (x, y) satisfies C in $\mathcal{L}$ as $\exists z\, C(x, z)$. So this set $P_x(C)$ must be eventually periodic, by theorem 3.1.1. Let $M(C)$ be its period and suppose the periodicity holds for x larger than $N(C)$. Now consider a nontrivial equality occurring in C and write it as $sy = rx+t$ for certain integers r, s and t. Now s cannot be zero, because either $r = 0$, and the equality would be trivial, or $r \neq 0$, and there would be at most one possible value for x, contradicting our assumption. So s is nonzero, and we can divide by s, obtaining the equality $y = px + q$ for certain $p, q \in \mathbb{Q}$. Now p cannot be negative, because that would cause y to be negative for x too large. So $p \geq 0$, and if we multiply by the denominators and bring the constant to the appropriate side, we obtain an equality of the form $ay + b = cx + d$, where $a, b, c, d \in \mathbb{N}$ and $a > 0$.
Given x, the value of y is determined completely by this equality, so we may replace $C(x, y)$ by $\exists z\, C(x, z) \wedge \underline{a}y + \underline{b} = \underline{c}x + \underline{d}$. Doing this for all the conjunctions, we obtain a disjunction of formulas of this form. Now pick N larger than N_0 and all the $N(C)$ and let M be the least common multiple of all the $M(C)$. In the sequel, x will be at least N. For a conjunction C of F_2, we can replace $\exists z\, C(x, z)$ by a disjunction of formulas of the form $x \equiv_M \underline{i}$, because M is an eventual period of $P_x(C)$, since $M(C) \mid M$. Do this for all the C, and distribute the $\wedge$ in $\exists z\, C(x, z) \wedge \underline{a}y + \underline{b} = \underline{c}x + \underline{d}$ over the disjunction of congruences we introduced, obtaining a disjunction of formulas of the form $x \equiv_M \underline{i} \wedge \underline{a}y + \underline{b} = \underline{c}x + \underline{d}$. If some congruence $x \equiv_M \underline{i}$ occurs multiple times, then the corresponding equalities should be equivalent, otherwise f would have multiple values for $x \equiv i \mod M$. Now pick as a_i, b_i, c_i and d_i the coefficients of the equality corresponding to $x \equiv_M \underline{i}$. We are done. □

3.5 Possible strategies

In this final section, we will combine the ideas of the previous sections to give two possible proof strategies for our conjecture that every self-interpretation of PrA is provably trivial. By lemma 3.2.1, it suffices to show that every self-interpretation of PrA is representably trivial in $\mathbb{N}$. So let ι be a self-interpretation of PrA based on some translation τ. By theorem 3.2.3, we may assume without loss of generality that $\mathcal{D}_\tau$ is some tautology and that the equivalence $\mathcal{F}_\tau(=)(x,y) \leftrightarrow x = y$ holds in $\mathbb{N}$. If $x, y \in \mathbb{N}$, then we write $x +_* y$ for the (really!) unique z such that $\mathbb{N} \models \mathcal{F}_\tau(+)(x,y,z)$. Similarly, we write 0_* and 1_* for the (again really) unique numbers that satisfy $\mathcal{F}_\tau(0)$ respectively $\mathcal{F}_\tau(1)$ in $\mathbb{N}$. We will also use these notations in $\mathcal{L}$-formulas, and it should be obvious how to transform them into real $\mathcal{L}$-formulas.

We define the $\mathcal{L}$-formula $x <_* y$ as $\exists z\ x +_* (z +_* 1_*) = y$. Note that this is simply the translation of $\exists z\ x + (z+1) = y$, which defines the order in PrA. We also define the translated successor function s_* by the $\mathcal{L}$-formula $\mathsf{S}_*(x,y) :\Leftrightarrow y = x +_* 1_*$. Because $y = x + 1$ defines a function in PrA, its translation S_* must define a function as well. We will also refer to $<_*$ and s_* as the *internal ordering* and the *internal successor function* respectively.

We have already proven that the inner model in $\mathbb{N}$ given by ι must be isomorphic to the standard model. Since the standard model has only one automorphism, there is only one isomorphism from the standard model to the inner model in $\mathbb{N}$ given by ι, and it must be the function $f : x \mapsto \mathsf{s}_*^x(0_*)$. Our goal is now to *define* f by an $\mathcal{L}$-formula $F(x,y)$; this will be a representable isomorphism from id to ι.

