The Weak Pigeonhole Principle in Models of Bounded Arithmetic

Neil Thapen
Merton College, Oxford

Trinity Term 2002

The Weak Pigeonhole Principle in Models of Bounded Arithmetic

Neil Thapen, Merton College, Oxford

Submitted for the degree of Doctor of Philosophy

Trinity Term 2002

Abstract

We develop the theory $S^1_2 + \forall a \, \mathrm{PHP}^a_{a^2}(\mathrm{PV})$, where the surjective weak pigeonhole principle $\mathrm{PHP}^a_{a^2}(\mathrm{PV})$ says that there is no polynomial time computable surjection from a onto a^2. We show that the $\forall\Sigma^b_1$ consequences of this theory can be witnessed in probabilistic polynomial time, but that the converse is unlikely to hold. Furthermore, if the cryptosystem RSA is secure then this theory does not prove the injective weak pigeonhole principle for polynomial time functions. We use this observation to show that if RSA is secure then the theory PV + (sharply bounded collection for PV formulas) lies strictly between PV and S^1_2 in strength. We also give some unconditional independence results for the relativized version of $S^1_2 + \forall a \, \mathrm{PHP}^a_{a^2}(\mathrm{PV})$. In particular, it does not prove the injective weak pigeonhole principle for an undefined function symbol.

We define a hierarchy of theories of arithmetic with a top, and use them to study the structure of a model M of bounded arithmetic by studying the relationships between its initial segments. We show that if the weak pigeonhole principle fails then for suitable $b > a^{\mathbb{N}}$ the initial segment $M \restriction b$ is definable inside $M \restriction a$ and hence is the unique end-extension of $M \restriction a$ to a model of arithmetic with a top (of this form). Conversely if any model K of arithmetic with a top is definable inside M then either K is isomorphic to an initial segment of M or vice versa, and the weak pigeonhole principle implies that the first of these holds. We also use some tools from general model theory to show that if the weak pigeonhole principle holds in a model of a weak theory of arithmetic with a top, then initial segments have more than one end-extension; in a model of a strong theory of arithmetic with a top, we can construct uncountable end-extensions of countable initial segments.

Acknowledgements

I would like to thank my supervisor Alex Wilkie for his infectious enthusiasm for mathematics, and Jan Krajíček for many stimulating conversations and for guiding me through the subject. I would also like to thank Jan Krajíček and the Mathematical Institute in Prague for their hospitality on my visits to the Czech Republic.

I have enjoyed my years as a graduate and am grateful to James McEvoy, Russell Barker, Cecily Crampin, Nicholas Peatfield and Henry Braun for giving them flavour, and to Ruanne Barnabas for giving them purpose. Finally I would like to thank my parents for their support and the British taxpayer for EPSRC grant 98001658, which made this work possible.

Contents

1 Introduction

This chapter contains a small amount of background material and a brief account of my research and the conclusions I have drawn from it. Technical details follow in the body of the thesis.

This work is motivated by the question, in which subtheories of $\mathrm{I}\Delta_0 + \Omega_1$ can the weak pigeonhole principle be proved? $\mathrm{I}\Delta_0 + \Omega_1$ was first considered by Wilkie and Paris [29] as part of a programme of research into the power of arithmetic without exponentiation. They began by studying $\mathrm{I}\Delta_0$, but this can define no function that grows faster than a polynomial and hence lacks some attractive properties of arithmetic. It cannot, for example, prove that you can uniformly substitute a term for a variable in a formula. This operation has a growth rate of something like $x \mapsto x^{\log_2 x}$, which is precisely what the axiom Ω_1 provides.

There are many mathematical principles that can be stated in the language of $\mathrm{I}\Delta_0$ (and many more in the language of $\mathrm{I}\Delta_0 + \Omega_1$, which can express anything in the polynomial hierarchy) but whose normal proofs involve the existence of much larger objects than these theories allow. Two examples are the unboundedness of the primes, normally proved using a factorial, and the pigeonhole principle, normally proved using counting, where the code for the counting function will be exponential in the size of the set counted. Paris, Wilkie and Woods in [21] showed that $\mathrm{I}\Delta_0 + \Omega_1$ proves a weak version (WPHP) of the pigeonhole principle, of the form "there is no injection from n^2 into n" and that $\mathrm{I}\Delta_0$ together with this principle proves that the set of primes is unbounded.

It is open whether $\mathrm{I}\Delta_0 + \Omega_1$ is any stronger than $\mathrm{I}\Delta_0$, in terms of which Π_1 sentences it can prove. WPHP is a candidate for a Π_1 sentence unprovable in $\mathrm{I}\Delta_0$. We say something about this in the later chapters of this thesis, but to keep things in order we will first look at a different approach to bounded arithmetic.

This is as a source for logical characterizations of complexity theoretic problems. Buss in [3] defined a hierarchy Σ_i^b of bounded formulas and a hierarchy S_2^i of theories, whose union is $\mathrm{I}\Delta_0 + \Omega_1$, such that the functions Σ_i^b definable in S_2^i are precisely those at the ith level of the polynomial hierarchy.

The axiom Ω_1 corresponds to the polynomial relationship between the length of the input to a machine and the length of its computation. It is open whether the S^i_2 hierarchy collapses to a finite level and an answer either way would have consequences in complexity theory. If it collapses, then so does the polynomial hierarchy; if it does not, then we have a reasonably well-behaved model of arithmetic in which $\mathbf{P} \neq \mathbf{NP}$.

In chapter 2 we give these definitions and also define a theory PV which directly axiomatizes the important properties of polynomial time functions. We prove the witnessing theorem for S^1_2 (theorem 2.17) by first proving an easy witnessing theorem for PV (theorem 2.10) and then showing that S^1_2 is conservative over PV for a certain class $\forall\exists$PV of formulas (corollary 2.16). We make use here of a collection principle BB(PV) (definition 2.12), provable in S^1_2, that has the consequence that Σ^b_1 formulas are equivalent to bounded existential PV formulas. The second section of this chapter contains a general result about witnessing theorems, that if a theory and a complexity class correspond (in the way that S^1_2 and $\mathbf{P}$ do) then that complexity class has a complete language. The last two sections set out some definitions of theories of arithmetic with a top that we will use later.

In this situation it is natural to consider WPHP for Σ^b_1 functions or polynomial time functions (or functions given by oracles) and to ask at what level of the S^i_2 hierarchy (or the relativized hierarchy) these are provable. We also have to distinguish between the injective WPHP, saying that there is no injection in a given class from n^2 into n, and the surjective WPHP, saying that there is no surjection in a given class from n onto n^2 (these are easily seen to be equivalent for Δ_0 functions in $\mathrm{I}\Delta_0$).

In chapter 3 we define four versions of the weak pigeonhole principle and prove them (for undefined or Σ^b_1 definable functions) in S^3_2 (corollary 3.3). This is followed by a result that we will use throughout this thesis, that, given a and $b > a^2$, any injection $a^2 \hookrightarrow a$ (surjection $a \twoheadrightarrow a^2$) can be amplified to an injection $b \hookrightarrow a$ (surjection $a \twoheadrightarrow b$) of the same complexity (lemmas 3.6 and 3.7). We then look in some detail at the theory $S^1_2 + \forall a\, \mathrm{PHP}^a_{a^2}(\mathrm{PV})$, that is, S^1_2 together with the surjective WPHP for PV formulas. We show that the $\forall\Sigma^b_1$ consequences of this theory can be witnessed in probabilistic polynomial time (theorem 3.11) but that it is unlikely that the Δ^b_1 definable sets in this theory capture everything in the complexity class **ZPP** (theorem 3.13). On the other hand, we show that if we can easily witness the injective WPHP for PV formulas, then we can crack RSA (lemma 3.15). We conclude that if RSA is secure then surjective WPHP does not prove injective WPHP in this case. See also chapter 5 where we prove a similar result unconditionally, in the relativized case.

In chapter 4 we return to the principle BB(PV). In the first section we show that if $\mathrm{PV} + \mathrm{BB}(\mathrm{PV})$ is as strong as S^1_2 then the S^i_2 hierarchy collapses. In the second section we show that if the surjective WPHP holds in a suitable model of PV, then initial segments of that model have more than one end-extension to models of PV (lemma 4.4). On the other hand we show that if PV proves BB(PV) then such end-extensions are unique in models of PV in which the injective WPHP fails (the proof of theorem 4.5). Together with the results of chapter 3 this means that if RSA is secure then $\mathrm{PV} + \mathrm{BB}(\mathrm{PV})$ lies strictly between PV and S^1_2 in strength.

In chapter 5 we prove some independence results for relativized theories. They are of the form: theory T cannot prove that a certain property holds of a structure defined by our new symbols on an interval $[0, a)$. We prove them by relating them to a problem in complexity theory: what can a polynomial time Turing machine discover about a finite structure that is given by oracles? In the first section we give an old general criterion for unprovability from the relativized version of S^2_2 (theorem 5.3) in the second section we show that the injective WPHP for a new function symbol is not provable from the relativized version of $S^1_2 + \forall a \, \mathrm{PHP}^a_{a^2}(\mathrm{PV})$ (corollary 5.6) and look for a general criterion for unprovability from this theory.

In the remaining half of the thesis we turn from looking at problems explicitly involving complexity theory and polynomial time to looking in detail at the models M of fragments of $\mathrm{I}\Delta_0 + \Omega_1$. Our method is to consider M as the union of its chain of initial segments (by which we mean the initial segments of the form $M \restriction a$, that is, with domain $\{x \in M : x < a\}$, for $a \in M$). These will be models of some kind of theory of arithmetic with a top element. Rather than study M or its theory directly, we study models of arithmetic with a top and ways of fitting such models together. We do not lose anything important by doing this because in general we are interested in bounded sentences, and any bounded sentence true in M will be true in some initial segment of M (although we have to be careful what we mean by "bounded").

Models of arithmetic with a top (by which we mean various related theories) have three main advantages here over models of fragments of $\mathrm{I}\Delta_0$. Firstly, it is very clear how much quantification is being used by a formula. Secondly, there is no need to distinguish between bounded and unbounded formulas, since every quantifier is bounded by the top element. This allows us to manipulate such structures using tools from model theory that would not work otherwise; for example initial segments of models of $\mathrm{I}\Delta_0$ have definable Skolem functions. Thirdly, models of arithmetic with a top exist on bounded domains inside models of arithmetic (with or without a top) and so can be manipulated using bounded formulas, allowing us to do some formalized model theory; see chapter 7.

We could summarize the results of the second half of this thesis as follows: WPHP holds in a model of arithmetic if and only if larger initial segments of the model are "more complex than" or "contain more information than" small initial segments.

In chapter 6 we show that if any version of WPHP fails between a and a^2 in a model K of S^1_0 of the form $[0, a^{|a|})$ then for any $k \in \mathbb{N}$ there is an end-extension J of K to a model of S^1_0 of the form $[0, a^{|a|^k})$ definable inside K (theorem 6.6). Furthermore J is the unique such end-extension, up to isomorphism over K. A consequence is that in a model of S^1_2 in which WPHP fails, increasing the interval our quantifiers range over (up to a certain point) does not increase the complexity of the Σ^b_1 sets we can define (corollary 6.8).

In chapter 7 we look for a converse to the "definable end-extension" part of theorem 6.6. We show that if a model J of arithmetic with a top is definable inside a model K of S^1_2, then J is definably isomorphic to an initial segment of K, or vice versa (theorem 7.4). If WPHP is true in K then it is the first of these that holds and the initial segment is unique; hence in models of S^1_2+WPHP we can precisely count sets if they come with sufficient internal structure (corollary 7.8). However, a precise converse of this part of theorem 6.6 is impossible (see the remark after corollary 7.6).

In chapter 8 we consider the consequences for a structure of the presence or absence of a definable surjection from a subset P onto the whole structure. This is a generalized version of the surjective WPHP. We use some tools from abstract model theory, but most of the interesting applications are to models of $\mathrm{PA}^{\mathrm{top}}$ and hence, indirectly, to $\mathrm{I}\Delta_0$. In the first section we look for converses to the "unique end-extension" part of theorem 6.6. This works for models of $\mathrm{PA}^{\mathrm{top}}$ (corollary 8.3) and we obtain a partial converse for models of S^1_0 (corollary 8.7). In the second section we characterize WPHP in terms of the possible cardinalities of initial segments of a model (corollary 8.13) and construct an uncountable model of S_2 in which the polynomial size sets are precisely the countable sets (corollary 8.14).

This completes the description of the body of this thesis. The material in chapters 4, 6 and 8 and part of chapter 3 has already appeared as [28].

2 Bounded Arithmetic

In the first section of this chapter we define the Σ_i^b formulas, the S_2^i hierarchy and a theory PV which directly axiomatizes the important properties of polynomial time functions. We prove the witnessing theorem for S_2^1 (theorem 2.17) by first proving an easy witnessing theorem for PV (theorem 2.10) and then showing that S_2^1 is conservative over PV for a certain class $\forall\exists$PV of formulas (corollary 2.16). We make use here of a collection principle BB(PV) (definition 2.12), provable in S_2^1, that has the consequence that Σ_1^b formulas are equivalent to bounded existential PV formulas. The second section contains a general result about witnessing theorems, that if a theory and a complexity class correspond (in the way that S_2^1 and **P** do) then that complexity class has a complete language. The last two sections set out some definitions of theories of arithmetic with a top that we will use later.

2.1 The theories S_2^i and PV

Our initial language consists of the constants 0 and 1 and the function symbols $+$, $\cdot$, $<$, $|\ |$ and $\#$. The intended interpretation of the length $|x|$ of x is the number of digits in the binary expansion of x. So for example 2^i has length $i+1$ and 2^i-1 has length i. The intended interpretation of the smash function $\#$ is $x\#y = 2^{|x|\cdot|y|}$. We will sometimes use the (informal) notation $\#x$ for the cut $2^{|x|^{\mathbb{N}}}$.

We take a theory BASIC fixing the algebraic properties of these symbols. See [3] or [13] for a list of axioms or, for a similar theory for a relational language, see definition 2.24.

In this thesis we will use two different definitions of bounded quantifiers and bounded formulas. Term-bounded quantifiers are appropriate when we want to model the effects of having computational resources available that are polynomial in the size of the input, and variable-bounded quantifiers when we want to model the effects of fixed computational resources. This corresponds to the difference between arithmetic with the smash function and arithmetic with a top. Term-bounded formulas will make no sense in a model with a top, and in general the bounded sets definable by term-bounded formulas are the same as those definable by variable-bounded formulas, so we will not always be careful to distinguish between the two types.

Definition 2.1

1. *A term-bounded quantifier is a quantifier of the form* $\exists x \leqslant t$ *or* $\forall x \leqslant t$ *where* t *is a term in* $+$, $\cdot$, $|\ |$ *and* $\#$ *that does not contain* x.

2. *A variable-bounded quantifier is a quantifier of the form* $\exists x \leqslant y$ *or* $\forall x \leqslant y$ *where* y *is a variable distinct from* x.

3. *A sharply bounded quantifier is a quantifier of the form* $\exists x \leqslant |t|$ *or* $\forall x \leqslant |t|$ *where* t *is a term that does not contain* x.

Definition 2.2

1. *A formula is* Δ_0 *if all of its quantifiers are variable bounded.*

2. *A formula is sharply bounded if all of its quantifiers are sharply bounded.*

3. *A formula is* Σ_i^b *if it contains no unbounded quantifiers and* $i-1$ *alternations of term-bounded quantifiers beginning with an existential quantifier and ignoring sharply bounded quantifiers. The* Π_i^b *formulas are defined dually.*

4. *A set in a structure is* Δ_i^b *if it is defined by both a* Σ_i^b *and a* Π_i^b *formula.*

Definition 2.3 *For* $i \geqslant 1$ *the theory* S_2^i *is* BASIC *together with the* Σ_i^b-LIND *axioms, consisting of*

$$\phi(0) \wedge \forall x < |y| \, (\phi(x) \rightarrow \phi(x+1)) \rightarrow \phi(|y|)$$

for each Σ_i^b *formula* ϕ *with parameters.*

The theory S_2 *is the union of the theories* S_2^i.

The most important property of S_2^1 is its ability to manipulate sequences of numbers. In particular, the function $w_i = x$, "w encodes a sequence of numbers and x is the ith element in the sequence", is Σ_1^b definable.

Proposition 2.4 (Buss [3]) *There is a* Σ_1^b *formula* $\mathrm{Comp}(e,x,t,w)$ *that expresses "*$w_1,\dots,w_{|t|}$ *encode the first* $|t|$ *configurations of the Turing machine* e *run on input* x*". Furthermore*

$$S_2^1 \vdash \forall e,x,t\, \exists! w\, \mathrm{Comp}(e,x,t,w).$$

Hence if f_e^k *is any polynomial time function computed by the Turing machine* e *with time exponent* k *for* $e,k \in \mathbb{N}$ *then the function* f_e^k *is* Σ_1^b *definable in* S_2^1*, by the formula*

$$f_e^k(x) = y \iff \exists w\, \mathrm{Comp}(e,x,2^{|x|^k},w) \wedge w_{|x|^k} = y.$$

Here we express $2^{|x|^k}$ using nested smash functions. We need to refer to the machine e (rather than just to the function that it computes) because we want to extend the function in the natural way to take nonstandard inputs.

Definition 2.5 *The language* L_{PV} *of* PV *function symbols consists of a function symbol* f_e^k *for every (standard) Turing machine* e *and every (standard) time exponent* k.

The theory $S_2^1(\mathrm{PV})$ *consists of* S_2^1 *with the addition of all the* PV *function symbols and with extra axioms*

$$f_e^k(x) = y \leftrightarrow \exists w\, \mathrm{Comp}(e,x,2^{|x|^k},w) \wedge w_{|x|^k} = y$$

expressing that the function symbols have the correct interpretations.

Clearly $S_2^1(\mathrm{PV})$ is conservative over S_2^1. We will use these two theories interchangeably.

Definition 2.6 *The theory* PV *consists of the universal consequences of* $S_2^1(\mathrm{PV})$ *in the language* L_{PV}.

Proposition 2.7 PV *proves "polynomial time recursion". That is, for all* $g,h,k \in L_{\mathrm{PV}}$*, if*

$$\mathrm{PV} \vdash \forall \bar{y}, i, x\, |g(i,x,\bar{y})| \leqslant |x| + |k(\bar{y})|$$

then there is $f \in L_{\mathrm{PV}}$ *such that*

$$\mathrm{PV} \vdash \forall \bar{y}, z, i, x\, ,\ f(0,x,\bar{y}) = h(x,\bar{y}) \wedge \forall i < |z|\, f(i+1,x,\bar{y}) = g(i,f(i,x),\bar{y}).$$

Just as the recursive functions can be defined syntactically as the closure of certain basic functions under composition and recursion, so the polynomial time functions can be defined as the closure of certain basic functions under composition, renaming and permuting of variables and a form of polynomial time recursion. Usually PV is defined as a theory formalizing this way of building up polynomial time functions, but our definition turns out to be equivalent. See [6], [3], [13].

Definition 2.8

1. *The* PV *formulas are the quantifier free formulas in the language* L_{PV}*.*
2. *The* $\exists^b$PV *formulas are those that consist of term-bounded existential quantifiers (*$+$*,* $\cdot$ *and* $\#$ *are in polynomial time, so there are* PV *function symbols for them) followed by a* PV *formula.*
3. *The theory* $\exists^b\mathrm{PV}{-}\mathrm{LIND}$ *consists of* PV *together with the axiom*

   $$\phi(0) \wedge \forall x < |y| \, (\phi(x) \to \phi(x+1)) \to \phi(|y|)$$

   for each $\exists^b$PV *formula* ϕ *with parameters.*

Proposition 2.9 PV *proves that the* PV *definable sets are closed under conjunction, negation and sharply bounded quantification.*

Proof Testing the truth of a sharply bounded quantifier only needs a polynomial number of steps. □

Since PV is a universal theory, we can easily prove a witnessing theorem for it (essentially just Herbrand's theorem). We will then go on to show that S^1_2 is $\forall\exists$PV conservative over PV, and deduce Buss' witnessing theorem for S^1_2. For a general treatment of proofs of this form see Avigad [1].

Theorem 2.10 *If* $\mathrm{PV} \vdash \forall x \, \exists y \, \phi(x,y)$ *for a* PV *formula* ϕ *then there is a* PV *function symbol* f *such that* $\mathrm{PV} \vdash \forall x \, \phi(x, f(x))$.

Proof Suppose the theorem fails and $\mathrm{PV} \vdash \forall x \, \exists y \, \phi(x,y)$ but for each PV function symbol f we have

$$\mathrm{PV} \nvdash \forall x \, \phi(x, f(x)).$$

Consider the theory $\Gamma(c) = \mathrm{PV} + \{\neg\phi(c, f(c)) : f \in L_{\mathrm{PV}}\}$. This theory is finitely satisfiable, since otherwise for a finite set $f_1, \ldots, f_n$ of PV function symbols we would have

$$\mathrm{PV} \vdash \forall x \bigvee_{i=1}^{n} \phi(x, f_i(x))$$

which implies

$$\mathrm{PV} \vdash \forall x\, \phi(x, F(x))$$

if we let F be the PV function symbol for the polynomial time machine that applies $f_1, \ldots, f_n$ to x in turn and stops when it finds $f_i(x)$ such that $\phi(x, f_i(x))$ holds (which we can check in polynomial time).

Let J be a model of $\Gamma(c)$ and let I be the submodel of J whose domain is the closure of $\{c\}$ in J under all PV function symbols. Then $I \vDash \mathrm{PV}$ (since PV is a universal theory) and $I \vDash \forall y\, \neg\phi(c, y)$. But this contradicts our assumption that $\mathrm{PV} \vdash \forall x\, \exists y\, \phi(x, y)$. $\square$

Theorem 2.11 (Zambella [30]) *The theory* $\mathrm{PV} + \exists^b\mathrm{PV}{-}\mathrm{LIND}$ *is* $\forall\exists\mathrm{PV}$ *conservative over* PV.

Proof Let $M \vDash \mathrm{PV}$ be countable. We will construct an $\exists$-elementary (in the language of PV) extension of M to a model N of PV such that for every PV formula with parameters there is a PV term f with parameters such that

$$N \vDash \forall x\, \exists y\, \phi(x, y) \rightarrow \forall x\, \phi(x, f(x)).$$

It will follow that $N \vDash \exists^b\mathrm{PV}{-}\mathrm{LIND}$ and this is sufficient for the theorem.

Let T_0 be the universal theory of M in a language expanded to include names for all the members of M (so that any model of T_0 will be an $\exists$-elementary extension of M). Let $c_0, c_1, \ldots$ be a countable collection of new constant symbols; add these to the language and enumerate as $\phi_0, \phi_1, \ldots$ all the PV formulas in this expanded language.

We will construct a sequence $T_0 \subseteq T_1 \subseteq T_2 \subseteq \ldots$ of consistent, universal theories. Suppose that T_i has been constructed. If $T_i \vdash \forall x\, \exists y\, \phi_i(x, y)$ then let $T_{i+1} = T_i$. Otherwise, let $T_{i+1} = T_i + \forall y\, \neg\phi_i(c_j, y)$ where c_j is a constant that does not occur in T_i or ϕ_i. In either case, T_{i+1} is consistent.

Let $T = \bigcup_{i \in \mathbb{N}} T_i$. Then T is consistent and has a model M'. Let N be the substructure of M' given by the set of elements named by terms in the expanded language. Then $N \vDash T$, since T is a universal theory.

Now let $\phi(x, y)$ be any PV formula with parameters from N and suppose $N \vDash \forall x \, \exists y \, \phi(x, y)$. By our construction of N, ϕ must appear in our list as ϕ_i for some i. It cannot be the case that $\forall y \, \neg \phi_i(c, y) \in T$ for any constant c, so we must have $T \vdash \forall x \, \exists y \, \phi_i(x, y)$. By Herbrand's theorem there are PV function symbols $f_1, \ldots, f_n$ with parameters such that

$$T \vdash \forall x \bigvee_{j=1}^{n} \phi_i(x, f_j(x))$$

and since we can check whether $\phi_i(x, f_j(x))$ holds in polynomial time we can combine these, as in the proof of theorem 2.10, into one PV function f such that

$$T \vdash \forall x \, \phi_i(x, f(x))$$

as desired.

We now use this property to prove that $N \vDash \exists^b\mathrm{PV}{-}\mathrm{LIND}$. Let $\phi(x, y)$ be any PV formula, in which we suppose for the sake of clarity that y is implicitly bounded. Suppose

$$N \vDash \exists y \, \phi(0, y) \wedge \forall x \!<\! |a| \, (\exists y \, \phi(x, y) \rightarrow \exists y' \, \phi(x + 1, y')).$$

We can rewrite the second conjunct as

$$N \vDash \forall x \!<\! |a| \, \forall y \, \exists y' \, (\phi(x, y) \rightarrow \phi(x + 1, y'))$$

and hence, by the construction of N, for some PV function f with parameters

$$N \vDash \forall x \!<\! |a| \, \forall y \, (\phi(x, y) \rightarrow \phi(x + 1, f(x, y))).$$

Using the recursion available in a model of PV we can iterate the function f $|a|$-many times and from the witness y for $\exists y \, \phi(0, y)$ successively find witnesses y for $\exists y \, \phi(1, y)$, $\exists y \, \phi(2, y)$, $\ldots$, $\exists y \, \phi(|a|, y)$. $\square$

Definition 2.12 *Sharply bounded collection for* PV *formulas, or* BB(PV), *is the axiom scheme*

$$\forall x\,\forall y\,,\forall i<|x|\,\exists z<y\,\phi(i,z)\rightarrow\exists w\,\forall i<|x|\,\phi(i,w_i)$$

for all PV *formulas* ϕ *with parameters.*

This is provable in S^1_2. Over PV it implies that every Σ^b_1 formula is equivalent to an $\exists^b$PV formula, since it allows us to move all the existential quantifiers to the front. However we have found no proof that the converse holds:

Open Problem 2.13 *Is* BB(PV) *implied over* PV *by the principle, every* Σ^b_1 *formula is equivalent to an* $\exists^b$PV *formula?*

Lemma 2.14 $\mathrm{PV}+\exists^b\mathrm{PV}{-}\mathrm{LIND}\vdash\mathrm{BB(PV)}$.

Proof Suppose that $\forall i<|x|\,\exists z<y\,\phi(i,z)$. Then use $\exists^b$PV$-$LIND on j in the formula $\exists w<y^j\,\forall i<j\,\phi(i,w_i)$, which is equivalent to a $\exists^b$PV formula by proposition 2.9. □

Corollary 2.15 *The theory* $\mathrm{PV}+\exists^b\mathrm{PV}{-}\mathrm{LIND}$ *proves that every* Σ^b_1 *formula is equivalent to an* $\exists^b$PV *formula. Hence* $\mathrm{PV}+\exists^b\mathrm{PV}{-}\mathrm{LIND}\vdash S^1_2$.

Together with theorem 2.11 this gives

Corollary 2.16 S^1_2 *is* $\forall\exists$PV *conservative over* PV.

We derive the witnessing theorem for S^1_2 as a corollary of this.

Theorem 2.17 (Buss [3]) *Suppose* $S^1_2\vdash\forall x\,\exists y\,\phi(x,y)$ *for a* Σ^b_1 *formula* ϕ. *Then there is a* PV *function symbol* f *such that* $S^1_2\vdash\forall x\,\phi(x,f(x))$.

Proof There is an $\exists^b$PV formula ψ such that $S^1_2\vdash\forall x\,\forall y\,,\phi(x,y)\leftrightarrow\psi(x,y)$. Hence $S^1_2\vdash\forall x\,\exists y\,\psi(x,y)$ and so by corollary 2.16, $\mathrm{PV}\vdash\forall x\,\exists y\,\psi(x,y)$. By theorem 2.10, $\mathrm{PV}\vdash\forall x\,\psi(x,f(x))$ for some PV function symbol f. Hence $S^1_2\vdash\forall x\,\phi(x,f(x))$. □

Corollary 2.18 *The subsets of* $\mathbb{N}$ *in* $\mathbf{P}$ *are precisely the subsets that are* Δ_1^b *definable in* S_2^1*, that is, definable by a* Σ_1^b *formula that is provably equivalent to a* Π_1^b *formula in* S_2^1*.*

We will also consider *relativized* theories, to which we have added extra function or relation symbols that are not defined in terms of our normal language. These are analogous to complexity classes that have been relativized to an oracle.

We look at these for two reasons. The first is that a structure in $+$, $\cdot$ etc. is rigid in a certain sense. The different parts are interconnected in subtle ways and it is difficult to modify an existing model or create a new one. But there are lots of ways to make new models by adding new symbols that satisfy the relativized theory. This should become clear in chapters 5 and 7. It is easy to prove independence results for relativized theories and hard for unrelativized theories. The second reason is to study the strength of our induction axioms by looking at what they can prove about a larger class of structures than the ones we can define in our language.

Definition 2.19 *Let* α *be a new function or relation symbol, or a set of such symbols. The* $\Sigma_i^b(\alpha)$ *formulas are the same as the* Σ_i^b *formulas except that we allow the symbols from* α *to occur.* $S_2^i(\alpha)$ *is* S_2^i *with the addition of the* LIND *axiom for all* $\Sigma_i^b(\alpha)$ *formulas, and axioms stating that the new functions are bounded by terms in* $\#$ *and so do not grow too quickly.* $S_2(\alpha)$ *and* $\mathrm{I}\Delta_0(\alpha)$ *are defined similarly.*

The functions computed by polynomial time oracle machines are $\Sigma_1^b(\alpha)$ definable in $S_2^1(\alpha)$ in the same way that the functions computed by machines without oracles are Σ_1^b definable in S_2^1, which leads to the following definition:

Definition 2.20 *The language* $L_{\mathrm{PV}}(\alpha)$ *contains a function symbol for every polynomial time Turing machine with oracles for the functions and relations in* α*. In particular* $L_{\mathrm{PV}}(\alpha)$ *contains all the functions in* α *and the characteristic functions of all the relations in* α*. The theory* $\mathrm{PV}(\alpha)$ *consists of the universal consequences of* $S_2^1(\alpha)$ *in this language, analogously with the definition of* PV *in terms of* S_2^1*. We will often refer to members of* $L_{\mathrm{PV}}(\alpha)$ *as* $\mathrm{PV}(\alpha)$ *function symbols.*

All the results in this section also hold for the relativized versions of the formulas and theories; no substantial changes to the proofs are needed.

2.2 A witnessing theorem implies a complete language

We include here a general result about witnessing theorems, by which we mean theorems showing that the sets or functions definable in a certain theory correspond precisely to the sets or functions in a certain complexity class; our canonical example is corollary 2.18. Our result is roughly that if there is such a correspondence then the complexity class has a complete language. We will use this in chapter 3 to show that it is unlikely that there is a witnessing theorem matching the theory $S^1_2 + \forall a\, \mathrm{PHP}^a_{a^2}(\mathrm{PV})$ with probabilistic polynomial time.

We give two versions. The first works for a specific kind of theory but for any complexity class; the second for a specific kind of complexity class but for any theory. We present these results in relativized form.

Theorem 2.21 *Let $\mathcal{F}$ be a class of functions and $T =_{def} S^1_2(\alpha) + \tau$ be a sound extension of $S^1_2(\alpha)$ by a single axiom τ consisting of a block of universal quantifiers and term-bounded existential quantifiers followed by a sharply bounded formula, possibly containing α. We assume that α is a one-place relation symbol. Suppose there is an oracle set A (which we will use to interpret α) such that*

1. *If $f \in \mathcal{F}$ then f is $\Sigma^b_1(\alpha)$ definable in T, that is, there is a $\Sigma^b_1(\alpha)$ formula $\phi(x,y)$ such that $T \vdash \forall x\, \exists! y\, \phi(x,y)$ and $\langle \mathbb{N}, A\rangle \vDash \forall x\, \phi(x, f(x))$;*

2. *If $T \vdash \forall x\, \exists y\, \phi(x,y)$ for $\phi \in \Sigma^b_1(\alpha)$, then there is $f \in \mathcal{F}$ such that $\langle \mathbb{N}, A\rangle \vDash \forall x\, \phi(x, f(x))$.*

Then $\mathcal{F}$ contains a complete function under logspace reducibility.

Proof Assume τ is of the form

$$\forall a_1\, \exists b_1 < t_1(a_1) \ldots \forall a_k\, \exists b_k < t_k(a_1, \ldots, a_k)\, \theta(\bar{a}, \bar{b})$$

where θ is sharply bounded and may contain the symbol α. Let $h_1, \ldots, h_k$ be new function symbols, and write $\theta^*(\bar{a}, \bar{h}(\bar{a}))$ for the formula

$$\theta(\bar{a}, h_1(a_1), h_2(a_1, a_2), \ldots, h_k(a_1, \ldots, a_k)) \\ \wedge h_1(a_1) < t_1(a_1) \wedge \ldots \wedge h_k(a_1, \ldots, a_k) < t_k(a_1, \ldots, a_k)$$

so that for a sentence σ not containing any hs,

$$T \vdash \sigma \iff S_2^1(\alpha) + \forall \bar{a}\, \theta^*(\bar{a}, \bar{h}(\bar{a})) \vdash \sigma.$$

Now let $\rho(x, e, t, w, q, r, z)$ be a formula expressing the conjunction of

1. w is a length-$|t|$ computation of the oracle Turing machine e with oracle tapes for $\alpha, h_1, \ldots, h_k$ on input x with oracle queries q and replies r and with output z;
2. At the jth computation step, queries $[\alpha(q_j^0)\,?], [h_1(q_j^1) = ?]$, $[h_2(q_j^1, q_j^2) = ?], \ldots, [h_k(q_j^1, \ldots, q_j^k) = ?]$ are made, with replies $r_j^0, r_j^1, \ldots, r_j^k$ respectively;
3. At each such step, $(r_j^0 = 1 \wedge \alpha(q_j^0)) \vee (r_j^0 = 0 \wedge \neg\alpha(q_j^0))$, $\theta(q_j^1, \ldots, q_j^k, r_j^1, \ldots, r_j^k)$ and $r_j^1 < t_1(q_j^1) \wedge \ldots \wedge r_j^k < t_k(q_j^1, \ldots, q_j^k)$ hold.

Notice that this formula states that the replies to the queries to α are correct, but not that the replies to queries to h_i are correct. In fact the symbols $h_1, \ldots, h_k$ do not appear in ρ, and we only use them in the description as convenient names for the oracle tapes.

The formula ρ is $\Sigma_1^b(\alpha)$ and

$$S_2^1(\alpha) + \forall \bar{a}\, \theta^*(\bar{a}, \bar{h}(\bar{a})) \vdash \forall x, e, t\, \exists w, q, r, z\, \rho(x, e, t, w, q, r, z).$$

This is proved by $\Sigma_1^b(\alpha)-$LIND on $|t|$. To use this we must give bounds for the existential quantifiers: w is bounded by a polynomial in t (assuming that the Turing machine looks at one tape square at a time), q and z are bounded by w and r is bounded by a polynomial in q. We can guarantee that 3 holds by replying truthfully to oracle queries.

So, by assumption, when α is interpreted as A there is $f \in \mathcal{F}$ which witnesses this sentence. That is,

$$f : (x, e, t) \mapsto (w, q, r, z)$$

and in addition there is $g \in \mathcal{F}$ with

$$g : (x, e, t) \mapsto z_{k+1}$$

where z_{k+1} is the $k + 1$st element of z considered as a sequence. We claim that g is our complete function.

To see this, take any function $i \in \mathcal{F}$. By assumption i is defined in T by some $\Sigma^b_1(\alpha)$ formula $\phi(x, y)$, so

$$\begin{aligned} S^1_2(\alpha) + \tau &\vdash \forall x \, \exists y \, \phi(x, y) \\ \Rightarrow S^1_2(\alpha) + \forall \bar{a} \, \theta^*(\bar{a}, \bar{h}(\bar{a})) &\vdash \forall x \, \exists y \, \phi(x, y) \\ &\Rightarrow S^1_2(\alpha) \vdash \forall x, \, \exists \bar{a} \, \neg\theta^*(\bar{a}, \bar{h}(\bar{a})) \vee \exists y \, \phi(x, y) \end{aligned}$$

and by the relativized version of the witnessing theorem 2.17 there is a polynomial time oracle machine e with input x and output $(\bar{a}, y)$ such that for all inputs x and oracles $A, \bar{H}$

$$\langle \mathbb{N}, A, \bar{H} \rangle \vDash \neg\theta^*(\bar{a}, \bar{h}(\bar{a})) \vee \phi(x, y).$$

Let $(w, q, r, z) = f(x, e, 2^{|x|^k})$, where $|x|^k$ is the polynomial time bound of e. Then z is of the form $(a_1, \ldots, a_k, y)$. Without loss of generality, in its computation the machine e has queried $[h_1(a_1) = ?], \ldots, [h_k(a_1, \ldots, a_k) = ?]$ and by construction the replies are such that $\theta^*(\bar{a}, \bar{h}(\bar{a}))$ holds; use these replies to define oracle functions $H_1, \ldots, H_k$. With this $\bar{H}$ we must have $\langle \mathbb{N}, A, \bar{H} \rangle \vDash \phi(x, y)$ since we cannot have $\langle \mathbb{N}, A, \bar{H} \rangle \vDash \neg\theta^*(\bar{a}, \bar{h}(\bar{a}))$, and $y = z_{k+1} = g(x, e, 2^{|x|^k})$. So $i(x) = g(x, e, 2^{|x|^k})$, and $x \mapsto (x, e, 2^{|x|^k})$ can be computed in logarithmic space. □

Our second version of this result uses the properties of the complexity class rather than the properties of the theory. Since we want it to work for more than one such class, we need a uniform way of defining complexity classes. We use the idea of a *leaf language* class; for a more thorough introduction to these see [2], where they are called *C-classes*.

A computation tree is the non-deterministic equivalent of a computation path. We will assume computation trees are always complete binary trees. The two different branches leading from each node correspond to the non-deterministic choice. The leaf at the end of a path is labelled 0 or 1 depending on whether that sequence of choices ends in acceptance or rejection of the input.

Definition 2.22 *For a nondeterministic polynomial time oracle Turing machine (NDTM) M, $x \in \mathbb{N}$ and $E \subseteq \mathbb{N}$ let $T(M, x, E)$ be the word consisting of the labels on the leaves of the computation tree of M on input x with oracle E, in lexicographic order of the computation paths.*

For disjoint sets $A, B \subseteq \{0, 1\}^*$ the leaf language class $\mathcal{C}^E(A, B)$ is the set of languages $L \subseteq \mathbb{N}$ for which there exist a NDTM M such that for all x

$$x \in L \Longleftrightarrow T(M, x, E) \in A$$
$$x \notin L \Longleftrightarrow T(M, x, E) \in B.$$

For example $\mathbf{P} = \mathcal{C}(1^*, 0^*)$; $\mathbf{NP} = \mathcal{C}(A, 0^*)$ where A is the set of strings containing at least one 1; $\mathbf{BPP} = \mathcal{C}(B, C)$ where B is the set of strings containing at least twice as many 1s as 0s and C is the set of strings containing at least twice as many 0s as 1s.

The following is partly inspired by a proof in [12].