The first strategy is inspired by the technique we used in section 3.3. Let $X \subset \mathbb{N}$ be the set defined by

$$\forall y\ (y < x \rightarrow y <_* x).$$

We can copy the proof of lemma 3.3.2 to show that X is cofinal in the ordering $<_*$ and infinite.[20] Since X is clearly $\mathcal{L}$-definable, it is eventually periodic with some period $P \geq 1$ and must, for x large enough, contain at least one equivalence class mod P. Let $b \in \mathbb{N}$ be such that $X_0 := \{b + Pu \mid u \in \mathbb{N}\} \subset X$.

We define the *internal difference function on* X_0, which we will denote by $\delta_*(u)$, by the $\mathcal{L}$-formula

$$\Delta_*(u,y) :\Leftrightarrow (\underline{b} + \underline{P}u) +_* y = \underline{b} + \underline{P}(u+1).$$

Note that, for $u \in \mathbb{N}$, we have $b + P(u+1) \in X_0 \subset X$. So since $b + Pu < b + P(u+1)$, we have $b + Pu <_* b + P(u+1)$, and $\delta_*(u)$ always uniquely exists.

We now make the following observation:

Lemma 3.5.1. *For every $\mathcal{L}$-definable set $A \subset \mathbb{N}$, the set $f(A) \subset \mathbb{N}$ is $\mathcal{L}$-definable as well.*

Proof. Suppose A is definable by some $\mathcal{L}$-formula $\phi(x)$. Then if $a \in A$, we have $N \models \phi(\underline{a})$, and also PrA $\vdash \phi(\underline{a})$, because PrA is the true theory of $\mathbb{N}$. But now its translation must also be provable in PrA, because ι is a self-interpretation of PrA. It takes little effort to see that this translation is equivalent to $\phi^\tau\left(\underline{f(a)}\right)$, so we have PrA $\vdash \phi^\tau\left(\underline{f(a)}\right)$ and also $\mathbb{N} \models \phi^\tau\left(\underline{f(a)}\right)$. Similarly, we can show that $a \notin A$ implies $\mathbb{N} \models \neg\phi^\tau\left(\underline{f(a)}\right)$. So $f(A)$ is $\mathcal{L}$-definable by ϕ^τ. □

[20] Because $\langle \mathbb{N}, <_* \rangle$ is isomorphic to $\langle \mathbb{N}, < \rangle$, these two properties are in fact equivalent in the current situation.

Note that the above proof does *not* need the $\mathcal{L}$-definability of f; it is therefore completely independent of the validity of conjecture 2.3.2. It is not at all clear whether the converse of lemma 3.5.1 also holds. This turns out to be an extremely interesting question, as the following theorem shows.

Theorem 3.5.2. *Let ι be a self-interpretation of $\mathtt{PrA}$ and use the notations introduced above. Then the following are equivalent:*

- *(i) ι is representably trivial in $\mathbb{N}$;*
- *(ii) the function δ_* has finite range;*
- *(iii) for every $\mathcal{L}$-definable set $A \subset \mathbb{N}$, the set $f^{-1}(A) \subset \mathbb{N}$ is $\mathcal{L}$-definable as well.*

Proof. **(i) $\implies$ (iii).** Suppose we have an $\mathcal{L}^-$-formula $F(x, y)$ representing the function $x \mapsto \mathsf{s}_*^x(0_*)$. If some $\mathcal{L}^-$-formula ϕ defines A, then we can define $f^{-1}(A)$ simply by the $\mathcal{L}^-$-formula $\exists y\ (F(x, y) \wedge \phi(y))$.

(iii) $\implies$ (ii). Suppose the function δ_* does not have finite range. Then by this assumption, the set $Y : \{u \in \mathbb{N} \mid \forall y\ (y < u \to \delta_*(y) <_* \delta_*(u))\}$ is infinite. It is also clearly $\mathcal{L}$-definable, so it must be eventually periodic with some period $\ell \geq 1$, and we can pick a $c \in \mathbb{N}$ such that $Y_0 := \{c + \ell v \mid v \in \mathbb{N}\} \subset Y$. If B is some finite set of natural numbers and $g : B \to \mathbb{N}$ is some function, we write $\sum^*_{i\in B} g(i)$ for the $+_*$-sum of all the $g(i)$, where the i ranges over B. That is, $\sum^*$ is completely analogous to $\sum$, but works with internal addition instead of the usual addition. Now for all $z \in \mathbb{N}$, we have the following *internal* telescoping series:

$$b + Pc + P\ell z = (b + Pc) +_* \sum^*_{0\le u<\ell z} \delta_*(c + u) \geq_* \sum^*_{0\le u<\ell z} \delta_*(c + u) = \sum^*_{0\le v<z} \sum^*_{0\le i<\ell} \delta_*(c + (\ell v + i))$$

$$\geq_* \sum^*_{0\le v<z} \sum^*_{0\le i<1} \delta_*(c + (\ell v + i)) = \sum^*_{0\le v<z} \delta_*(c + \ell v).$$