Theorem 2.23 *Let T be any theory in any language, $E \subseteq \mathbb{N}$ and Φ, Γ any two classes of formulas in one free variable x. Suppose there is a leaf language class $\mathcal{C} = \mathcal{C}^E(A, B)$ such that*

1. *T-proofs are recognizable in polynomial time;*

2. *$\mathcal{C}$ consists precisely of the subsets of $\mathbb{N}$ expressed by formulas $\phi(x) \in \Phi$ for which there is some formula $\gamma(x) \in \Gamma$ such that $T \vdash \phi(x) \leftrightarrow \gamma(x)$;*

3. *If a formula $\phi \in \Phi$ expresses a language in $\mathcal{C}$ and π is a T-proof of $\phi(x) \leftrightarrow \gamma(x)$, for $\gamma \in \Gamma$, there is a NDTM M_π with access to the oracle E which witnesses, as in the definition of a leaf language class, that the language expressed by ϕ is in $\mathcal{C}$. Furthermore pairs (π, M_π) can be recognized in polynomial time.*

Then $\mathcal{C}$ has a complete language.

Proof Call a tuple $(\pi, M, a, 2^{|a|^k})$ *well-formed* if π is a T-proof of $\phi(x) \leftrightarrow \gamma(x)$ for some $\phi \in \Phi$ and $\gamma \in \Gamma$, M is a description of a NDTM, $M = M_\pi$, and $|a|^k$ is a time bound for M on input a.

Let K be the language consisting of all well-formed tuples $(\pi, M_\pi, a, 2^{|a|^k})$ for which $T(M_\pi, a, E) \in A$. We claim that $K \in \mathcal{C}$.

Without loss of generality, there is a language $J \in \mathcal{C}$ with $J \neq \mathbb{N}$. So there is a NDTM I and a number $z \in \mathbb{N} \setminus J$ such that $T(I, z, E) \in B$. Let N be the machine which, on input $\sigma = (\pi, M, a, 2^{|a|^k})$ first checks whether σ is well-formed. If so, it then simulates the machine M acting on input a. If not, it instead simulates the machine I acting on input z. By our assumptions, N is computable in polynomial time with oracle E.

Now consider N running on input σ. If $\sigma \in K$, then σ is well-formed, $T(M, a, E) \in A$, and N simulates M. So $T(N, \sigma, E) \in A$. If $\sigma \notin K$, then either σ is not well-formed, so that N simulates I and $T(N, \sigma, E) = T(I, z, E) \in B$ or σ is well-formed but $T(M, a, E) \in B$, so $T(N, \sigma, E) \in B$. Hence the machine N witnesses that $K \in \mathcal{C}$.

Now take any language $L \in \mathcal{C}$. For some formulas $\phi(x) \in \Phi, \gamma(x) \in \Gamma$ expressing L there is a T-proof π that $\phi(x) \leftrightarrow \gamma(x)$. Let $|x|^k$ be a time bound for the machine M_π and define the logspace computable function

$$g : a \mapsto (\pi, M_\pi, a, 2^{|a|^k}).$$

The tuple $g(a)$ is always well-formed, so $a \in L$ if and only if $T(M_\pi, a, E) \in A$ (by part 3 of our assumption about $\mathcal{C}$), and this is true if and only if $g(a) \in K$ (by the definition of K). $\square$

2.3 Arithmetic with a top

Our language will consist of the three place relations $x = y + z$ and $x = y \cdot z$, the two place relations $x < y$ and $|x| = y$ and the constants $0, 1, 2$. We use a relational language because $+$ and $\cdot$ do not define total functions and to ensure that initial segments of models of our theory are still models.

We define a theory BASIC$'$ to fix the simple properties of these symbols. Notice that no axioms (except for 2) guarantee the existence of anything larger than their parameters.

Definition 2.24 *The theory* BASIC$'$ *consists of the following axioms:*

1. $<$ *is a discrete linear ordering;*
2. $0, 1, 2$ *are the first three elements in this ordering;*
3. $+, \cdot$ *define partial functions and* $|\ |$ *a total function (hence where these are defined we will use function notation);*
4. $x + 1 = y \leftrightarrow y$ *is the successor of* x*;* $x \cdot 1 = x$*;*
5. $x + y = z \leftrightarrow y + x = z$*;* $x \cdot y = z \leftrightarrow y \cdot x = z$*;*
6. *If* $(x+y)+z = w$ *then* $y+z$ *and* $x+(y+z)$ *are defined and* $x+(y+z) = w$*; similarly for* $\cdot$*;*
7. *If* $x \cdot (y + z)$ *is defined then* $x \cdot y$*,* $x \cdot z$ *and their sum are defined and* $x \cdot (y + z) = x \cdot y + x \cdot z$*; similarly for multiplication on the left;*
8. $x + 0 = x$ *and* $x \cdot 0 = 0$ *(the rest of the normal inductive definitions of* $+$ *and* $\cdot$ *follow from axioms 4,5,6 and 7);*
9. $|0| = 0$ *and* $|1| = 1$*;*
10. $x \neq 0 \to (|2 \cdot x| = |x| + 1 \wedge |2 \cdot x + 1| = |x| + 1)$ *and when the left hand side of either of the conjucts is defined, so is the right hand side;*
11. *If* $x < y \wedge (z + y$ *is defined* $)$ *then* $(z + x$ *is defined and* $z + x < z + y)$*; if* $x < y \wedge 0 < z \wedge (z \cdot y$ *is defined* $)$ *then* $(z \cdot x$ *is defined and* $z \cdot x < z \cdot y)$*;*
12. $x \leqslant y \to |x| \leqslant |y|$*;*
13. $x < y \leftrightarrow \exists z\, (0 < z \leqslant y \wedge x + z = y)$*;*
14. $|x| + 1 < |y| \to 2 \cdot x$ *exists;*
15. $\exists y\, (x = 2 \cdot y \vee x = 2 \cdot y + 1)$.

Definition 2.25 R *is the theory consisting of* BASIC$'$ *together with an axiom stating that there is a greatest element. A model of* R *is said to be of the form* $[0, e + 1)$ *if it has a greatest element* e.

We will define a new class of formulas, the $\bar{\Sigma}_i^b$ formulas. These are very similar to the Σ_i^b formulas but their syntax is more appropriate for models with a top element, since we will use variable-bounded rather than term-bounded quantifiers. In models with a top this comes to the same thing as using unbounded quantifiers. We also need to change the definition of "sharply bounded" in this context:

Definition 2.26 *A quantifier is sharply bounded if it appears in the form $\forall x \leqslant |y|^k$ or $\exists x \leqslant |y|^k$ where x and y are variables and $k \in \mathbb{N}$.*

This definition is equivalent to our earlier definition of sharply bounded, in the structures that we consider. It does not quite work as it stands, since $|\ |$ and $\cdot$ are relation rather than function symbols. So for example $\exists x \leqslant |y|^2 \, \phi(x, y)$ written out fully is

$$\exists a \, \exists b \, (a = |y| \wedge b = a \cdot a \wedge \exists x \, (x \leqslant b \wedge \phi(x, y))).$$

This will not cause any problems, since in models of our theories $|\ |$ will always define a function and multiplication will generally be a total function when restricted to lengths.

Definition 2.27 *A formula is $\bar{\Sigma}_i^b$ if it contains no unbounded quantifiers and $i-1$ alternations of variable-bounded quantifiers beginning with a bounded existential quantifier and ignoring sharply bounded quantifiers. The $\bar{\Pi}_i^b$ formulas are defined dually. A set in a structure is $\bar{\Delta}_i^b$ if it is defined by both a $\bar{\Sigma}_i^b$ and a $\bar{\Pi}_i^b$ formula.*

Definition 2.28 *For $i \geqslant 1$, S_0^i is the theory consisting of R together with the length induction axiom*

$$[(|z|^k exists) \wedge \phi(0) \wedge \forall x < |z|^k \, (\phi(x) \rightarrow \phi(x+1))] \rightarrow \phi(|z|^k)$$

for all $\bar{\Sigma}_i^b$ formulas ϕ and all $k \in \mathbb{N}$; z is a parameter and ϕ may possibly contain other parameters. The set of length induction axioms for $\bar{\Sigma}_i^b$ formulas is called $\bar{\Sigma}_i^b$–LIND.

The theory S^1_0 is strong enough to prove that we can consider numbers as codes for binary sequences. In S^i_0 we can prove that any short binary sequence defined by a $\bar{\Delta}^b_i$ formula is coded by some number. The proofs are standard; we just need to be careful that we do not need to use any large numbers. They are included as the last section of this chapter.

Definition 2.29 $\mathrm{PA}^{\mathrm{top}}$ *is R together with an axiom scheme stating that every Δ_0 set, and hence every definable set, has a least element.*

Given a model of $K \vDash \mathrm{PA}^{\mathrm{top}}$ of the form $[0, a)$, we can construct a (unique) end-extension of K to a model of $\mathrm{PA}^{\mathrm{top}}$ of the form $[0, a^2)$ by defining natural $+$ and $\cdot$ relations on $K \times K$. Repeating this construction countably many times, we get

Proposition 2.30 *Any model of* $\mathrm{PA}^{\mathrm{top}}$ *has an end-extension to a model of* $\mathrm{I}\Delta_0$. *Hence* $\mathrm{PA}^{\mathrm{top}}$ *and* $\mathrm{I}\Delta_0$ *prove the same* Π_1 *sentences.*

This is proved in [8], where it is attributed indirectly to Paris.

We will also use relativized versions of these formulas and theories. These are defined analogously with $\Sigma^b_i(\alpha)$ and $S^i_2(\alpha)$ (see definition 2.19). The only difference is that we do not need to add axioms limiting the growth rate of any new functions we introduce, since their ranges are automatically bounded by the top element.

2.4 Bootstrapping S^1_0

We show that we can perform the basic operations of coding and decoding sequences in S^1_0.

Lemma 2.31 *Let $K \vDash S^1_0$ have greatest element e. Then the following are true in K.*

1. *The relation* $\mathrm{parity}(x) = \delta$ *given by*

$$(\delta = 1 \wedge \exists y\,(2 \cdot y + 1 = x)) \vee (\delta = 0 \wedge \exists y\,(2 \cdot y = x))$$

   defines a function.

2. *The relation* $\lfloor \frac{x}{2} \rfloor = y$ *given by*

$$2 \cdot y + \mathrm{parity}(x) = x$$

defines a function.

3. *The relation* $2^i = x$ *given by*

$$\exists y\,,\ |y| = i \wedge |x| = i + 1 \wedge x = y + 1$$

defines a function, for $i < |e|$.

4. *For all* $i < |e|$ *and all* y, $|y| \leqslant i \leftrightarrow y < 2^i$.

5. *The relation* $\mathrm{decomp}(x, i) = (y, z)$ *given by*

$$|y| \leqslant i \wedge x = y + 2^i \cdot z$$

defines a function, for $i < |e|$.

6. *For all* i, j *with* $i + j < |e|$, *we have* $2^i \cdot 2^j = 2^{i+j}$.

7. *The relation* $\mathrm{MSP}(x, i) = z$ *(standing for Most Significant Part) given by*

$$\exists y\, \mathrm{decomp}(x, i) = (y, z)$$

defines a function, for $i < |e|$.

8. *The relation* $\mathrm{bit}(x, i) = \delta$ *given by*

$$\delta = \mathrm{parity}(\mathrm{MSP}(x, i - 1))$$

defines a function, for $1 \leqslant i \leqslant |e|$.

9. *Let* $\phi(i)$ *be any* $\bar{\Delta}^b_m$ *relation. Then, if* $K \vDash \bar{\Sigma}^b_m\mathrm{-LIND}$, *we have*

$$\forall x\, \exists w\, (\forall 1 \leqslant i \leqslant |e|\,,\ \mathrm{bit}(w, i) = 1 \leftrightarrow (\mathrm{bit}(x, i) = 1 \wedge \phi(i))).$$

10. $\forall x\, \forall x'$, $(\forall i < |e|\, \mathrm{bit}(x, i) = \mathrm{bit}(x', i)) \rightarrow x = x'$.

11. $\forall i < |e|\, \forall 1 \leqslant j \leqslant |e|\, (\mathrm{bit}(2^i - 1, j) = 1 \leftrightarrow j < i)$.

12. *Let $\phi(i)$ be any $\bar{\Delta}^b_m$ relation. Then, if $K \vDash \bar{\Sigma}^b_m$−LIND is of the form $[0, 2^i)$ for some i, there is w in K such that*

$$\forall 1 \leqslant i \leqslant |e|\,,\ \mathrm{bit}(w,i) = 1 \leftrightarrow \phi(i).$$

Proof We make implicit use of the axiom guaranteeing that $+$, $\cdot$ and $|\ |$ define partial functions.

1. By axiom 15 at least one of $\mathrm{parity}(x) = 0$ and $\mathrm{parity}(x) = 1$ is true. Suppose both were true and for some $y, y' \in K$ we had $x = 2 \cdot y = 2 \cdot y' + 1$. Then

$$\begin{aligned} 2 \cdot y' &< 2 \cdot y && \text{by axiom 4} \\ \Rightarrow y' &< y && \text{by axiom 11} \\ \Rightarrow \exists z > 0\ y' + z &= y && \text{by axiom 13.} \end{aligned}$$

If $z > 1$ then by axiom 11, $y' + z > y' + 1$; so if either $z = 1$ or $z > 1$ we have $y \geqslant y' + 1$. Hence

$$\begin{aligned} 2 \cdot y &\geqslant 2 \cdot (y' + 1) \text{ [and RHS is defined]} && \text{by axiom 11} \\ \Rightarrow 2 \cdot y &\geqslant 2 \cdot y' + 2 \text{ [and RHS is defined]} && \text{by axiom 7} \\ \Rightarrow 2 \cdot y' + 1 &\geqslant 2 \cdot y' + 2 && \text{by assumption} \\ \Rightarrow 2 \cdot y' + 1 &\geqslant (2 \cdot y' + 1) + 1 && \text{by axiom 6} \end{aligned}$$

but this contradicts axiom 4.

2. By axiom 15 there is always at least one such y. Suppose

$$\begin{aligned} 2 \cdot y + \mathrm{parity}(x) = x &= 2 \cdot y' + \mathrm{parity}(x). \\ \Rightarrow 2 \cdot y &= 2 \cdot y' && \mathrm{parity}(x) \text{ is 0 or 1} \\ \Rightarrow y &= y' && \text{by axiom 11.} \end{aligned}$$

3. We will first use $\bar{\Sigma}^b_1$−LIND to show $\forall i < |e|\, \exists x\, 2^i = x$. For the base case, $|0| = 0 \wedge |1| = 1 \wedge 1 = 0{+}1$; hence $2^0 = 1$. Also $|1| = 1 \wedge |2| = 2 \wedge 2 = 1{+}1$; hence $2^1 = 2$.

Now suppose $i+1<|e|$, $i>0$ and $2^i=x$; that is, for some y we have $|y|=i \wedge |x|=i+1 \wedge x=y+1$. Then $|y|+1<|e|$ so $2\cdot y$ exists, by axiom 14. Then

$$\begin{aligned}
|2\cdot y| = |y|+1 < |e| & \qquad \text{by axiom 10}\\
\Rightarrow 2\cdot y < e & \qquad \text{by axiom 12}\\
\Rightarrow 2\cdot y+1 \text{ exists} & \qquad \text{by axiom 4}\\
|2\cdot y+1| = |y|+1 < |e| & \qquad \text{by axiom 10}\\
\Rightarrow 2\cdot y+1 < e & \qquad \text{by axiom 4}\\
\Rightarrow 2\cdot y+2 \text{ exists} & \qquad \text{by axiom 12.}
\end{aligned}$$

Let $y'=2\cdot y+1$ and $x'=2\cdot y+2$. Then $|y'|=i+1$, $x'=y'+1$ and $|x'|=|2\cdot(y+1)|=|x|+1=i+2$ so $2^{i+1}=x'$. This completes the induction.

For uniqueness, suppose $2^i=x$, $2^i=x'$ and $x'>x$. That is, for some y,y' with $y'>y$ we have

$$x'=y'+1 \wedge x=y+1 \wedge |y|=|y'|=i \wedge |x|=|x'|=i+1.$$

Now $y' \geqslant y+1=x$ so $|y'| \geqslant |x|$ by axiom 12. Hence $i \geqslant i+1$, contradicting axiom 4.

4. Let $x=2^i$. Then for some z, $|z|=i \wedge x=z+1$. So

$$y<2^i \Rightarrow y \leqslant z \Rightarrow |y| \leqslant i \qquad \text{by axiom 12}$$

and

$$y \geqslant 2^i \Rightarrow |y| \geqslant i+1 \Rightarrow |y|>i.$$

5. We will use $\bar{\Sigma}^b_1$−LIND on i to show that, given x,

$$\forall i<|e|\,\exists y,z\,\mathrm{decomp}(x,i)=(y,z).$$

For the base case $|0| \leqslant 0$, $2^0=1$ and $x=0+2^0\cdot x$ so we can put $y=0$ and $z=x$.

For the inductive case suppose $i+1 < |e|$, $|y| \leqslant i$ and $x = y + 2^i \cdot z$ for some y, z. Let $z' = \lfloor \frac{z}{2} \rfloor$. Then $z = 2 \cdot z' + \text{parity}(z)$. By the proof of (3) above, $2 \cdot 2^i = 2^{i+1}$. So

$$x = 2^i \cdot (2 \cdot z' + \text{parity}(z)) + y = 2^{i+1} \cdot z' + 2^i \cdot \text{parity}(z) + y.$$

Let $y' = 2^i \cdot \text{parity}(z) + y$. Now $2^i \cdot \text{parity}(z) \leqslant 2^i$ and by (4), since $|y| \leqslant i$, we have $y < 2^i$. Hence $y' < 2^{i+1}$ and by (4) again $|y'| \leqslant i+1$. This completes the induction.

For uniqueness, suppose $y, y' < 2^i$, $x = y + 2^i \cdot z = y' + 2^i \cdot z'$ and, without loss of generality, $y < y'$. We cannot have $z \leqslant z'$ since then by axiom 11 we would have $y + 2^i \cdot z < y' + 2^i \cdot z'$. Hence $z' < z$ so by axiom 13 for some $u, v > 0$ we have $y' = y + u$, $z = z' + v$ and $u \leqslant y' < 2^i$. So

$$\begin{aligned} x = y + 2^i \cdot z' + 2^i \cdot v = y + u + 2^i \cdot z' \qquad & \text{by axiom 7} \\ \Rightarrow y + 2^i \cdot v = y + u \qquad & \text{by axiom 11} \\ \Rightarrow 2^i \cdot v = u \qquad & \text{by axiom 11.} \end{aligned}$$

This contradicts $0 < u < 2^i$.

6. Fix i and use $\bar{\Sigma}^b_1$−LIND on j. For the base case, we have $2^i \cdot 2^0 = 2^i$. Now suppose it is true for j and that $i + j + 1 < |e|$. By the proof of (3), $\forall t < |e|\, 2^{t+1} = 2 \cdot 2^t$ so

$$2^i \cdot 2^{j+1} = 2^i \cdot 2^j \cdot 2 = 2^{i+1} \cdot 2 = 2^{i+j+1}.$$

This completes the induction.

7. Trivial.

8. Trivial.

9. Fix x. We will show by $\bar{\Sigma}^b_m$−LIND that for all $0 \leqslant j \leqslant |e|$,

$$\begin{aligned} \exists w\, \forall 1 \leqslant i \leqslant j\, [(\text{bit}(w,i) = 1 \leftrightarrow (\text{bit}(x,i) = 1 \wedge \phi(i))) \\ \wedge (j < |e| \rightarrow \text{MSP}(w,j) = \text{MSP}(x,j))]. \end{aligned}$$

Note that this can be written as a $\bar{\Sigma}^b_m$ formula.