For $x \in \mathbb{N}$, write $\delta(x)$ for $f^{-1}(\delta_*(x))$. Now since all the numbers $c + \ell v$ are in $Y_0 \subset Y$, all terms in the above rightmost sum must be different. But that means that all the $\delta(c + \ell v)$ must be different as well. We get

$$\sum^*_{0\le v<z} \delta_*(c + \ell v) = \sum^*_{0\le v<z} \mathsf{s}_*^{\delta(c+\ell v)}(0_*) \geq_* \sum^*_{0\le v<z} \mathsf{s}_*^{v}(0_*) = \mathsf{s}_*^{1/2\cdot z(z-1)}(0_*) = f\left(\tfrac{1}{2}z(z-1)\right).$$

Combining the above two internal inequalities, we see that $b + Pc + P\ell z \geq_* f\left(\frac{1}{2}z(z-1)\right)$, which means exactly that $f^{-1}(b + Pc + P\ell z) \geq \frac{1}{2}z(z-1)$.
Now take $A = \{b + Pc + P\ell z \mid z \in \mathbb{N}\}$; it is obviously $\mathcal{L}$-definable. If $z < z'$, we have $b + Pc + P\ell z < b + Pc + P\ell z'$, and since the right hand side is in $X_0 \subset X$, we see that $b + Pc + P\ell z <_* b + Pc + P\ell z'$. Taking f^{-1} on both sides gives us $f^{-1}(b + Pc + P\ell z) < f^{-1}(b + Pc + P\ell z')$. So the sequence $f^{-1}(b+Pc), f^{-1}(b+Pc+P\ell), f^{-1}(b+Pc+P\ell\cdot 2), \ldots$ is an increasing sequence containing all the elements from $f^{-1}(A)$. By the above considerations, it grows at least quadratically. But now $f^{-1}(A)$ cannot be $\mathcal{L}$-definable, because by theorem 3.1.1, it would have to be eventually periodic, and in particular, its terms would grow at most linearly. So (iii) doesn't hold.

(ii) $\implies$ (i). Suppose the function $\delta_*(u)$ has finite range, then by corollary 3.1.4, it must be eventually periodic. Suppose the periodicity holds for $u \geq c$ and denote its period by ℓ. We set $R = b + Pc$, $K := \sum^*_{0\leq i<\ell} \delta_*(c+i)$, $r := f^{-1}(R)$ and $k := f^{-1}(K)$. Now we again apply an internal telescoping series to find:

$$\begin{aligned} R + P\ell z &= (b+Pc) + P\ell z = (b+Pc) +_* \sum^*_{0\leq u<\ell z} \delta_*(c+u) = R +_* \sum^*_{0\leq u<\ell z} \delta_*(c+u) \\ &= R +_* \sum^*_{0\leq v<z} \sum^*_{0\leq i<\ell} \delta_*(c+(\ell v+i)) = R +_* \sum^*_{0\leq v<z} \sum^*_{0\leq i<\ell} \delta_*(c+i) \\ &= R +_* \sum^*_{0\leq v<z} K = \mathsf{s}_*^r(0_*) +_* \sum^*_{0\leq v<z} \mathsf{s}_*^k(0_*) = \mathsf{s}_*^{r+kz}(0_*) = f(r+kz). \end{aligned}$$

This means that for $0 \leq i < k$, we have $\mathsf{s}_*^i(R+P\ell z) = \mathsf{s}_*^i(f(r+kz)) = f(r+(kz+i))$. Because the function s_*^i for a fixed i is $\mathcal{L}$-definable, we have a way of defining our function $f(x)$ for $x \geq r$. For $x < r$, we can define it manually. So we may take $F(x,y)$ to be

$$\bigvee_{l=0}^{r-1} \left[x = \underline{l} \wedge y = \underline{f(l)}\right] \vee \bigvee_{i=0}^{k-1} \left[\exists z \ \left(x = \underline{r+i} + \underline{k}z \wedge y = \mathsf{s}_*^i\left(\underline{R} + \underline{P\ell}z\right)\right)\right]. \qquad \square$$

This theorem tells us that, no matter what mathematical reality turns out to be, we find ourselves in an interesting situation. Either conjecture 2.3.2 is true, and conjecture 2.3.1 follows as well. In the other case, there is a self-interpretation of `PrA` such that the set of $\mathcal{L}$-definable sets suddenly becomes *larger* when we pass to the inner model. That is, there are sets that aren't definable by internal means, but that *do* form a definable set when viewed externally. This would be quite an intriguing phenomenon in itself.