For the base case $j = 0$, we can put $w = x$. For the inductive case, suppose $j < |e|$ and we have found a suitable w for stage j. Let $\mathrm{decomp}(w, j) = (y, z)$. $\mathrm{MSP}(x, j) = \mathrm{MSP}(w, j) = z$ so $\mathrm{bit}(x, j + 1) = \mathrm{parity}(z)$. Now we divide into two cases.

Firstly, suppose that either $\mathrm{bit}(x, j + 1) = 0$ or $\phi(j + 1)$ holds. Then we let $w' = w$ so that $\mathrm{bit}(w', j + 1) = \mathrm{parity}(z) = \mathrm{bit}(x, j + 1)$.

Secondly, suppose that $\mathrm{bit}(x, j + 1) = 1$ and $\neg\phi(j + 1)$ holds. Then $z = 2 \cdot \lfloor \frac{z}{2} \rfloor + 1$. Let $w' = y + 2^j \cdot (2 \cdot \lfloor \frac{z}{2} \rfloor)$ (this exists, by axiom 11). Then $\mathrm{bit}(w', j + 1) = 0$ as required.

In either case, if $j+1 < |e|$ then $\mathrm{MSP}(w', j+1) = \lfloor \frac{z}{2} \rfloor = \mathrm{MSP}(x, j+1)$.

Lastly we must show that once we have chosen w' the bits $1, \ldots, j$ of w' are the same as those of w. This is true in case 1, because then $w' = w$. So suppose we are in case 2 and have

$$w = y + 2^j + 2^{j+1} \cdot \lfloor \frac{z}{2} \rfloor \wedge w' = y + 2^{j+1} \cdot \lfloor \frac{z}{2} \rfloor.$$

Let $k < j$, so $j = k + l$ for some $l \geqslant 1$. Let $\mathrm{decomp}(w', k) = (u, v)$. Then $w' = u + 2^k \cdot v$, $w = u + 2^k \cdot v + 2^j = u + 2^k \cdot (v + 2^l)$ and $\mathrm{parity}(v) = \mathrm{parity}(v + 2^l)$, since $l \geqslant 1$. Hence $\mathrm{bit}(w', k + 1) = \mathrm{bit}(w, k + 1)$.

10. We will use $\bar{\Sigma}^b_1-$LIND to show

$$\forall 1 \leqslant i \leqslant |e| \, \exists j \, (j + i = |e| \wedge \mathrm{MSP}(x, j) = \mathrm{MSP}(x', j)).$$

In the base case $i = 1$ and $j + 1 = |e|$. Let $z' = \mathrm{MSP}(x', j)$ and $z = \mathrm{MSP}(x, j)$ so that $x = y + 2^j \cdot z$. Suppose $z \geqslant 2$. Then $2^j \cdot 2$ exists and by axiom 10

$$\begin{aligned} |2^j \cdot 2| &= |2^j| + 1 \text{ [and RHS is defined]} \\ &= j + 1 + 1 > |e| \end{aligned}$$

which is a contradiction. Hence $z < 2$ so $z = \mathrm{parity}(z) = \mathrm{bit}(x, j + 1)$. Similarly $z' = \mathrm{bit}(x', j + 1)$. So $z = z'$.

Now suppose $1 < i < |e|$, $j + i = |e|$ (so $j - 1$ exists) and $\mathrm{MSP}(x, j) = \mathrm{MSP}(x', j) = z$. Let $\mathrm{decomp}(x, j-1) = (u, v)$ so

$$\begin{aligned} x &= u + 2^{j-1} \cdot v \\ &= u + 2^{j-1} \cdot \mathrm{parity}(v) + 2^{j-1} \cdot 2 \cdot \lfloor \frac{v}{2} \rfloor. \end{aligned}$$

By uniqueness of MSP, we have $\lfloor \frac{v}{2} \rfloor = z$. By definition, $\mathrm{parity}(v) = \mathrm{bit}(x, j)$. Hence

$$\mathrm{MSP}(x, j-1) = v = \mathrm{bit}(x, j) + 2 \cdot z$$

and similarly

$$\mathrm{MSP}(x', j-1) = \mathrm{bit}(x', j) + 2 \cdot z = \mathrm{MSP}(x, j-1).$$

This completes the induction.

11. We use $\bar{\Sigma}^b_1$−LIND on i. The base case is $i = 0$. Then $2^i - 1 = 0$ so all of its bits are 0. For the inductive case, suppose $i + 1 < |e|$ and let $x = 2^i - 1$. Now

$$2^{i+1} = 2 \cdot 2^i = 2^i + (2^i - 1) + 1$$

and hence $2^{i+1} - 1 = x + 2^i$. Let $y = 2^{i+1} - 1$. Then $\mathrm{bit}(y, i+1) = 1$ and clearly $\forall j > i + 1\, \mathrm{bit}(y, j) = 0$. Lastly, just as in the last part of the proof of (9), $\forall 1 \leqslant k \leqslant i\, \mathrm{bit}(y, k) = \mathrm{bit}(x, k) = 1$.

12. By "K is of the form $[0, 2^i)$ for some i" we mean that the top element e is $(2^{j-1} - 1) \cdot 2 + 1$. By the proof of (11) we can show that the binary expansion of e consists of a sequence of 1s; then use (9). □

3 The Weak Pigeonhole Principle

We define four versions of the weak pigeonhole principle and prove them (for undefined or Σ_1^b definable functions) in S_2^3 (corollary 3.3). This is followed by a result that we will use throughout this thesis, that, given a and $b > a^2$, any injection $a^2 \hookrightarrow a$ (surjection $a \twoheadrightarrow a^2$) can be amplified to an injection $b \hookrightarrow a$ (surjection $a \twoheadrightarrow b$) of the same complexity (lemmas 3.6 and 3.7). We then look in some detail at the theory $S_2^1 + \forall a\, \mathrm{PHP}^a_{a^2}(\mathrm{PV})$, that is, S_2^1 together with the surjective WPHP for PV formulas. We show that the $\forall\Sigma_1^b$ consequences of this theory can be witnessed in probabilistic polynomial time (theorem 3.11) but that it is unlikely that the Δ_1^b definable sets in this theory capture everything in the complexity class **ZPP** (theorem 3.13). On the other hand, we show that if we can easily witness the injective WPHP for PV formulas, then we can crack RSA (lemma 3.15). We conclude that if RSA is secure then surjective WPHP does not prove injective WPHP in this case. See also chapter 5 where we prove a similar result unconditionally, in the relativized case.

Definition 3.1 *For $a < b$ the following are different versions of the pigeonhole principle. By the* weak pigeonhole principle, WPHP, *we mean one of these when $b \geqslant a^2$.*

1. *The bijective* PHP *states that f is not a bijection from b onto a. We will not use this very often and do not have any special notation for it.*

2. *The injective* PHP *states that f is not an injection from b into a. We write this as*

$$\mathrm{PHP}^b_a(f) \equiv \exists x < b\, f(x) \geqslant a \vee \exists x_1, x_2 < b\, (x_1 \neq x_2 \wedge f(x_1) = f(x_2)).$$

3. *The surjective* PHP *states that f is not a surjection from a onto b. We write this as*

$$\mathrm{PHP}^a_b(f) \equiv \exists y < b\, \forall x < a\, f(x) \neq y.$$

4. *The multifunction* PHP *states that the relation R is not the graph of an injective multifunction from b into a. We write this as*

$$\begin{aligned}\overline{\mathrm{mPHP}}^{b}_{a}(R) \equiv\ & \exists x < b\, \forall y < a\, \neg R(x,y) \\ & \vee \exists x_1, x_2 < b\, \exists y < a\, (x_1 \neq x_2 \wedge R(x_1,y) \wedge R(x_2,y)).\end{aligned}$$

There is an alternative form of this version, saying that the function f is not a surjection from a set $s \subseteq [0,a)$ onto b. We write this as:

$$\mathrm{mPHP}^{b}_{a}(f,s) \equiv \exists y < b\, \forall x < a\, (s(x) \rightarrow f(x) \neq y).$$

We write $\mathrm{PHP}^{x}_{y}(\Gamma)$ *for the set consisting of* $\mathrm{PHP}^{x}_{y}(f)$ *for every function in the class* Γ *(or every relation or set, as appropriate).*

In most contexts mPHP and $\overline{\mathrm{mPHP}}$ are equivalent. For example suppose R is a Σ^b_1 definable injective multifunction from b into a. Let f be the inverse of R and s the range of R; then f and s are Σ^b_1 definable and f is an surjection from s onto b. We will generally prefer to use functions and the notation of mPHP but we give our proof of WPHP for the form $\overline{\mathrm{mPHP}}$.

We usually consider these as axiom schemes asserting WPHP for every function in a certain class. In this sense, version 4 is the strongest and implies versions 2 and 3 and either of these implies version 1.

3.1 Upper and lower bounds

Theorem 3.2 (Maciel, Pitassi, Woods [17]) $S^3_2(\alpha) \vdash \forall a\, \overline{\mathrm{mPHP}}^{a^2}_{a}(\Sigma^b_1(\alpha))$.

Proof We assume that a is a power of 2; this assumption can be removed using the techniques described later in this section, by amplifying the function then restricting the domain and padding out the range.

Suppose ϕ is a $\Sigma^b_1(\alpha)$ injective multifunction from a^2 into a. The idea of the proof is to divide the domain $[0,a^2)$ of ϕ into a intervals $[0,a)$,$[a,2a), \ldots,$ $[a^2-a,a^2)$ and divide the range $[0,a)$ of ϕ into two intervals $[0,a/2),[a/2,a)$. Either one of the intervals in the domain maps entirely into the interval $[a/2,a)$, or every interval in the domain contains an element mapping to something in the interval $[0,a/2)$. In either case, we can obtain a definable injective multifunction from a set of size a into one of the intervals in the range. In the first case this is direct. In the second case we do it by first applying the multifunction which maps $x < a$ to every element in the interval $[x \cdot a, x \cdot a + a)$ and then applying ϕ and restricting the range of the resulting multifunction to $[0, a/2)$. In either case, we can compose this again with our original injection $a^2 \hookrightarrow a$ to get an injective multifunction from a^2 into $[0, a/2)$ or $[a/2, a)$.

Let S_w be the injective multifunction mapping $x < a$ to $a \cdot w + x$ if $w < a$ or mapping x to the interval $[x \cdot a, x \cdot a + a)$ if $w = a$. What we have shown is that for some $w \leqslant a$, the composition $\phi \circ S_w \circ \phi$ is an injective multifunction from $[0, a^2)$ into either $[0, a/2)$ or $[a/2, a)$.

The proof works by iterating this construction $|a|$ many times; each time we can halve the range of our multifunction, while keeping the domain the same size. We will consider numbers w in $[0, (a+1)^i)$ as codes for sequences of numbers $w_1, \ldots, w_i$ in $[0, a+1)$. For $1 \leqslant i \leqslant |a|$ and $w < (a+1)^i$ define the relation $\phi^i_w(x, y)$ as

$$\begin{aligned}\exists u < a^i,\ \phi(x, u_1) \wedge u_i = y \wedge \forall 1 \leqslant j < i, (w_j < a \wedge \phi(a \cdot w_j + u_j, u_{j+1}))\\ \vee\ (w_j = a \wedge \exists z < a\, \phi(a \cdot u_j + z, u_{j+1}));\end{aligned}$$

in the notation above this is $y \in (\phi \circ S_{w_i} \circ \ldots \circ S_{w_2} \circ \phi \circ S_{w_1} \circ \phi)(x)$ (where we apply our multifunctions from right to left, that is, apply ϕ, then S_{w_1}, then ϕ again, and so on).

Consider the formula

$$\begin{aligned}&\exists b < a\, \exists w < (a+1)^i\\ &\quad (\phi^i_w \text{ is an injective multifunction from } [0, a^2) \text{ into } [b, b + a/2^{i-1})).\end{aligned}$$

This is $\Sigma^b_3(\alpha)$. It is true for $i = 1$ and we have shown that if it is true for $i < |a|$ then it is true for $i+1$. Hence in a model of $S^3_2(\alpha)$ it is true for $i = |a|$; but this is impossible since the range will be finite. □

Corollary 3.3 *Every version of* $\mathrm{WPHP}(\Sigma^b_1(\alpha))$ *is provable in* $S^3_2(\alpha)$.

We also have a lower bound for the relativized WPHP, which we derive as a corollary of theorem 5.3 from chapter 5.

Theorem 3.4 *No version of* WPHP(f) *is provable for an undefined function symbol* f *in* $S^2_2(f)$.

Proof Let Φ be the following sentence, in the language consisting only of a two-place function symbol f:

$$\forall x_1, x_2, x_3, x_4\,[f(x_1, x_2) = f(x_3, x_4) \to (x_1 = x_3 \wedge x_2 = x_4)] \\ \wedge\, \forall y\, \exists x_1, x_2\, f(x_1, x_2) = y$$

stating that f is a bijection between the cartesian product of the universe with itself and the universe. There is an infinite model in which Φ is true, namely the natural numbers if we interpret f as the normal bijective pairing function. Hence by theorem 5.3 there is a model M of $S^2_2(f)$ and an element a of M such that $M \vDash (\langle [0,a), f\rangle \vDash \Phi)$. So f is a bijection $a \times a \longleftrightarrow a$ and we can define functions in M that violate all versions of WPHP. $\square$

3.2 Amplification

Theorem 3.5 (Paris, Wilkie, Woods [21]) *If* $K \vDash \mathrm{PA}^{\mathrm{top}}$ *is of the form* $[0, a^\varepsilon)$ *for* $a, \varepsilon \in K$, $\varepsilon \geqslant 2$ *and* K *defines a function* f *which is a surjection* $a \twoheadrightarrow a^\varepsilon$, *then* K *has an end-extension to* $J \vDash \mathrm{PA}^{\mathrm{top}}$ *of the form* $[0, a^{2^\varepsilon})$.

The proof of this theorem is essentially an amplification of f to a surjection $a \twoheadrightarrow a^{2^\varepsilon}$, and is similar to the amplification of a pseudorandom number generator to a pseudorandom function generator using a complete binary tree (see for example [7]). We give an amplification construction below that is based on a similar idea. However we use less induction than the construction in [21] and thus are only able to use one "branch" of a binary tree. Thus later on in chapter 6 when we give our version of theorem 3.5 it will only allow us to extend the size of a strucure from a^ε to a^{ε^2}, rather than to a^{2^ε}. Notice that our amplification construction is similar to the one used to polynomially increase the stretching factor of a pseudorandom number generator.

We remark that a proof of WPHP in S_2 can be derived as a corollary of theorem 3.5. In a model M of S_2, if there is $a \in M$ and a function violating WPHP definable inside $M \restriction a$, apply the theorem to $M \restriction a^{|a|}$ and get a model of $\mathrm{PA^{top}}$ of the form $[0, a^a)$. Then WPHP fails inside the logarithmic part of this model, which is impossible.

We prove our amplification lemma in S^j_0 (with a top) so that we can keep tight control over the range of quantification used. See section 2.3 for details of the definitions.

Lemma 3.6 *Suppose $j \geqslant 1$ and $K \vDash S^j_0$ is of the form $[0, a^\varepsilon)$ for $a, \varepsilon \in K$, $\varepsilon \geqslant 2$ and suppose $r(x)$ and $f(x, y)$ are $\bar{\Sigma}^b_j$ formulas violating* $\mathrm{mPHP}^a_{a^2}$*, that is, such that $r \subseteq [0, a)$ and*

$$\forall y < a^2\, \exists x < a\, (r(x) \wedge f(x, y)) \wedge \forall x < a\, (r(x) \rightarrow \exists! y < a^2\, f(x, y)).$$

Define $g(x, y)$ as

$$\exists w\, , w_1 = x \wedge \forall 1 \leqslant i < \varepsilon\, (r(w_i) \wedge f(w_i, y_i + a \cdot w_{i+1})) \\ \wedge\, r(w_\varepsilon) \wedge f(w_\varepsilon, y_\varepsilon + a \cdot 0)$$

Here w_i is the ith element of w, considering $w \in [0, a^\varepsilon)$ as code for an ε-length sequence of elements of $[0, a)$; similarly for y_i. Define $s(x)$ as

$$x < a \wedge \exists y\, g(x, y).$$

Then $s(x)$ and $g(x, y)$ are $\bar{\Sigma}^b_j$ formulas violating $\mathrm{mPHP}^a_{a^\varepsilon}$*. Furthermore if $r = [0, a)$ then $s = [0, a)$ and if f defines an injection on domain r then g defines an injection on domain s. Hence we can amplify surjections $a \twoheadrightarrow a^2$ to surjections $a \twoheadrightarrow a^\varepsilon$, injections $a^2 \hookrightarrow a$ to injections $a^\varepsilon \hookrightarrow a$ (by considering the inverses of f and g) and bijections $a \leftrightarrow a^2$ to bijections $a \leftrightarrow a^\varepsilon$.*

Proof The idea is to take a binary tree consisting of 2ε nodes labelled $w_1, \ldots, w_\varepsilon, y_1, \ldots, y_\varepsilon$. The root is labelled w_1 and for $1 \leqslant i \leqslant \varepsilon$ the node w_i has a left-hand child y_i and a right-hand child w_{i+1}, with the exception that w_ε has no right hand child. The definition of $g(x, y)$ above describes the way we assign numbers to the nodes. We consider the surjection defined by f as a function F with range r and domain $[0, a) \times [0, a)$, so that $F(v_1) = (v_2, v_3) \Leftrightarrow f(v_1, v_2 + a \cdot v_3)$. We start at the top of the tree and set $w_\varepsilon = F^{-1}(y_\varepsilon, 0)$, then $w_{\varepsilon-1} = F^{-1}(y_{\varepsilon-1}, w_\varepsilon)$, $w_{\varepsilon-2} = F^{-1}(y_{\varepsilon-2}, w_{\varepsilon-1})$ and so on. We take x to be the value w_1 at the root of the tree.

To prove that g is a surjection from $[0,a)$ onto $[0,a^\varepsilon)$, fix y and use $\bar{\Sigma}^b_j$–LIND on k in the formula

$$\exists w\,, \forall \varepsilon - k \leqslant i < \varepsilon\, (r(w_i) \wedge f(w_i, y_i + a \cdot w_{i+1})) \\ \wedge\, r(w_\varepsilon) \wedge f(w_\varepsilon, y_\varepsilon + a \cdot 0).$$

This mimics the process described above of finding labels w_i as we work down the tree. Take w witnessing this formula in the case $k = \varepsilon - 1$ and let $x = w_1$. Then $g(x, y)$.

Similarly, induction up the tree shows that if $x \in s$ and for some y, y' both $g(x,y)$ and $g(x,y')$ hold then $y_k = y'_k$ for all k, so $y = y'$. Hence g restricted to s defines a function.

If $r = [0,a)$, then let $x < a$. Put $w_1 = x$. By induction up the tree we can find w, y such that for all i $F(w_i) = (y_i, w_{i+1})$, since F is defined on all of $[0,a)$. Hence $s = [0,a)$.

If f is an injection on r, suppose for some x, y, y' we have that $g(x,y)$ and $g(x,y')$ hold and are witnessed by w and w' respectively. Induction down the tree shows that $w_i = w'_i$ for all i (or f would not be an injection) so $x = x'$. $\square$

This result immediately transfers to Σ^b_j functions in models of S^j_2 and to relativized functions and theories.

Lemma 3.7

1. *For any $f \in L_{\mathrm{PV}}$ there exists $F \in L_{\mathrm{PV}}$ such that*

   $$S^1_2 \vdash \forall a\, \forall b\, \forall c\, (\forall y < a^2\, \exists x < a\; f(c,x) = y \rightarrow \forall y < b\, \exists x < a\; F(c,b,a,x) = y).$$

   That is, any PV *surjection $a \twoheadrightarrow a^2$ with parameter c can be amplified uniformly to a surjection $a \twoheadrightarrow b$ with parameters c, b, a.*

2. *For any $g \in L_{\mathrm{PV}}$ there exists $G \in L_{\mathrm{PV}}$ such that*

   $$\mathrm{PV} \vdash \forall c\, \forall b\, \forall a < b\, \forall x_1 < x_2 < b\, \exists y_1 < y_2 < a^2\,, \\ (G(c,b,a,x_1) = G(c,b,a,x_2) \vee G(c,b,a,x_1) \geqslant a) \\ \rightarrow (g(c,y_1) = g(c,y_2) \vee g(c,y_1) \geqslant a).$$

That is, any PV *injection* $a^2 \hookrightarrow a$ *with parameter* c *can be amplified uniformly to an injection* $b \hookrightarrow a$ *with parameters* c, b, a.

Proof In part 1 the machine computing F first finds the smallest ε such that $a^\varepsilon > b$, then constructs, and uses the surjection f to label, a binary tree of length ε as in the proof of 3.6, beginning by labelling the root with the input, working up the tree, and reading the output from the leaves. As in that proof, Σ_1^b−LIND shows that F is a surjection.

In part 2 the machine computing G works in a similar way, except that it begins by labelling the leaves of the tree with the input (considered as a length-ε sequence) then working down the tree, and reading the output from the root. Again Σ_1^b−LIND proves that G is an injection if g is. Since we can write this implication in $\forall\exists$PV form it is also provable in PV, by corollary 2.16. $\square$

The surjective WPHP for PV functions is in many ways the most interesting version of WPHP. It has links with definability (as we see in the next chapter) and with randomness (in the next section). Another attractive property is that we can get rid of unwanted parameters in proofs involving it by amplifying the range of the function used until we can enlarge its domain to contain its parameters and still violate WPHP. More precisely,

Lemma 3.8 *For any* PV *function symbol* $f(c, x)$*, there is a* PV *function symbol* $G(x)$ *(with no other parameters) such that*

$$S_2^1 \vdash \forall b\, (\exists a < b\, \exists c < b\, \forall y < a^2\, \exists x < a\, f(c, x) = y \rightarrow \forall y < b^8\, \exists x < b^4\, G(x) = y),$$

that is, if f *violates surjective* WPHP *below* b *then* G *violates surjective* WPHP *at* b^4.

Proof Suppose $f(c, x) : a \twoheadrightarrow a^2$ (x here is a placeholder). Then by lemma 3.7 we have a surjection $F(c, b^8, a, x) : a \twoheadrightarrow b^8$. Define G so that

$$G : (x_1, x_2, x_3, x_4) \mapsto F(x_1, (x_2 + 1)^8, x_3, x_4).$$

Since c, $b - 1$ and a are all less than b, the range of $F(c, b^8, a, x)$ on $[0, a)$ is contained in the range of $G(\bar{x})$ on $[0, b)^4$. $\square$

Let $u(e, x, t)$ be the universal PV function symbol, which calculates the output of the Turing machine with code e run on input x for time $|t|$.

Corollary 3.9 *The theories surjective* WPHP(PV) *with parameters,*

$$\forall a \, \forall e \, \forall t \, \mathrm{PHP}^{a}_{a^2}(u(e, x, t))$$

and surjective WPHP(PV) *without parameters,*

$$\{\forall a \, \mathrm{PHP}^{a}_{a^2}(f(x)) : f \in L_{\mathrm{PV}}\}$$

are equivalent over S^1_2.

Proof The forwards implication is trivial. For the other direction, if for some e, t, a the function $u(e, x, t)$ violates WPHP at a, choose $b > e, t, a^2$ and use lemma 3.8 to find a parameter free function violating WPHP at b^4. $\square$

It would be nice if some of our later results (in particular the consequences of witnessing with a probabilistic algorithm) for surjective WPHP(PV) could be carried over to surjective WPHP(Σ^b_1). The next lemma shows that, in terms of their Σ^b_1 consequences, the two theories are not as dissimilar in strength as they might be. To use surjective WPHP(Σ^b_1) in this way we need to amend our definition of WPHP slightly to say that: no Σ^b_1 definable *relation* is the graph of a surjective function from a onto a^2. We cannot just use the version for functions because we cannot guarantee that, in an arbitrary model, our Σ^b_1 formula will define a function (if it did always define a function there would be a PV function symbol for it anyway, by the witnessing theorem).

Lemma 3.10 *Suppose* $\chi(b, u, v)$ *is a* Σ^b_1 *formula, which we will treat as a two-place formula* χ_b *with a parameter. Suppose*

$$S^1_2 \vdash \forall a, b \, [\mathrm{PHP}^{a}_{a^2}(\chi_b) \to \exists y \, \theta(a, b, y)]$$

where PHP *is of the form:* χ_b *is not the graph of a surjective function* $a \twoheadrightarrow a^2$. *Then* $\forall a, b \, \exists y \, \theta(a, b, y)$ *is provable in* $S^1_2 + \forall a \, \mathrm{PHP}^{a}_{a^2}(\mathrm{PV})$.

Proof Re-writing our assumption slightly, we have

$$S^1_2 \vdash \forall a, b\,[\neg\mathrm{PHP}^a_{a^2}(\chi_b) \vee \exists y\,\theta(a,b,y)].$$

Now $\neg\mathrm{PHP}^a_{a^2}(\chi_b)$ is the conjunction

$$\forall v\!<\!a^2\,\exists u\!<\!a\,\chi_b(u,v) \wedge \forall u\!<\!a\,\exists v\!<\!a^2\,\chi_b(u,v) \\ \wedge \forall v_1\!<\!v_2\!<\!a^2\,\forall u\!<\!a\,\neg(\chi_b(u,v_1)\wedge\chi_b(u,v_2))$$

so in particular, using only the middle conjunct,

$$S^1_2 \vdash \forall a, b\,[\forall u\!<\!a\,\exists v\!<\!a^2\,\chi_b(u,v) \vee \exists y\,\theta(a,b,y)]$$

and by the witnessing theorem 2.17 there is a PV function f such that

$$S^1_2 \vdash \forall a, b\,[\forall u\!<\!a\,(f(a,b,u) < a^2 \wedge \chi_b(u, f(a,b,u))) \vee \exists y\,\theta(a,b,y)].$$

Suppose the conclusion of the lemma fails, and there is a model M with

$$M \vDash S^1_2 + \forall x\,\mathrm{PHP}^x_{x^2}(\mathrm{PV}) + \forall y\,\neg\theta(a,b,y)$$

for some $a, b \in M$. Since $M \nvDash \exists y\,\theta(a,b,y)$, three things must hold, corresponding to the three conjuncts in WPHP:

1. $M \vDash \forall v\!<\!a^2\,\exists u\!<\!a\,\chi_b(u,v)$;

2. $M \vDash \forall u\!<\!a\,(f(a,b,u) < a^2 \wedge \chi_b(u, f(a,b,u)))$;

3. $M \vDash \forall v_1\!<\!v_2\!<\!a^2\,\forall u\!<\!a\,\neg(\chi_b(u,v_1)\wedge\chi_b(u,v_2))$.

By $\mathrm{PHP}^a_{a^2}(f)$ in M, there exists $v_1 \in M$, $v_1 < a^2$ with $\forall u\!<\!a\,f(a,b,u) \neq v_1$. By (1) for some $u < a$ we have $\chi_b(u,v_1)$. Now let $v_2 = f(a,b,u)$. By (2) $v_2 < a^2$ and $\chi_b(u,v_2)$, and of course $v_1 \neq v_2$. But this contradicts (3). □

The above results all relativize with no significant changes to the proofs.

3.3 The complexity of witnessing WPHP

We prove a relativized version of the following theorem, since we will make heavy use of it in chapter 5.

Theorem 3.11 (Wilkie [13]) *If $S^1_2(\alpha)+\forall a\,\mathrm{PHP}^a_{a^2}(\mathrm{PV}(\alpha)) \vdash \forall x\,\exists y\,\theta(x,y)$, for θ a $\Sigma^b_1(\alpha)$ formula, then there is a probabilistic polynomial time oracle Turing machine which, for any input x and any oracle A, outputs with probability at least $2/3$ some y such that $\langle \mathbb{N}, A\rangle \vDash \theta(x,y)$.*

Proof Let $u(e,w,t)$ be the universal function symbol, calculating the output of the program with code e (and access to the oracle α) run for time $|t|$ on input w. Suppose

$$S^1_2(\alpha) + \forall a\,\forall e\,\forall t\,\mathrm{PHP}^a_{a^2}(u(e,w,t)) \vdash \forall x\,\exists y\,\theta(x,y),$$

where w is a placeholder. Moving WPHP to the right hand side and using Parikh's theorem (see for example [13]) we have that for some $k \in \mathbb{N}$,

$$S^1_2(\alpha) \vdash \forall x\,,\, (\exists a,e,t < 2^{|x|^k}\,\forall v < a^2\,\exists w < a\; u(e,w,t) = v) \vee \exists y\,\theta(x,y).$$

We use the relativized version of lemma 3.8 to obtain $G \in L_{\mathrm{PV}}(\alpha)$ such that if the universal function symbol defines a surjection $a \twoheadrightarrow a^2$ for some $a < 2^{|x|^k}$ using parameters $e,t < 2^{|x|^k}$ then G defines a surjection $(2^{|x|^k})^4 \twoheadrightarrow (2^{|x|^k})^8$ using no parameters. So

$$S^1_2(\alpha) \vdash \forall x\,,\, [\forall v < 2^{8|x|^k}\,\exists w < 2^{4|x|^k}\; G(w) = v] \vee \exists y\,\theta(x,y).$$

Hence by the relativized witnessing theorem there are $\mathrm{PV}(\alpha)$ functions g_0 and g_1 such that for any oracle A,

$$\langle \mathbb{N}, A\rangle \vDash \forall x\,\forall v < 2^{8|x|^k}\,,\; g_0(x,v) < 2^{4|x|^k} \wedge (G(g_0(x,v)) = v \vee \theta(x, g_1(x,v))).$$

So given x, if we choose v at random in $[0, 2^{8|x|^k})$, with high probability we will have $\theta(x, g_1(x,v))$ since in the standard model very few of the elements of $[0, 2^{8|x|^k})$ will be in the range of G on the domain $[0, 2^{4|x|^k})$. □

A predicate B is in the class **ZPP** if there is a polynomial time probabilistic machine which gives the right answer to the question "is x in B?" with high probability and gives no answer otherwise (it never gives the wrong answer). For an oracle set A the class $\mathbf{ZPP}^A$ is defined similarly except that the machine is given access to an oracle for A.

Corollary 3.12 *Suppose that $\phi(x)$ and $\psi(x)$ are $\Sigma_1^b(\alpha)$ formulas such that*

$$S_2^1(\alpha) + \forall a\, \mathrm{PHP}^a_{a^2}(\mathrm{PV}(\alpha)) \vdash \forall x\, (\phi(x) \leftrightarrow \neg\psi(x)).$$

Then for any oracle A the set X defined in $\langle \mathbb{N}, A\rangle$ by ϕ is in $\mathbf{ZPP}^A$. In particular, if $S_2^1 + \forall a\, \mathrm{PHP}^a_{a^2}(\mathrm{PV})$ Δ_1^b-defines a predicate X then $X \in \mathbf{ZPP}$.

Proof We have that

$$S_2^1(\alpha) + \forall a\, \mathrm{PHP}^a_{a^2}(\mathrm{PV}(\alpha)) \vdash \forall x\, (\phi(x) \vee \psi(x))$$

so by the theorem there is a probabilistic polynomial time oracle machine that, on input x, outputs with high probability a w witnessing either ϕ or ψ. Consider the probabilistic polynomial time oracle machine that first computes w and then outputs "yes" if w witnesses $\phi(x)$ (this can be checked in polynomial time); "no" if w witnesses $\psi(x)$ (again in polynomial time); and "don't know" otherwise. This machine will return the right answer to "$x \in X$?" with high probability and will never return a wrong answer. $\square$

However the converse does not hold, at least in the relativized case.

Theorem 3.13 *There is an oracle set A and a set $X \in \mathbf{ZPP}^A$ such that X is not $\Delta_1^b(\alpha)$ definable in $S_2^1(\alpha) + \forall a\, \mathrm{PHP}^a_{a^2}(\mathrm{PV}(\alpha))$, where we interpret α by A.*

Proof In [2] an oracle is constructed with respect to which **ZPP** does not have a complete language. Let A be such an oracle. Then if the theorem were false, together with corollary 3.12 this would mean that the sets in $\mathbf{ZPP}^A$ are precisely the sets $\Delta_1^b(\alpha)$ definable in $S_2^1(\alpha) + \forall a\, \mathrm{PHP}^a_{a^2}(\mathrm{PV}(\alpha))$. This gives a contradiction because by theorem 2.23, since $\mathbf{ZPP}^A$ is a leaf-language class (see [2]) it must have a complete language. (To use 2.23 we also need the fact that the proof of theorem 3.11 tells us how to construct the probabilistic machine required.)

We could also use theorem 2.21. Suppose the theorem is false. Then for any probabilistic polynomial time function f (with oracle A) the set

$$\{(x, i) : \mathrm{bit}(f(x), i) = 1\}$$

is $\Delta_1^b(\alpha)$ definable in $S_2^1(\alpha) + \forall a \, \mathrm{PHP}^a_{a^2}(\mathrm{PV}(\alpha))$, since it is in $\mathbf{ZPP}^A$. Using the comprehension available in $S_2^1(\alpha)$ we can reassemble the function f from its bits, so that f is $\Sigma_1^b(\alpha)$ definable in our theory.

Together with theorem 3.11 this means that the probabilistic polynomial time functions (with oracle A) are precisely the functions $\Sigma_1^b(\alpha)$ definable in the theory; hence by theorem 2.21 there is a complete function g for this class. Then $\{x \in \mathbb{N} : g(x) = 1\}$ is a complete language for $\mathbf{ZPP}^A$. □

Definition 3.14 (Rivest, Shamir, Adleman [25]) *An instance of RSA consists of a modulus n which is the product of two large primes p and q, exponents e and d which are mutually inverse modulo $(p-1)(q-1)$, a message $m < n$ and a cyphertext $c < n$ such that $c \equiv m^e \bmod n$ and $m \equiv c^d \bmod n$. We say that we can crack RSA if, given n, e and c, we can efficiently find m.*

Lemma 3.15 (Krajíček and Pudlák [15]) *Suppose there is an efficient algorithm witnessing injective* WPHP *for* PV *function symbols with parameters, that is, given any polynomial time function f there is an algorithm which, on input c, a such that $\forall x < a^2 \, (f(c,x) < a)$, outputs $x_1 < x_2 < a^2$ such that $f(c, x_1) = f(c, x_2)$. Then we can crack RSA.*

Proof (This is a more direct version of the proof in [15].) We are given c, n and e. Without loss of generality $(c, n) = 1$ or we could factorize n, recover p and q and hence find $(p-1)(q-1)$, d and m. Hence c has an inverse modulo n.

Let s be the order of $c \bmod n$, which will be the same as the order of $m \bmod n$. Now e and s must be coprime. Otherwise let $(e, s) = t > 1$ and $u = s/t$. Then $u < s$ and $s | eu$, so $c^u \equiv m^{eu} \equiv 1$, contradicting the leastness of s.

Use the algorithm on the function $x \mapsto c^x \bmod n$ to find $x_1 < x_2 < n^2$ with $c^{x_1} \equiv c^{x_2} \bmod n$. Let $r_0 = x_2 - x_1 \neq 0$. Then $c^{r_0} \equiv 1$ so $s | r_0$.

Remove all factors of e from r_0 by calculating

$$r_1 = \frac{r_0}{(e, r_0)}, \quad r_2 = \frac{r_1}{(e, r_1)}, \quad \ldots$$

to get r with $(e, r) = 1$. This takes at most $\log r_0$ divisions.

Since $(e, s) = 1$ and $s|r_0$ we must have $s|r_1$. Similarly $s|r_2$ etc. So $s|r$. Calculate d' such that $d'e = 1 + tr$ some t; finally calculate $c^{d'} \bmod n$. Then, since $s|r$ so $m^r \equiv 1$,

$$c^{d'} \equiv m^{ed'} \equiv m^{1+tr} \equiv (m^r)^t m \equiv m. \qquad \square$$

Corollary 3.16

1. *If S^1_2 proves injective* WPHP *for* PV *functions with parameters, then RSA is vulnerable to deterministic polynomial time attack.*

2. *If S^1_2 together with the surjective* WPHP *for* PV *functions proves the injective* WPHP *for* PV *functions with parameters, then RSA is vulnerable to probabilistic polynomial time attack.*

Proof Injective WPHP for PV functions with parameters is the $\forall\Sigma^b_1$ scheme

$$\forall c \, \forall a \, [\exists x_1 < x_2 < a^2 \; f(c, x_1) = f(c, x_2) \vee \exists x < a^2 \; f(c, x) \geqslant a]$$

for PV function symbols f.

If it is provable in S^1_2 then by theorem 2.17 there is a deterministic polynomial time algorithm satisfying the assumptions of lemma 3.15.

If it is provable in $S^1_2 + \forall a \, \mathrm{PHP}^a_{a^2}(\mathrm{PV})$ then by theorem 3.11 there is a probabilistic polynomial time algorithm satisfying the assumptions of lemma 3.15. $\square$

4 Models of PV

We return to the principle BB(PV) (see definition 2.12). In the first section of this chapter we show that if PV + BB(PV) is as strong as S^1_2 then the S^i_2 hierarchy collapses. In the second section we show that if the surjective WPHP holds in a suitable model of PV, then initial segments of that model have more than one end-extension to models of PV (lemma 4.4). On the other hand we show that if PV proves BB(PV) then such end-extensions are unique in models of PV in which the injective WPHP fails (the proof of theorem 4.5). Together with the results of chapter 3 this means that if RSA is secure then PV + BB(PV) lies strictly between PV and S^1_2 in strength.

4.1 More about sharply bounded collection

Recall that sharply bounded collection for PV formulas, or BB(PV), was defined in section 2.1 as the axiom scheme

$$\forall x\, \forall y\, , \forall i{<}|x|\, \exists z{<}y\, \phi(i,z) \rightarrow \exists w\, \forall i{<}|x|\, \phi(i,w_i)$$

for all PV formulas ϕ with parameters.

We give some evidence in this section that the theory PV + BB(PV) is strictly weaker than S^1_2. The following is adapted from Zambella and uses a similar construction to our proof of the witnessing theorem 2.17.

Lemma 4.1 (Zambella [30]) *Any model $N \vDash$ PV has a $\exists^b$PV-elementary extension to a structure $M \vDash$ PV + BB(PV) such that M is the closure in M under all* PV *function symbols of the union of N with a subset $K \subseteq M$, where K has the property that for every $x \in K$ there is $y \in N$ such that $x < |y|$.*

Proof Expand the language to include a name for every element of N and let T^- be the universal theory of N in this language, so that any model of T^- is automatically an $\exists$-elementary, and hence $\exists^b$PV-elementary, extension of M. Add a new set of constant symbols $\{c_a : a \in N\}$ to the language and let T_0 be the union of T^- with the sentences $\{c_a < |a| : a \in N\}$ (this is the only relationship between c_a and a; this notation is brought in so that we can guarantee that $K \leqslant \log N$). Enumerate as $\langle \phi_1(x,y), t_1\rangle, \langle \phi_2(x,y), t_2\rangle, \ldots$ all the pairs consisting of a PV formula in two free variables and a closed term in our expanded language.