Let us now indicate a second strategy, which uses the result from section 3.4. According to theorem 3.4.1, we have the following normal form for s_*:

$$x \geq \underline{N} \to \left[\mathsf{S}_*(x,y) \leftrightarrow \bigvee_{i=0}^{M-1} \left(x \equiv_M \underline{i} \wedge \underline{a_i}y + \underline{b_i} = \underline{c_i}x + \underline{d_i}\right)\right],$$

where M, N and all the a_i, b_i, c_i and d_i are natural numbers, and the a_i are nonzero. In example 2.3.3, we can write $\mathsf{S}_*(x,y)$ as

$$(x \equiv_4 \underline{0} \wedge y = x + \underline{2}) \vee (x \equiv_4 \underline{1} \wedge y = \underline{2}x + \underline{2}) \vee (x \equiv_4 \underline{2} \wedge \underline{2}y = x) \vee (x \equiv_4 \underline{3} \wedge y = \underline{2}x + \underline{2}).$$

Here $N = 0$, or in other words, the 'decent' behaviour of s_* starts immediately.

Note that the sequence $0_*, \mathsf{s}_*(0_*), \mathsf{s}_*^2(0_*), \ldots$ is a permutation of $\mathbb{N}$. Thus, we may select some $R_0 \in \mathbb{N}$ such that R_0 occurs in this sequence after every element of $\{0, 1, \ldots, N-1\}$. Now for every x such that $x \geq_* R_0$, the formula $\mathsf{S}_*(x,y)$ holds precisely if

$$\bigvee_{i=0}^{M-1} \left(x \equiv_M \underline{i} \wedge \underline{a_i}y + \underline{b_i} = \underline{c_i}x + \underline{d_i}\right)$$

holds. As we have already seen, we can write $\underline{a_i}y + \underline{b_i} = \underline{c_i}x + \underline{d_i}$ as the linear form $y = p_i x + q_i$ for certain $p_i, q_i \in \mathbb{Q}$. In order to know which of these linear forms we apply when, we want to know how the inner model walks through the clauses of the above disjunction. More precisely, we want to know how x behaves itself modulo M if we apply s_* repeatedly. It turns out that the answer to this question immediately tells us whether conjecture 2.3.2 is correct or not.

Theorem 3.5.3. *Let ι be a self-interpretation of* `PrA` *and use the notations introduced above. Then the following are equivalent:*

- *(i) ι is representably trivial in $\mathbb{N}$;*
- *(iv) the sequence $0_*, \mathsf{s}_*(0_*), \mathsf{s}_*^2(0_*), \ldots$ is eventually periodic modulo M;*
- *(v) for all $n \in \mathbb{Z}_{\geq 1}$, the sequence $0_*, \mathsf{s}_*(0_*), \mathsf{s}_*^2(0_*), \ldots$ is eventually periodic modulo n.*

Proof. **(i) $\Longrightarrow$ (v).** Suppose we have an $\mathcal{L}$-formula $F(x, y)$ representing the function $x \mapsto \mathsf{s}_*^x(0_*)$. Now we can define the relation "y is the x^{th} term in our sequence" simply by $F(x, y)$. The function that sends a number to its remainder upon division by n is also $\mathcal{L}$-definable, so we can define the function that sends x to the x^{th} term in our sequence modulo n. This function has finite range, so by corollary 3.1.4, it must be eventually periodic, which proves this direction.

(v) $\Longrightarrow$ (iv). Trivial.