We will construct a chain $T_0 \subseteq T_1 \subseteq T_2 \subseteq \ldots$ of consistent universal theories. Suppose T_i has been constructed. Either $T_i \vdash \forall x < |t_{i+1}|\, \exists y\, \phi_{i+1}(x,y)$, or not. In the first case, put $T_{i+1} = T_i$. In the second case, choose $d \in N$ such that the constant c_d has not been used so far (outside the definition of T_0) and put

$$T_{i+1} = T_i \cup (c_d < |t_{i+1}|) \wedge \forall y \neg \phi_{i+1}(c_d, y).$$

To make sure that this is consistent we need to choose d large enough that $c_d < |d|$ is consistent with what we are adding (this is required by the definition of T_0) but we can do this because, since t_{i+1} is a term built up from elements of N and constants that have to be smaller than the length of some element of N, we can always find a d such that $t_{i+1} < d$ is consistent.

Let $T = \bigcup_i T_i$. T is a consistent theory so has a model M'. T is a universal theory, so if we let M be the substructure of M' given by all the closed terms in the expanded language, $M \vDash T$.

To show that M is a model of sharply bounded collection, let $\phi_i(x,y)$ and t_i be any PV formula and any term. Suppose $M \vDash \forall x < |t_i|\, \exists y\, \phi_i(x,y)$. Then we must have $T_{i-1} \vdash \forall x < |t_i|\, \exists y\, \phi_i(x,y)$, since otherwise a constant symbol c would have been introduced at stage i in the construction so that $M \vDash (c < |t_i|) \wedge \forall y \neg \phi_i(c,y)$. T is a universal theory so by Herbrand's theorem there is a finite set $f_1, \ldots, f_n$ of terms such that

$$T \vdash \forall x < |t_i| \bigvee_{k=1}^{n} \phi_i(x, f_k(x)).$$

As in the proof of theorem 2.10 we can combine these into one PV function F such that $\forall x < |t_i|\, \phi_i(x, F(x))$. Furthermore by the coding available in PV we can find $w \in M$ such that $\forall x < |t_i|\, w_x = F(x)$. Hence $M \vDash \forall x < |t_i|\, \phi_i(x, w_x)$. $\square$

Theorem 4.2 (Zambella [30]) *If* $\mathrm{PV} + \mathrm{BB}(\mathrm{PV}) \vdash S^1_2$, *then* $\mathrm{PV} \vdash S^1_2$.

Proof Let $N \vDash \mathrm{PV}$. Extend N to the model M of sharply bounded collection given by lemma 4.1. Suppose that $M \vDash S^1_2$. We will show that the comprehension scheme

$$\exists z \, \forall x < |a| \, (\exists y < b \, \phi(x, y) \leftrightarrow \mathrm{bit}(z, x) = 1)$$

holds in N for every PV formula ϕ. It follows that LIND holds in N for the formula $\exists y < b \, \phi(x, y)$.

Let ϕ be a PV formula with parameters in N. Then for some w in M, since $M \vDash S^1_2$,

$$M \vDash \forall x < |a| \,, w_x < b \wedge (\exists y < b \, \phi(x, y) \leftrightarrow \phi(x, w_x))$$

(this is sometimes called strong replacement, and was proved for Σ^b_1 formulas in S^1_2 by Buss [4]). By the construction of M, there exist a PV function symbol f with parameters from N and an element c of M with $c < |e|$ for some $e \in N$, such that $w = f(c)$; without loss of generality we may assume $f(u)_x < b$ for all $u, x \in M$. By the properties of PV, we can find $v \in N$ with

$$N \vDash \forall x < |a| \,, \mathrm{bit}(v, x) = 1 \leftrightarrow \exists i < |e| \, \phi(x, f(i)_x).$$

This v is precisely the element needed for our comprehension axiom for ϕ. For if $x < |a|$, $x \in N$ and $\mathrm{bit}(v, x) = 1$ then $N \vDash \exists i < |e| \, \phi(x, f(i)_x)$ so $N \vDash \exists y < b \, \phi(x, y)$. Conversely if $\mathrm{bit}(v, x) \neq 1$ then $N \vDash \forall i < |e| \, \neg\phi(x, f(i)_x)$ and the same is true in M by $\exists^b$PV-elementariness, so in particular $M \vDash \neg\phi(x, f(c)_x)$, that is, $M \vDash \neg\phi(x, w_x)$. Hence $M \vDash \forall y < b \, \neg\phi(x, y)$ and again by elementariness the same is true in N. □

Theorem 4.3 (Zambella [30], Buss [5]) *If* $\mathrm{PV} \vdash S^1_2$ *then* $\mathrm{PV} \vdash S_2$.

Hence if $\mathrm{PV} + \mathrm{BB}(\mathrm{PV})$ proves S^1_2 the bounded arithmetic hierarchy collapses.

4.2 WPHP in models of PV

We first show that if a model M of PV satisfies surjective WPHP for PV function symbols then any initial segment of M that is not too large has two different end-extensions to models of PV. Recall that $u(e, x, c)$ is the universal PV function symbol, which calculates the output of the Turing machine with code e run on input x for time $|c|$. We will use the notation $\#a$ as shorthand for the cut $2^{|a|^{\mathbb{N}}}$.

Lemma 4.4 *Let $M \vDash \mathrm{PV}$. Suppose that ε, a are nonstandard elements of M, that $\varepsilon < |a|$, that $\#a$ is not cofinal in M and that $M \vDash \forall c\, \mathrm{PHP}^{a^\varepsilon}_{a^{2\varepsilon}}(u(x, 0, c))$ (where x is a placeholder). Then there is $N \vDash \mathrm{PV}$ such that*

$$(M \restriction a) \subseteq_e N \subset (M \restriction \#a)$$

and in particular for some $v < a^{2\varepsilon}$ in M, $v \notin N$. Furthermore N and $M \restriction \#a$ are not isomorphic.

Proof Let N be the closure of $[0, a)$ in M under all PV function symbols, so $N \vDash \mathrm{PV}$, since PV is universally axiomatized. We that claim N is as required. Let $c > \#a$. By WPHP, for some $v < a^{2\varepsilon}$ in M,

$$M \vDash \forall y < a^\varepsilon\, u(y, 0, c) \neq v.$$

Consider the type

$$\Gamma(z) = \{\forall b_1, \ldots, b_n \subseteq [0, a)\, f(\bar{b}) \neq z : f \in L_{\mathrm{PV}}, n \in \mathbb{N}\}.$$

No element of N realizes this type, but v does realize it. Otherwise, we would have a function f running in time $|a|^k$, for some $k \in \mathbb{N}$, such that $v = f(\bar{b})$ for some $\bar{b} \subseteq [0, a)$. Then there is clearly some input $y < a^\varepsilon$ to the universal machine that we can construct from $(f, k, b_1, \ldots, b_n)$, with $f \in \mathbb{N}$, such that the universal machine run on y simulates f with input $\bar{b}$ and halts in time $|a|^k$. Hence $v = u(y, 0, c)$, contradicting the choice of v. □

Theorem 4.5 *Assume that* PV *proves* BB(PV). *Then* PV *also proves that surjective* WPHP(PV) *implies injective* WPHP(PV) *with parameters.*

Proof Let $M \vDash \mathrm{PV}$ satisfy surjective WPHP(PV), but be such that for some $a, c \in M$ and some PV function symbol f, $f(c, x)$ is an injection in M from a^2 into a. Choose $b > a, c$ so that $b = 2^\beta$ for some β. Amplify f to an injection $g : 2^{\beta^2} \hookrightarrow b$ with parameters in $[0, 2^{\beta^2})$. Taking an elementary extension if necessary, assume that $\#b$ is not cofinal in M.

Take ε small and nonstandard. By lemma 4.4 we can find $N \vDash \mathrm{PV}$, a submodel of M with $[0, 2^{\beta^2}) \subseteq_e N$ but $v \notin N$ for some v in $[0, 2^{\beta^2 \varepsilon})$.

Define a PV function symbol H (with parameters in $[0, 2^{\beta^2})$) mapping $x < 2^{\beta^2 \varepsilon}$ to the unique $y < b^{\varepsilon}$ such that

$$\forall 1 \leqslant i \leqslant \varepsilon \, y_i = g([x]_i)$$

where y_i is the numeral in $[0, b)$ consisting of the β bits occurring in places $((i-1)\beta + 1), \ldots, i\beta$ of y, and $[x]_i$ the numeral in $[0, 2^{\beta^2})$ consisting of the β^2 bits occurring in places $((i-1)\beta^2 + 1), \ldots, i\beta^2$ of x. This is an alternative way of amplifying g, and as before by the $\forall\exists$PV conservativity of S^1_2 over PV, H is an injection $2^{\beta^2 \varepsilon} \hookrightarrow b^{\varepsilon}$ in both M and N. H can be thought of as mapping ε-length sequences from $[0, 2^{\beta^2})$, coded in $[0, 2^{\beta^2 \varepsilon})$, injectively into ε-length sequences from $[0, b)$, coded in $[0, b^{\varepsilon})$; so H maps sequences coded in M that may not be in N to sequences coded in N.

Let $u = H(v)$. In M, we can expand u back to an ε-length sequence from $[0, 2^{\beta^2})$. Formally, $M \vDash \forall 1 \leqslant i \leqslant \varepsilon \, \exists! w < 2^{\beta^2} \, g(w) = u_i$, since we may take $w = [v]_i$ and g is injective. But this formula is also true in N, since $u \in N$, $M \restriction 2^{\beta^2}$ and $N \restriction 2^{\beta^2}$ are the same and the formula only refers to the properties of this initial segment. However $v \notin N$ and H is an injection in N, so in N we have

$$\forall 1 \leqslant i \leqslant \varepsilon \, \exists! w < 2^{\beta^2} \, g(w) = u_i \wedge \forall x < 2^{\beta^2 \varepsilon} \, H(x) \neq u$$

and this contradicts sharply bounded collection. $\square$

Notice that this in fact contradicts a weaker version of collection than BB(PV), since ε could be much smaller than β.

Corollary 4.6 *If RSA is secure against deterministic polynomial time attack, then* $\mathrm{PV} + \mathrm{BB(PV)} \nvdash S^1_2$. *If RSA is secure against randomized polynomial time attack, then* $\mathrm{PV} \nvdash \mathrm{BB(PV)}$.

Proof The second part is by theorem 4.5. For the first part, by theorems 4.2 and 4.3 if $\mathrm{PV} \vdash \mathrm{BB(PV)}$ then $\mathrm{PV} \vdash S_2$. Hence the injective WPHP is provable in PV, since it is provable in S_2. Hence it can be witnessed in polynomial time. $\square$

5 Witnessing and independence

We prove some independence results for relativized theories. They are of the form: theory T cannot prove that a certain property holds of a structure defined by our new symbols on an interval $[0, a)$. We prove them by relating them to a problem in complexity theory: what can a Turing machine discover about a finite structure that is given by oracles? In the first section we give an old general criterion for unprovability from the relativized version of S^2_2 (theorem 5.3) in the second section we show that the injective WPHP for a new function symbol is not provable from the relativized version of $S^1_2 + \forall a\, \mathrm{PHP}^a_{a^2}(\mathrm{PV})$ (corollary 5.6) and look for a general criterion for unprovability from this theory.

Related problems have been studied in cryptography, where these structures are known as "black box" structures. The proof of theorem 5.9 is based partly on an idea from Shoup [27].

5.1 Unprovability in $S^2_2(\alpha)$

Definition 5.1 *We say that a Turing machine M is given a structure $K = \langle [0, a), \alpha\rangle$ as input if it starts with the number a written on its input tape and is given access to an oracle for the relations and functions in α, where we consider constants to be 0-place functions.*

Lemma 5.2 (Riis [26]) *Let $\phi(\bar{x}, \bar{y})$ be an open formula in a relational language α disjoint from our language of arithmetic. Let Φ be the sentence $\forall \bar{x}\, \exists \bar{y}\, \phi(\bar{x}, \bar{y})$. Suppose Φ has an infinite model. Then there is no oracle machine $M \in \mathbf{P}^{\mathbf{NP}}$ which, when given any structure $\langle [0, a), \alpha\rangle$ as input, outputs a tuple $\bar{x} < a$ such that $\langle [0, a), \alpha\rangle \vDash \forall \bar{y}\, \neg\phi(\bar{x}, \bar{y})$.*

Hence by the full, relativized version of the witnessing theorem 2.17 if Φ has an infinite model then $S^2_2(\alpha) \nvdash \forall a\, (\langle [0, a), \alpha\rangle \vDash \neg\Phi)$.

Proof We follow the presentation in [13]. For the sake of simplicity we will only consider the case in which our new language is a single binary relation symbol α.

Let $\langle K, A\rangle$ be an infinite structure satisfying $\forall \bar{x}\, \exists \bar{y}\, \phi(\bar{x}, \bar{y})$, where we interpret α by the binary relation A. Suppose for a contradiction that there is a deterministic oracle machine M and a non-deterministic oracle machine N such that M^N, given any structure $\langle [0,a), \alpha \rangle$ as input, outputs a witness to $\langle [0,a), \alpha \rangle \models \neg\Phi$. Let the running times of M and N be bounded by polynomials p and q respectively.

Fix $a \in \mathbb{N}$ sufficiently large and let S be the set of partial functions $[0,a) \to K$. We will construct $\sigma \in S$ as follows: we begin with σ empty, and start a computation of M on a; we will add a small number of pairs to σ at each step in the computation.

We may assume that at each step an oracle query $[b?]$ is made to N. Consider all possible extensions $\tau \in S$ of σ and all possible computation paths w of the non-deterministic machine N on b. If, for some such τ and some such w, τ is defined on all elements $z_1, z_2 \in [0,a)$ for which the computation w queries $[\alpha(z_1,z_2)?]$ and if w is an accepting path if we reply to such queries with $A^\tau(z_1,z_2) =_{def} A(\tau(z_1),\tau(z_2))$, then return "yes" and add $\langle z_1, \tau(z_1)\rangle$ and $\langle z_2, \tau(z_2)\rangle$ to σ for all pairs z_1, z_2 queried in w. This adds at most $|w| \leqslant q(p(|a|))$ elements to σ. If there is no such extension τ and path w, return "no" and leave σ unchanged.

By step i of the computation, we will have a partial function σ such that $|\sigma| \leqslant i \cdot q(p(|a|))$ and for any $\tau \in S$ with $\sigma \subseteq \tau$, on input a the two machines M^N with oracle A^σ and M^N with oracle A^τ will be in the same configuration after i steps.

Let σ' be the end result of this construction, when the computation of M has finished. Then $|\sigma'| \ll a$ and for any partial function τ extending σ', on input a the machine M^N with oracle A^τ outputs a witness $\bar{b}$ such that

$$\forall \bar{y} < a \, \neg(\langle [0,a), A^\tau \rangle \models \phi(\bar{b}, \bar{y})).$$

But this is a contradiction since by the assumption that $\langle K, \alpha \rangle \models \forall \bar{x} \, \exists \bar{y} \, \phi(x,y)$ there is some tuple $\bar{c} \in K$ for which $\langle K, A \rangle \models \phi(\sigma'(\bar{b}), \bar{c})$ (without loss of generality $\bar{b}$ is in the domain of σ') and we may extend σ' to $\tau \in S$ with $\bar{c} = \tau(\bar{d})$ for some $\bar{d} \subseteq [0,a)$; so

$$\bar{d} < a \wedge (\langle [0,a), A^\tau \rangle \models \phi(\bar{b}, \bar{d})). \qquad \square$$

Theorem 5.3 (Riis [26]) *Let Φ be any sentence containing only symbols from a language α disjoint from the language of arithmetic. If Φ has an* *infinite model then*

$$S_2^2(\alpha) \nvdash \forall a\,(\langle[0,a),\alpha\rangle \vDash \neg\Phi).$$

Proof We can convert Φ into a form Ψ to which lemma 5.2 applies by first Skolemizing and then replacing each function symbol f with a relation symbol R_f, each time conjoining $\forall\exists$ formulas stating that R_f defines a function. The lemma shows that there is a model of $S_2^2 + (\langle[0,a),\alpha\rangle \vDash \Psi)$ with induction for all these relation symbols. We can now re-introduce symbols for the functions, since we have ensured that the relevant relations define functions in this model. $\square$

5.2 Unprovability in $S_2^1(\alpha) + \forall a\, \mathrm{PHP}^a_{a^2}(\mathrm{PV}(\alpha))$

We study the limits of what we can prove in $S_2^1(\alpha) + \forall a\, \mathrm{PHP}^a_{a^2}(\mathrm{PV}(\alpha))$ about the structures $\langle[0,a),\alpha\rangle$ defined by the language α on initial segments of models. The ultimate goal is to find an analogue of Riis' sufficient condition for unprovability in $S_2^2(\alpha)$ (theorem 5.3). The best candidate for a condition is, roughly, that Φ has arbitrarily large models in which the number of witnesses to Φ is small. However we run into problems in proving this, because in this setting we do not have the machinery Riis uses to get rid of function symbols.

We obtain our results by showing that it is hard for a probabilistic polynomial time oracle machine to find certain elements of a structure it is given as input, then applying theorem 3.11. We need an inequality:

Lemma 5.4 *If $a, t \in \mathbb{N}$ and $4t^2 < a$ then $(1 - t/a)^t > 3/4$.*

Proof Using the binomial expansion,

$$\begin{aligned}\left(1 - \frac{t}{a}\right)^t &= 1 - t\left(\frac{t}{a}\right) + \frac{t(t-1)}{2}\left(\frac{t}{a}\right)^2 - \frac{t(t-1)(t-2)}{3\cdot 2}\left(\frac{t}{a}\right)^3 \\ &\qquad + \ldots + \left(-\frac{t}{a}\right)^t.\end{aligned}$$

Since $t^2 < a$, we have

$$\binom{t}{i+1}\left(\frac{t}{a}\right)^{i+1} = \frac{t-i}{i+1}\cdot\frac{t}{a}\cdot\binom{t}{i}\left(\frac{t}{a}\right)^i < \binom{t}{i}\left(\frac{t}{a}\right)^i.$$

So the absolute value of each term in our expansion is smaller than the last, and we may remove (in pairs) all but the first two terms to give

$$\left(1 - \frac{t}{a}\right)^t > 1 - \frac{t^2}{a} > \frac{3}{4}. \qquad \square$$

Lemma 5.5 *There is no probabilistic polynomial time oracle machine which, given a structure $\langle [0,a), f\rangle$ as input where f is a two-place function, outputs with probability at least $2/3$ distinct pairs (x_1, x_2), (y_1, y_2) witnessing that f satisfies injective* WPHP*, that is, such that $f(x_1, x_2) = f(y_1, y_2)$.*

Proof Suppose that this is false and that such a machine M exists, running with time exponent $k \in \mathbb{N}$. Choose $a \in \mathbb{N}$ so that a is bigger than $4|a|^{2k}$, and let $t = |a|^k$.

For $c \in \{0,1\}^t$ let M_c be the deterministic machine obtained by replacing the random choices in M with the fixed sequence c of coin tosses. There are a^{a^2} many possible functions f on our domain. By our assumption, for each such function, for $2/3$ of the possible coin tosses, the machine M_c will find elements witnessing WPHP. Hence there must be some sequence c of coin tosses such that M_c will find witnesses for $2/3$ of the possible functions.

Fix such a c and start a computation of M_c. We will construct an oracle function f step by step. We assume that at each step in the computation one oracle query $[f(x_1, x_2) =?]$ is made, with $x_1, x_2 < a$. We construct f by replying to each such query with a random number in $[0, a)$, if the pair x_1, x_2 has not been asked before, or by giving the same reply as before, if it has. At the end of the computation we will have fixed the value of f at at most t places; we choose the remaining values at random in $[0, a)$. There were a^{a^2} possible sequences of random choices we could have made, and each one leads to a different f. Hence choosing f in this way gives the same distribution as choosing it uniformly at random.