(iv) $\Longrightarrow$ (i). Suppose that the sequence $0_*, \mathsf{s}_*(0_*), \mathsf{s}_*^2(0_*), \ldots$ is eventually periodic modulo M. Pick $R \geq R_0$ sufficiently large such that the sequence $R, \mathsf{s}_*(R), \mathsf{s}_*^2(R), \ldots$ is *purely* periodic modulo M, and denote its period by ℓ. We denote the remainders of the first ℓ terms of this sequence upon division by M by $k_0, \ldots, k_{\ell-1}$. In example 2.3.3, we can take $R = 0$, and the sequence $R, \mathsf{s}_*(R), \mathsf{s}_*^2(R)$ simply is the second line of the array. Modulo 4, it becomes periodic immediately, with period $0, 2, 1, 0, 2, 3$ (these are the k_j's), and $\ell = 6$.
Consider some j with $0 \leq j < \ell$. All terms in the sequence $\mathsf{s}_*^j(R), \mathsf{s}_*^{j+\ell}(R), \mathsf{s}_*^{j+\ell\cdot 2}(R), \ldots$ are congruent to k_j modulo M. Furthermore, we can get from one term in the sequence to the next by applying the linear form $y = p_i x + q_i$ successively for $i = k_j, \ldots, k_{\ell-1}, k_0, \ldots, k_{j-1}$. So the function that sends $\mathsf{s}_*^{j+\ell u}(R)$ to the next term $\mathsf{s}_*^{j+\ell(u+1)}(R)$ is itself given by a linear form, say $y = \alpha_j x + \beta_j$, where $\alpha_j, \beta_j \in \mathbb{Q}$. Moreover, α_j must be nonnegative, since all the p_i are. We claim that $\alpha_j = 1$, and we will prove this by eliminating the following two cases.

1. $\alpha_j < 1$. In that case the sequence $\mathsf{s}_*^j(R), \mathsf{s}_*^{j+\ell}(R), \mathsf{s}_*^{j+\ell\cdot 2}(R), \ldots$ converges to $\gamma := \frac{\beta_j}{1-\alpha_j}$. Because all terms of the sequence are natural numbers, γ must be in $\mathbb{N}$ and the sequence becomes eventually constant and equal to γ. But this means that an ℓ-fold application of s_* to γ gives us γ again, which is absurd, because s_* is the internal successor function.
2. $\alpha_j > 1$. Then the sequence $\mathsf{s}_*^j(R), \mathsf{s}_*^{j+\ell}(R), \mathsf{s}_*^{j+\ell\cdot 2}(R), \ldots$ grows exponentially. For $u \in \mathbb{N}$, we have $\mathsf{s}_*^{j+\ell u}(R) = \mathsf{s}_*^{\ell u}\left(\mathsf{s}_*^j(R)\right) = \mathsf{s}_*^j(R) +_* \mathsf{s}_*^{\ell u}(0_*) = \mathsf{s}_*^j(R) +_* f(\ell u)$. But since the set of multiples of ℓ is $\mathcal{L}$-definable, the set of numbers of the form $f(\ell u)$ is $\mathcal{L}$-definable as well, by lemma 3.5.1, and therefore the set of terms of our sequence must be $\mathcal{L}$-definable as well. By theorem 3.1.1, the latter set must be eventually periodic, which cannot be the case as the sequence shows exponential growth.

So α_j is indeed equal to 1, and we get $\mathsf{s}_*^{j+\ell u}(R) = \mathsf{s}_*^j(R) + \beta_j u$ for $u \in \mathbb{N}$. Because the terms $\mathsf{s}_*^{j+\ell u}(R)$ should be different natural numbers, β_j must be some *positive* natural number.
Now it is time to give an $\mathcal{L}$-formula $F(x, y)$ that represents the function $x \mapsto \mathsf{s}_*^x(0_*)$. Define $r = f^{-1}(R)$ then the previous paragraph gives us

$$\mathsf{s}_*^{r+j+\ell u}(0_*) = \mathsf{s}_*^{j+\ell u}(\mathsf{s}_*^r(0_*)) = \mathsf{s}_*^{j+\ell u}(R) = \mathsf{s}_*^j(R) + \beta_j u$$

for $u \in \mathbb{N}$. This shows that a number of the form $r + j + \ell u$ should be sent to $\mathsf{s}_*^j(R) + \beta_j u$. Because j could range $0, 1, \ldots, \ell - 1$, we have our definition for $x \geq r$. For $x < r$, we can define it manually. So we may take $F(x,y)$ to be

$$\bigvee_{l=0}^{r-1} \Big[x = \underline{l} \wedge y = \underline{f(l)} \Big] \;\vee \bigvee_{j=0}^{\ell-1} \Big[\exists u \;\Big(x = \underline{r+j} + \underline{\ell} u \wedge y = \underline{\mathsf{s}_*^j(R)} + \underline{\beta_j} u \Big) \Big]. \qquad \square$$

Given what `PrA` has shown us about $\mathcal{L}$-definability, the truth of (v), and hence also of (iv), seems rather plausible. However, the presence of *iteration* makes this particular question more complicated. A priori, the behaviour of x modulo M isn't determined by the normal form for s_*. Indeed, if all we know is that $x \equiv i \mod M$, then we can only determine $\mathsf{s}_*(x)$ modulo $p_i M$. On the other hand, it would be quite strange if a sequence arising from an $\mathcal{L}$-definable function does not behave in the manner typical for $\mathcal{L}$-definable objects, that is, periodically.

Conclusion

Let us reflect briefly in what we have done in this thesis. We formulated two conjectures concerning interpretations and $\mathtt{PrA}$. Unfortunately, we haven't been able to prove them, and they still stand as conjectures. This doesn't mean that all our efforts have been in vain. Let us list the insights and results we have gained.

- The question about the interpretability of $\mathtt{PrA}$ in $\mathtt{PrA}^-$ is connected to the study of self-interpretations of $\mathtt{PrA}$. More precisely, if all self-interpretations of $\mathtt{PrA}$ are provably trivial, then $\mathtt{PrA}$ is not interpretable in $\mathtt{PrA}^-$.
- The model $\mathcal{R}_\infty$ shows two things. First of all, we can use the interpretations $\mathcal{D}_n$ from theorem 2.2.1 at most to approximate $\mathtt{PrA}$. There is a chance we never arrive at a structure satisfying the whole of $\mathtt{PrA}$. Secondly, $\mathcal{R}_\infty$ illustrates the independence of the division axioms $\mathtt{PrAx8}_p$, for primes p.
- Suppose we have an interpretation from some theory U to $\mathtt{PrA}$. If we carry out this interpretation in $\mathbb{N}$, then we may forget concerns about the domain and about identity.
- Suppose we have a self-interpretation of $\mathtt{PrA}$. Then the inner model in $\mathbb{N}$ given by this interpretation is isomorphic to the standard model. Our conjecture 2.3.2 concerns the definability of this isomorphism.
- Several order types, like $\mathbb{Q}$, ω^2 and $\mathbb{N} + \mathbb{Z} \cdot \mathbb{Q}$ are not definable in the standard model.
- There is a decent normal form for $\mathcal{L}^-$-definable functions.
- We have discovered various statements that are equivalent to conjecture 2.3.2.

Let us end with some suggestions for further research.

- We begin by stating the obvious: find out whether conjecture 2.3.2 is true. There are, however, other interesting questions arising from this conjecture.
- Let U be some theory such that all its self-interpretations are provably trivial. We can consider the category that has theories as objects and arrows *modulo provable isomorphism* as objects.[21] Let V be a theory such that U and V are mutually interpretable. That is, there exist $\iota : U \to V$ and $\kappa : V \to U$. Now by our assumption, $\kappa \circ \iota$ is the identity arrow in this category, so $\kappa : V \to U$ is split epi. One can show that, in general, finite axiomatizability is preserved by split epis in our category. This essentially is the reasoning we employed in the proof of theorem 2.3.3. Are there other interesting mathematical properties of theories that are preserved by split epis? And can we find interesting examples of theories that have only provably trivial self-interpretations?
- If conjecture 2.3.2 is false, then there is a self-interpretation of $\mathtt{PrA}$ such that certain (externally) $\mathcal{L}^-$-definable sets grow at least quadratically when viewed internally. Can we strengthen this result, e.g. to find definable sets that internally grow faster than a fixed polynomial? Or faster than all polynomials? Would such results help us to prove conjecture 2.3.2?

[21] Note that this is *not* the same category as the one we mentioned in footnote 8.

Index of symbols

Index of terms

References

[1] Herbert B. Enderton. *A Mathematical Introduction to Logic.* Academic Press, 1972.

[2] Clemens Grabmayer. Die Entscheidungscomplexität logischer Theorien: Eine Studie anhand der Presburger Arithmetik. Master's thesis, Johannes Kepler Universität Linz, 1997.

[3] A. A. Muchnik. The definable criterion for definability in presburger arithmetic and its applications. *Theoretical Computer Science*, 290(3):1433–1444, 2003.

[4] Julia Robinson. Definability and Decision Problems in Arithmetic. *Journal of Symbolic Logic*, 14(2):98–114, 1949.

[5] Craig Smoryński. *Logical Number Theory, An Introduction*, volume I. Springer-Verlag, 1991.

[6] Albert Visser. Hume's Principle, Beginnings. *Review of Symbolic Logic*, 4(1):114–129, 2011.