The probability that the computation successfully found x_1, x_2, y_1, y_2 with $f(x_1, x_2) = f(y_1, y_2)$ is at most the probability that, if t elements of $[0, a)$ are chosen at random, two of them are the same. The probability that t such elements are all different is

$$1 \cdot \frac{a-1}{a} \cdot \frac{a-2}{a} \cdot \ldots \cdot \frac{a-(t-1)}{a} > \left(1 - \frac{t}{a}\right)^t > \frac{3}{4}$$

using lemma 5.4.

Hence we have shown that if we choose f uniformly at random the machine M_c will only find witnesses with probability $< 1/4$. This contradicts the choice of c. $\square$

Corollary 5.6 *The injective* WPHP *for a function given by an oracle f is not provable in $S^1_2(f) + \forall a \, \mathrm{PHP}^a_{a^2}(\mathrm{PV}(f))$.*

Corollary 5.7 *For any $\Sigma^b_1(f)$ formula $\chi_a(x,y)$ containing only a,x,y as free variables,*

$$S^1_2(f) \nvdash \forall a \, , \, \mathrm{PHP}^a_{a^2}(\chi_a(x,y)) \to \mathrm{PHP}^{a^2}_a(f).$$

Proof Apply the relativized version of lemma 3.10. $\square$

The key idea in the proof of lemma 5.5 is that we can find a class of structures, a large number of which will fool a given deterministic polynomial time machine. It follows that there is no probabilistic polynomial time machine which works on every structure in the class. This is easy in lemma 5.5, because the only condition on our structures is that they be binary functions.

In general this will not be quite so easy, since we will want to find a large class of structures that all satisfy a particular theory. We do this by taking one structure that satisfies the theory and permuting it.

Definition 5.8 *Suppose that $a \in \mathbb{N}$ has been fixed. Let $K = \langle [0,a), <, P, 0 \rangle$ be the structure on the set $\{0, \ldots, a-1\}$ with the usual ordering relation, a one place "modulo predecessor" function P taking $x+1$ to x and and 0 to $a-1$, and a constant symbol 0 for the least element. Let S be the set of all permutations of $[0,a)$. For $\sigma \in S$, let $K^\sigma = \langle [0,a), <^\sigma, P^\sigma, 0^\sigma \rangle$ be K permuted by σ. That is, $x <^\sigma y$ if and only if $\sigma(x) < \sigma(y)$, $P^\sigma(x) = y$ if and only if $P(\sigma(x)) = \sigma(y)$ and $0^\sigma = \sigma^{-1}(0)$.*

For $i \in \mathbb{N}$ we define N^σ_i, the i-neighbourhood of 0^σ, to be the set of elements x of $[0,a)$ from which we can reach 0^σ with i or fewer applications of the function P^σ. (So $x \in N^\sigma_i$ if and only if $\sigma(x) \in [0,i]$.)

Theorem 5.9 *There is no probabilistic polynomial time oracle machine which, for all $a \in \mathbb{N}$ and all permutations σ of $[0, a)$, given input a and equipped with oracles for $<^\sigma$ and P^σ, with probability at least $2/3$ outputs the least element 0^σ of the structure $\langle [0, a), <^\sigma, P^\sigma, 0^\sigma \rangle$.*

Proof Suppose there is such a machine M with time bound $|x|^k$. Choose $a \in \mathbb{N}$ bigger than $4|a|^{2k}$ and let $t = |a|^k$. For a sequence $c \in \{0,1\}^t$, we will write M_c for the deterministic machine (essentially a branching program) which simulates M running with coin tosses c. The only part of M_c we consider is a combined oracle query and oracle reply tape, which at step i of the computation will contain an element w_i of $[0, a)$. We allow M_c to do one of three things at each step $i + 1$:

1. Write down some number w_{i+1} on the tape;
2. Query $[w_{i-1} <^\sigma w_i?]$ and expect w_{i+1} to be 0 or 1 accordingly;
3. Query $[P^\sigma(w_i) =?]$ and expect w_{i+1} to be the correct answer.

Let C be the set $\{0,1\}^t$ of possible coin tosses and S the set of permutations of $[0, a)$. We are interested in the number of pairs $(\sigma, c) \in S \times C$ for which the machine $M_c(<^\sigma, P^\sigma)$ succeeds, that is, outputs $w_t = 0^\sigma$ on input a. Call a pair *bad* if at some step i in the computation of $M_c(<^\sigma, P^\sigma)$ at least one of the elements $w_1, \ldots, w_i$ is in N^σ_{t-i}, the $(t-i)$-neighbourhood of 0^σ. All other pairs are *good.*

Clearly the set of bad pairs contains all the pairs for which $M_c(<^\sigma, P^\sigma)$ succeeds. By assumption, for each σ the machine $M_c(<^\sigma, P^\sigma)$ is successful for $2/3$ of all coin tosses. Hence there are at least $a! \cdot \frac{2}{3} \cdot 2^t$ bad pairs.

We will obtain a contradiction by showing that for each $c \in C$, if we choose a permutation σ at random then (σ, c) is good with high probability.

Fix c and choose σ step-by-step as follows. Set $\sigma_0 = \varnothing$, and begin a computation of M_c. Suppose that after step i in the computation σ_i is a partial permutation of $[0, a)$, defined on $w_1, \ldots, w_i$ (this list may contain repetitions) and nowhere else. To avoid a technical problem, we assume that the program always puts $w_1 = 0$ and $w_2 = 1$. The definition of σ_{i+1} depends on the next action M_c takes.

1. If M_c writes down an element w_{i+1}, do nothing if w_{i+1} has already occurred on the list. If w_{i+1} is new, choose y at random in $[0,a)\setminus\mathrm{ran}(\sigma_i)$ and let $\sigma_{i+1} = \sigma_i \cup \{\langle w_{i+1}, y\rangle\}$.

2. If $[w_{i-1} <^\sigma w_i?]$ is queried set w_{i+1} to be 0 or 1, depending on whether $\sigma_i(w_{i-1}) < \sigma_i(w_i)$ or not. Let $\sigma_{i+1} = \sigma_i$.

3. If $[P^\sigma(w_i) =?]$ is queried, let $y = P(\sigma_i(w_i))$. If $y = \sigma_i(x)$ for some x, set $w_{i+1} = x$ and let $\sigma_{i+1} = \sigma_i$. Otherwise choose x at random in $[0,a) \setminus \mathrm{dom}(\sigma_i)$, set $w_{i+1} = x$ and let $\sigma_{i+1} = \sigma_i \cup \{\langle x, y\rangle\}$.

After t steps the computation will have finished; extend σ_t randomly to a total permutation σ.

For each c, σ chosen this way will be distributed uniformly in S, since there are $a!$ different sequences of random choices we could have used and each one leads to a different σ. We claim that if σ is so chosen then at the ith step in a computation the probability that the computation has been "good so far" is at least $(1-t/a)^i$. That is,

$$\mathrm{Prob}[\forall j \leqslant i\,,\, w_j \notin N^\sigma_{t-i}] \geqslant \left(1 - \frac{t}{a}\right)^i .$$

This holds for $i = 1$, since $\sigma_1(w_1)$ is chosen at random from $[0,a)$ and $w_1 \in N^\sigma_{t-1}$ if and only if $\sigma_1(w_1) \in [0,t]$. Now suppose this holds for i, and consider the different things that can happen at the $i+1$st step. Neither case 2 nor case 3 above can make a computation that has been good so far turn bad. In particular, if $[P^\sigma(w_i) =?]$ is queried and $w_i \notin N^\sigma_{t-i}$ then $w_{i+1} = P^\sigma(w_i) \notin N^\sigma_{t-i-1}$, by definition. If case 1 occurs, then the computation only turns bad if a new w_{i+1} is written down and $\sigma(w_{i+1})$, chosen at random, is in $[0, t-i-1]$. The probability that this does not happen is at least $(1-t/a)$.

Hence once we have finished the computation and defined σ, the probability that the pair (σ, c) is good is at least $(1-t/a)^t$.

We have shown that for each c, at least $a! \cdot (1-t/a)^t > \frac{3}{4}a!$ permutations σ yield good pairs. Hence there are at most $a! \cdot \frac{1}{4} \cdot 2^t$ bad pairs, a contradiction. $\square$

Corollary 5.10 *The theory $S^1_2(<') + \forall x\, \mathrm{PHP}^x_{x^2}(\mathrm{PV}(<'))$ does not prove that every total order $<'$ has a least element on every interval $[0,a)$.*

Proof Suppose

$$\forall a\,,\,(<' \text{ is a total order on } [0,a)) \to \exists x < a\, \forall y < a\, (x \neq y \to x <' y)$$

is provable in the theory. If we introduce a Herbrand function p to replace the universal quantifier $\forall y$, we have that

$$\begin{aligned}&\forall a\,,\,(<' \text{ is a total order on } [0,a))\\ &\qquad\qquad \to \exists x < a\, (p(x) < a \wedge x \neq p(x) \to x <' p(x))\end{aligned}$$

is provable in the theory. This is a $\Sigma_1^b(<', p)$ formula, so by theorem 3.11 there is a probabilistic polynomial time machine M which will, when equipped with oracles for $<'$ and p and given an input a, find witnesses with high probability.

However by theorem 5.9 we can find a number $a \in \mathbb{N}$ and a structure $\langle [0,a), <^\sigma, P^\sigma, 0^\sigma \rangle$ in which the element 0^σ is the only witness to the formula above and the structure is such that M does not output 0^σ if only given access to $<^\sigma$ and P^σ. $\square$

The same proof will work for any language containing only constants, relations and one unary function symbol. It ought to be possible to generalize the result and get something like the following:

Definition 5.11 *Suppose $a \in \mathbb{N}$, and $K = \langle [0,a), \bar{r}, \bar{f}, \bar{c} \rangle$ is a structure with a finite number of relations, functions and constants $\bar{r}, \bar{f}, \bar{c}$. Let S be the set of all permutations of $[0,a)$. For $\sigma \in S$ we define K^σ, the permutation of K by σ, to be the structure $\langle [0,a), \bar{r}^\sigma, \bar{f}^\sigma, \bar{c}^\sigma \rangle$. Here, for each relation r_i and tuple $\bar{x} \subseteq [0,a)$, $r_i^\sigma(\bar{x})$ if and only if $r_i(\sigma(\bar{x}))$. Similarly $f_i^\sigma(\bar{x}) = y$ if and only if $f_i(\sigma(\bar{x})) = \sigma(y)$, and $c_i^\sigma = \sigma^{-1}(c_i)$.*

If $\bar{x}, \bar{y}$ are tuples in $[0,a)$, and $t \in \mathbb{N}$, we say that $\bar{y}$ is derivable from $\bar{x}$ by a circuit in K of size t if there is a sequence $w_1, \ldots, w_t$ of elements of $[0,a)$ which contains every element in the tuple $\bar{y}$ and is such that every element of $\bar{w}$ is either an element of the tuple $\bar{x}$, or the interpretation of a constant in K, or follows from earlier elements in the sequence $\bar{w}$ by a function in K.

Open Problem 5.12 *Suppose $a \in \mathbb{N}$, and $K = \langle [0,a), \bar{r}, \bar{f}, \bar{c} \rangle$ is a structure as above. Let W be any set of tuples from $[0,a)$, and for $\sigma \in S$ let $W^\sigma =$*

$\{\bar{x} \subseteq [0,a) : \sigma(\bar{x}) \in W\}$. *Suppose for some* $t \in \mathbb{N}$ *and some probability* q *there is a probabilistic machine* M *which for any* $\sigma \in S$, *when run on input* K^σ *for time* t, *outputs a member of* W^σ *with probability at least* q. *Show that for at least* qa^t *of the* t*-tuples* $\bar{x}$ *in* $[0,a)$, *some member of* W *is derivable from* $\bar{x}$ *by a circuit in* K *of size* t.

6 Constructing unique end-extensions

We show that if any version of WPHP fails between a and a^2 in a model K of S^1_0 of the form $[0, a^{|a|})$ then for any $k \in \mathbb{N}$ there is an end-extension J of K to a model of S^1_0 of the form $[0, a^{|a|^k})$ definable inside K (theorem 6.6). Furthermore J is the unique such end-extension, up to isomorphism over K. A consequence is that in a model of S^1_2 in which WPHP fails, increasing the interval our quantifiers range over (up to a certain point) does not increase the complexity of the Σ^b_1 sets we can define (corollary 6.8). (In fact our results are slightly more general than this.)

6.1 Categoricity, definability and coding

Definition 6.1 (Gaifman; see [22], [23], [9]) *A structure M is relatively categorical over a definable subset P with respect to a theory T if $M \vDash T$ and for every $N \vDash T$, if there is an isomorphism between $M \restriction P$ and $N \restriction P$ then this can be extended to an isomorphism between M and N.*

We set out what it means for a structure to be defined inside a model of bounded arithmetic. Recall that R is a theory of arithmetic with a top without induction; see definition 2.25.

Definition 6.2 *If S and $X \subseteq S$ are sets in a model $K \vDash R$, then X is said to be $\bar{\Delta}^b_j$ in S if both X and $S \setminus X$ are definable by $\bar{\Sigma}^b_j$ formulas.*

Definition 6.3 *Let $K \vDash R$. We say that a structure J (in our language) is $\bar{\Sigma}^b_j$ defined in K if there exist a $\bar{\Sigma}^b_j$ subset S of K and relations $=_J$, $<_J$, $\cdot_J$, $+_J$ and $|\ |_J$ that are $\bar{\Delta}^b_j$ in S such that J consists of the domain $S/_{=_J}$ with the relations induced by $<_J$, $\cdot_J$, $+_J$ and $|\ |_J$. For $a \in K$, we say that J is defined below a if $S \subseteq [0, a)$.*

Definition 6.4 *If K is a model of R and J is $\bar{\Sigma}^b_j$ defined in K, a $\bar{\Sigma}^b_j$ function from K to J is a function of the form*

$$x \mapsto \{y \in S : \phi(x, y)\}$$

for a $\bar{\Sigma}^b_j$ formula ϕ, which maps elements of K to $=_J$-equivalence classes.

We show that if a model of S_0^j violates WPHP in the right way, then we can code very long sequences inside the model. We will go on in the next section to make these sequences into our end-extension by defining the operations of arithmetic on them.

Lemma 6.5 *Suppose $K \vDash S_0^j$ is of the form $[0, a^\varepsilon)$, where $a = 2^\alpha$ some α and K does not satisfy* $\mathrm{mPHP}_a^{a^2}(\bar{\Sigma}_j^b)$. *Then for any $l \in \mathbb{N}$ there is a $\bar{\Sigma}_j^b$ subset S of $[0, a)$ such that we can define a $\bar{\Sigma}_j^b$ coding relation $\langle x \rangle_i = y$, which defines a function from $S \times \varepsilon^l$ to $[0, a)$, but is not defined for x outside S. This can be used to code any $\bar{\Sigma}_j^b$ definable ε^l-length sequence of elements of $[0, a)$ as an element of S (possibly two elements of S will code the same sequence).*

Proof We will call elements of $[0, a)$ "numerals", and use the coding function x_i to treat any element of K as a sequence of ε numerals. Let $r(x)$ and $f(x, y)$ be the $\bar{\Sigma}_j^b$ formulas violating mPHP. We first use lemma 3.6 to amplify f to a $\bar{\Sigma}_j^b$ definable surjection g from $s([0, a))$ onto $[0, a^\varepsilon)$, where s is also a $\bar{\Sigma}_j^b$ formula. Furthermore, by the lemma if $r([0, a)) = [0, a)$ then $s([0, a)) = [0, a)$ and if f is 1-1 then g is.

Thus we can encode an ε-length sequence of numerals as a single element of $s([0, a))$. To encode ε^l-length sequences, we use a complete ε-ary tree of (standard) height l. We label the leaves of the tree with the sequence $\beta_1 \ldots \beta_{\varepsilon^l}$ which we want to encode, and then label the other nodes so that if a node is labelled w, then $w \in s$ and its children are labelled $g(w)_1 \ldots g(w)_\varepsilon$. We define $\langle x \rangle_i = y$ to hold if, in the tree with x at the root, the leaf at the end of the path naturally given by $i < \varepsilon^l$ (considered as an l-tuple in $\varepsilon \times \ldots \times \varepsilon$) is labelled y. Let $S(x)$ be the formula $x < a \wedge \forall 1 \leqslant i \leqslant \varepsilon^l \, \exists y < a \, \langle x \rangle_i = y$.

To show formally that this is a coding relation with the required property, let $\phi(i, y)$ be a $\bar{\Sigma}_j^b$ formula, possibly with parameters, such that $\forall 1 \leqslant i \leqslant \varepsilon^l \, \exists y < a \, \phi(i, y)$. We claim that we can find $x \in S$ encoding a sequence satisfying ϕ, ie. $\forall 1 \leqslant i \leqslant \varepsilon^l \, \phi(i, \langle x \rangle_i)$. This is done by considering, in turn, each level of the tree described in the previous paragraph and showing that each node on that level has a suitable label.

First look at the level immediately below the leaves. Let $\phi'(i, y)$ be the formula stating that y is a suitable label for the ith node on this level:

$$s(y) \wedge \forall 1 \leqslant k \leqslant \varepsilon \, \phi((i-1) \cdot \varepsilon + k, g(y)_k)$$

(where g is our surjection $s([0,a)) \twoheadrightarrow [0,a^\varepsilon))$. Since we can encode, using first comprehension and then g^{-1}, any $\bar{\Sigma}^b_j$ definable ε-length sequence of numerals as a single element of s, we can show that all these nodes can be labelled, ie. $\forall 1 \leqslant i \leqslant \varepsilon^{l-1}\, \exists y < a\, \phi'(i,y)$. The formula $\phi'(i,y)$ is still $\bar{\Sigma}^b_j$, so we can repeat this step $l-1$ times for all the lower levels in the tree to find $y \in S$ (there may be more than one such y) encoding a suitable ε^l-length sequence via $\langle y \rangle_i$. $\square$

6.2 The construction

Note that in the theorem below the restriction that a should be a power of 2 can easily be dispensed with, since given $K \vDash S^i_0$ of the form $[0,b)$ we can always construct a model of the form $[0,b^2)$ from the cartesian product $K \times K$, and this model will certainly be determined up to isomorphism over K. So we can find a suitable power of 2 when we need one.

We give this proof in a rather general form, but the most useful case is $j=1$, $k=0$, $\varepsilon = |a|$.

Theorem 6.6 *Suppose $j,k,l \in \mathbb{N}$, $j \geqslant 1$, $k \geqslant 0$, $l \geqslant 2$. Let K be a model of S^{j+k}_0 of the form $[0,a^\varepsilon)$, where $a = 2^\alpha$ for some α. Suppose $K \vDash \neg\mathrm{mPHP}^{a^2}_a(\bar{\Sigma}^b_j)$. Then*

1. *K has an end extension to a model J of S^{k+1}_0 of the form $[0,a^{\varepsilon^l})$. Furthermore this end extension is definable inside K below a in the sense of definition 6.3.*

2. *If I is any end-extension of K to a model of S^j_0 of the form $[0,a^{\varepsilon^l})$, then I is relatively categorical over K with respect to the theory $S^j_0 + (a^{\varepsilon^l}-1$ is the greatest element).*

Proof We will first construct J, and then show it is an end extension. Each element of J will be constructed in the natural way as a sequence of numerals of length ε^l. We use the coding function $\langle x \rangle_i$ and the set S given by lemma 6.5, and the fact that we can define addition and multiplication numeral-wise.

Define, for b, c in S,

$$\begin{aligned}
b =_J c \quad &\Leftrightarrow \quad \forall 1 \leqslant i \leqslant \varepsilon^l \, \langle b \rangle_i = \langle c \rangle_i; \\
b <_J c \quad &\Leftrightarrow \quad \exists 1 \leqslant i \leqslant \varepsilon^l \, (\langle b \rangle_i < \langle c \rangle_i \wedge \forall i < t \leqslant \varepsilon^l \langle b \rangle_t = \langle c \rangle_t); \\
(|b| = c)_J \quad &\Leftrightarrow \quad (b = 0 \wedge c = 0) \vee \exists 1 \leqslant i \leqslant \varepsilon^l, \\
&\qquad \langle b \rangle_i \neq 0 \wedge c = \alpha \cdot (i-1) + |\langle b \rangle_i| \wedge \forall i < t \leqslant \varepsilon^l \langle b \rangle_t = 0.
\end{aligned}$$

These are $\bar{\Delta}^b_j$ in S because we only apply $\langle x \rangle_i$ to members of S, on which it behaves like a function, so that we can negate the subformulas containing it without having to increase the quantifier complexity. For example, to write the definition of equality in J more fully, we have

$$\begin{aligned}
b =_J c \quad &\Leftrightarrow \quad S(b) \wedge S(c) \wedge \forall 1 \leqslant i \leqslant \varepsilon^l \, \exists y \; \langle b \rangle_i = y \wedge \langle c \rangle_i = y \\
b \neq_J c \quad &\Leftrightarrow \quad S(b) \wedge S(c) \wedge \exists 1 \leqslant i \leqslant \varepsilon^l \, \exists y \, \exists z \; \langle b \rangle_i = y \wedge \langle c \rangle_i = z \wedge y \neq z.
\end{aligned}$$

To define addition, first note that in S^1_0 we can define $\bar{\Sigma}^b_1$ *modulo addition* and *carry* functions $A(x, y, z)$ and $C(x, y, z)$ in K such that if $x, y, z < a$, then $A(x, y, z), C(x, y, z) < a$ and $x + y + z = a \cdot C(x, y, z) + A(x, y, z)$. We add our ε^l-length sequences numeral by numeral, and use a variable w to encode the sequence of numerals carried.

Define, for $c, d, e \in S$,

$$\begin{aligned}
(c + d = e)_J \Leftrightarrow \exists w \, , \; S(w) \wedge [&\langle w \rangle_1 = C(\langle c \rangle_1, \langle d \rangle_1, 0) \\
&\wedge \forall 1 < i \leqslant \varepsilon^l \langle w \rangle_i = C(\langle c \rangle_i, \langle d \rangle_i, \langle w \rangle_{i-1})] \\
\wedge [&\langle e \rangle_1 = A(\langle c \rangle_1, \langle d \rangle_1, 0) \wedge \langle w \rangle_\varepsilon = 0 \\
&\wedge \forall 1 < i \leqslant \varepsilon^l \langle e \rangle_i = A(\langle c \rangle_i, \langle d \rangle_i, \langle w \rangle_{i-1})].
\end{aligned}$$

We can always find a $w \in S$ witnessing the first expression in square brackets, so we can define the complement in S of this relation by putting a negation in front of the second expression in square brackets. Hence it is $\bar{\Delta}^b_j$ in S.

Multiplication is defined in a similar way. We need to be able to encode $(2\varepsilon^l + 1) \times (\varepsilon^l)^2$ matrices of numerals, and by lemma 6.5 there is a $\bar{\Sigma}^b_j$ subset T of $[0, a)$ on which we can do this via a $\bar{\Sigma}^b_j$ coding relation $(x)^i_k = y$, for $i = 1, \ldots, (\varepsilon^l)^2$ and $k = 1, \ldots, 2\varepsilon^l + 1$. For any $b, c \in S$, there are two such matrices π, ζ which encode the multiplication of b by c as follows (of course there may be more than one member of T coding each of these matrices). Each row of π contains the product of a numeral of b and a numeral of c, suitably offset. In particular if $\langle b\rangle_k \langle c\rangle_i = u + av$, for some $u, v < a$, then row $(i-1)\varepsilon^l + k$ of π has u in the $(i+k)$th place, v in the $(i+k+1)$st place, and zero everywhere else. Each row i of ζ encodes the sum of rows 1 to i of π. Define, for $d \in S$, $(b \cdot c = d)_J$ if for some π and ζ as above $\forall 1 \leqslant k \leqslant \varepsilon^l \, \langle d\rangle_k = (\zeta)_k^{(\varepsilon^l)^2}$ and all the other entries in row $(\varepsilon^l)^2$ of ζ are 0.

This completes the construction of the model $J = S/_{=_J}$. For $x \in S$ we will write $[x]$ for the equivalence class of x under $=_J$. To establish the BASIC$'$ axioms, we first observe that by $\bar{\Sigma}^b_j$−LIND in K, $<_J$ is a total ordering on J. Then consider the formula

$$\phi(x, y) \Leftrightarrow S(y) \wedge \forall 1 \leqslant i \leqslant \varepsilon \, \langle y\rangle_i = x_i \wedge \forall \varepsilon < i \leqslant \varepsilon^l \, \langle y\rangle_i = 0$$

and the map

$$\sigma : x \mapsto \{y \in S : K \vDash \phi(x, y)\}.$$

This is a map $K \to J$ and the relations on S have been defined so that it is an isomorphism onto an initial segment of J. Hence J is an end-extension of K.

Standard methods now show that $J \vDash$ BASIC$'$.

Let $\gamma = \alpha \cdot \varepsilon$ and $[\gamma'] = \sigma(\gamma)$, so that the sharply bounded quantifiers are precisely those that are bounded by some standard power of γ (in K) or some standard power of $[\gamma']$ (in J).

We claim that, for $n \geqslant 0$, for every $\bar{\Sigma}^b_{n+1}$ (or $\bar{\Pi}^b_{n+1}$) formula $\theta(\bar{x})$, there is a $\bar{\Sigma}^b_{n+j}$ (respectively $\bar{\Pi}^b_{n+j}$) formula $\theta_J(\bar{x})$ such that for all $\bar{b} \in S$,

$$J \vDash \theta([\bar{b}]) \Leftrightarrow K \vDash \theta_J(\bar{b}). \tag{i}$$

and if θ is sharply bounded then we can find both a $\bar{\Sigma}^b_j$ formula θ^{Σ}_J and a $\bar{\Pi}^b_j$ formula θ^{Π}_J satisfying (i).

We prove the sharply bounded case first, by induction on the number of quantifiers in θ. We know how to translate quantifier free formulas into formulas that are $\bar{\Delta}^b_j$ in S, and this is precisely the property that we require. Now suppose θ is of the form $\forall y < [\gamma']^m \, \chi(\bar{x}, y)$ for some $m \in \mathbb{N}$, where χ is sharply bounded. Then for any $\bar{b} \in S$, by the definition of the isomorphism σ,

$$
\begin{aligned}
J \vDash \theta([\bar{b}]) &\Leftrightarrow \forall i \in K,\, i < \gamma^m,\, J \vDash \chi([\bar{b}], \sigma(i)) \\
&\Leftrightarrow \forall i \in K,\, i < \gamma^m,\, \forall c \in \sigma(i), K \vDash \chi_J^{\Pi}(\bar{b}, c) \\
&\Leftrightarrow K \vDash \forall i < \gamma^m\, \forall x\, (S(x) \wedge \phi(i, x) \rightarrow \chi_J^{\Pi}(\bar{b}, x))
\end{aligned}
$$

where χ_J^{Π} is given by the induction hypothesis and is $\bar{\Pi}_j^b$. Since the equivalence class $\sigma(i)$ is never empty, this is in turn equivalent to

$$
K \vDash \forall i < \gamma^m\, \exists x\, (S(x) \wedge \phi(i, x) \wedge \chi_J^{\Sigma}(\bar{b}, x))
$$

where χ_J^{Σ} is given by the induction hypothesis and is $\bar{\Sigma}_j^b$. We deal similarly with sharply bounded existential quantifiers.

For the remaining cases, sharply bounded quantifiers are dealt with as above. If θ is of the form $\exists y\, \chi(\bar{x}, y)$, where χ is a $\bar{\Sigma}_{n+1}^b$ formula for which we have found a suitable $\bar{\Sigma}_{n+j}^b$ translation χ_J, we have

$$
\begin{aligned}
J \vDash \exists y\, \chi([\bar{b}], y) &\Leftrightarrow J \vDash \chi([\bar{b}], [c]) \quad \text{for some } c \in S \\
&\Leftrightarrow K \vDash \exists z\, (S(z) \wedge \chi_J(\bar{b}, z)).
\end{aligned}
$$

We treat $\bar{\Pi}_{n+1}^b$ formulas in a similar way.

To show that J satisfies $\bar{\Sigma}_{k+1}^b$–LIND, suppose θ is a $\bar{\Sigma}_{k+1}^b$ formula and

$$
J \vDash \theta(0) \wedge \forall x < [\gamma']^m\, (\theta(x) \rightarrow \theta(x+1)).
$$

Then for the corresponding $\bar{\Sigma}_{j+k}^b$ formula θ_J, writing $\theta_J(\phi(x))$ for $\exists y\, (S(y) \wedge \phi(x, y) \wedge \theta_J(y))$,

$$
K \vDash \theta_J(\phi(0)) \wedge \forall x < \gamma^m\, (\theta_J(\phi(x)) \rightarrow \theta_J(\phi(x+1))).
$$

So by $\bar{\Sigma}_{j+k}^b$–LIND in K, $K \vDash \theta_J(\phi(\gamma^m))$ and hence $J \vDash \theta([\gamma']^m)$.

Finally, suppose I is an end-extension of K to a model of S_0^j of the form $[0, a^{\varepsilon^l})$. Let J be the end-extension of K to a model of S_0^1 given by the construction above if we take $k = 0$. We claim that I is isomorphic to J. We can use our ordinary coding function to consider any $x \in I$ as a sequence $x_1 \ldots x_{\varepsilon^l}$ of numerals in $[0, a)$. For $x, y \in I$, $y \in S$, put

$$
\phi'(x, y) \Leftrightarrow \forall 1 \leqslant i \leqslant \varepsilon^l\, \langle y \rangle_i = x_i,
$$

meaning "y codes x". Just as in lemma 6.5, $\langle x \rangle_i$ can be used in I to encode any $\bar{\Sigma}^b_j$ definable ε^l-length sequence of numerals as an element of S, so in particular, every element of I is coded by (at least one) element of S. Conversely, by normal comprehension in I, every element of S codes an element of I. Also all the elements in an $=_J$ equivalence class in S must code the same element of I. Hence the map

$$\sigma' : x \mapsto \{y \in S : I \vDash \phi'(x, y)\}$$

is a bijection $\sigma' : I \leftrightarrow J$, and the definitions of the relations on S are set up precisely so that it is an isomorphism. $\square$

Corollary 6.7 *For $i \geqslant 1$, if $K \vDash S^i_0$ is of the form $[0, a\#a)$ and defines a $\bar{\Sigma}^b_1$ function violating either injective or surjective* WPHP *between a and a^2, then K has an end-extension to a model M of S^i_2 in which $\#a$ is cofinal. Furthermore this end-extension is unique, in that for any model N of S^i_2 with $N \restriction a\#a$ isomorphic to K, this isomorphism extends to an isomorphism $M \cong N \restriction \#a$.*

Corollary 6.8 *For $i \geqslant 1$, let M be a model of S^i_2 in which either injective or surjective* WPHP *fails between a and a^2 for a $\bar{\Sigma}^b_1$ function with parameters in $[0, a\#a)$. Then every Σ^b_i subset of $[0, a\#a)$ definable in M with parameters from $[0, a\#a)$ (this is Σ^b_i without the bar, so we are allowing $\#$ to appear as a function symbol in the bounds on quantifiers) is $\bar{\Sigma}^b_i$ definable inside $M \restriction a\#a$ (that is, with all quantifiers bounded by $a\#a$).*

Proof Use the translation in the proof of theorem 6.6. $\square$

Hence if the weak pigeonhole principle fails in an initial segment we cannot define more complex sets by increasing the range of our quantifiers; all the complexity in the structure is already contained inside that initial segment.

Another way of looking at this is that if the weak pigeonhole principle fails then we can do computations from the polynomial hierarchy in constant space. However this is at the expense of introducing more sharply bounded universal quantifiers, and so increasing the time taken (in some sense).

The version of corollary 6.8 for $\mathrm{I}\Delta_0$ leads naturally to a proof by diagonalization that the (parameter-free) Δ_0 hierarchy does not collapse in a model in which WPHP fails [20]. See [18] for a discussion of the extent to which the theory of an initial segment $M \restriction b$ of a model of true arithmetic is determined by $M \restriction a$, for $a < b$, under various assumptions about the collapse of the linear or polynomial time hierarchies.

7 Definable structures

We look for a converse to the "definable end-extension" part of theorem 6.6. We show that if a model J of arithmetic with a top is definable inside a model K of S^1_0, where K is of the form $[0, a^{|a|})$ and the domain of J is a subset of the interval $[0, a)$, then either J is definably isomorphic to an initial segment of K, or vice versa (theorem 7.4). If WPHP is true in K then it is the first of these that holds and the initial segment is unique; hence in models of $S^1_2 + \mathrm{WPHP}$ we can precisely count sets if they come with sufficient internal structure (corollary 7.8). However, a precise converse of this part of theorem 6.6 is impossible (see the remark after corollary 7.6).

7.1 Constructing an isomorphism

The proof is in two steps. Lemmas 7.1, 7.2 and 7.3 show that if an initial segment of J is isomorphic to an initial segment of K then we can extend that isomorphism to one with an exponentially larger domain. Theorem 7.4 then uses the extra room we have in K and the ability this gives us to code short sequences of elements of J to find an isomorphism with a small domain to start us off.

Lemma 7.1 *Let $K \vDash S^1_0$ be of the form $[0, a)$, and suppose $J \vDash R$ is $\bar{\Sigma}^b_1$ definable in K (see definition 6.3). Suppose further that for some $t \in S$ and some $\varepsilon < |a|$ there is a $\bar{\Sigma}^b_1$ isomorphism from $K \restriction \varepsilon$ onto $J \restriction |[t]_J|$. Then $J \restriction [t]_J \vDash S^1_0$.*

Proof We claim that for each $\bar{\Sigma}^b_1$ formula ϕ, there is a $\bar{\Sigma}^b_1$ formula ϕ_J such that for all $\bar{b} \in S$,

$$J \restriction [t]_J \vDash \phi([\bar{b}]_J) \Leftrightarrow K \vDash \phi_J(\bar{b}).$$

We prove this by induction on the quantifier complexity of ϕ. From the definition of $\bar{\Sigma}^b_1$-definability we know how to translate open formulas into formulas that are $\bar{\Delta}^b_1$ in S, which is precisely the property required, and we can translate $\exists x\, \theta(\bar{y}, x)$ as $\exists x\, (S(x) \wedge \theta_J(\bar{y}, x))$. Lastly, suppose ϕ is of the form $\forall i < |[t]_J|^n\, \theta(\bar{y}, i)$, for θ a $\bar{\Sigma}^b_1$ formula and $n \in \mathbb{N}$. We extend our $\bar{\Sigma}^b_1$

isomorphism $K \restriction \varepsilon \cong J \restriction |[t]_J|$ naturally to an isomorphism $K \restriction \varepsilon^n \cong J \restriction |[t]_J|^n$, given by a $\bar{\Sigma}^b_1$ formula χ say, and define

$$\phi_J(\bar{y}) \Leftrightarrow \forall i < \varepsilon^n \, \exists x \, (\chi(i, x) \wedge \theta_J(\bar{y}, x)).$$

To show that $\bar{\Sigma}^b_1$−LIND holds in $J \restriction [t]_J$, suppose ϕ is a $\bar{\Sigma}^b_1$ formula, $n \in \mathbb{N}$ and $|[t]_J|^n$ exists in J. Let χ be a $\bar{\Sigma}^b_1$ formula defining the isomorphism $K \restriction \varepsilon^n \cong J \restriction |[t]_J|^n$. We will write $\phi_J(\chi(i))$ as shorthand for $\exists x \, (\chi(i, x) \wedge \phi_J(x))$. Suppose

$$J \restriction [t]_J \vDash \phi(0) \wedge \forall i < |[t]_J|^n \, (\phi(i) \rightarrow \phi(i+1)).$$

Then

$$K \vDash \phi_J(\chi(0)) \wedge \forall i < \varepsilon^n \, (\phi_J(\chi(i)) \rightarrow \phi_J(\chi(i+1))).$$

Hence $K \vDash \phi_J(\chi(\varepsilon^n))$, so $J \restriction [t]_J \vDash \phi(|[t]_J|^n)$. $\square$

Lemma 7.2 *If $J \vDash S^1_0$ is $\bar{\Sigma}^b_1$ defined in $K \vDash S^1_0$, then the relations $(x = 2^i)_J$ and $(\mathrm{bit}(x, i) = 1)_J$ are $\bar{\Delta}^b_1$ in S.*

Proof The normal definitions of these relations do not use any sharply bounded universal quantifiers (which would not in general translate into sharply bounded quantifiers in K). So, if 0_J is a representative of the 0 element of J, we can put

$$\begin{aligned}
(x = 2^i)_J \quad &\Leftrightarrow \quad S(x) \wedge S(i) \wedge \exists y \, (S(y) \wedge (y = i+1)_J \wedge (|x| = y)_J) \\
&\qquad\qquad \wedge \exists z \, (S(z) \wedge (z + 1 = x)_J \wedge (|z| = i)_J); \\
(x \neq 2^i)_J \quad &\Leftrightarrow \quad S(x) \wedge S(i) \wedge \exists y \, ((y = 2^i)_J \wedge \neg y =_J x); \\
(\mathrm{bit}(x, i) = 1)_J \quad &\Leftrightarrow \quad S(x) \wedge S(i) \wedge 0_J <_J i \wedge \exists w \, \exists y \, \exists z \, (S(w) \wedge S(y) \wedge S(z) \\
&\qquad\qquad \wedge (w = 2^{i-1})_J \wedge z <_J w \wedge (x = y \cdot 2 \cdot w + w + z)_J); \\
(\mathrm{bit}(x, i) \neq 1)_J \quad &\Leftrightarrow \quad S(x) \wedge S(i) \wedge 0_J <_J i \wedge \exists w \, \exists y \, \exists z \, (S(w) \wedge S(y) \wedge S(z) \\
&\qquad\qquad \wedge (w = 2^{i-1})_J \wedge z <_J w \wedge (x = y \cdot 2 \cdot w + z)_J).
\end{aligned}$$

These relations will have the correct properties because $J \vDash S^1_0$. $\square$

Lemma 7.3 *Suppose $K \vDash S^1_0$ is of the form $[0,a)$, $J \vDash R$ is $\bar{\Sigma}^b_1$ defined in K and for some $n \in \mathbb{N}$ there is a $\bar{\Sigma}^b_1$ isomorphism between $K \restriction |a|^{(n)}$ and an initial segment of J (where $|\ |^{(n)}$ means a nesting of $|\ |$s that is n levels deep). Then there is a $\bar{\Sigma}^b_1$ isomorphism, either from all of K onto an initial segment of J, or from an initial segment of K onto all of J.*

Proof Let t be (a representative of) the $<_J$-greatest element of J. We will inductively construct $\bar{\Sigma}^b_1$ isomorphisms with domains $|a|^{(n-1)}, \ldots, |a|, a$ stopping if at any point we reach t and exhaust J.

For the inductive step, suppose that $\phi(x,y)$ is a $\bar{\Sigma}^b_1$ formula giving an isomorphism from $K \restriction |a|^{(m)}$ onto an initial segment of J.

Let $i < |a|^{(m)}$ be greatest such that

$$2^i < |a|^{(m-1)} \wedge \exists x \, \exists y \, , \, S(x) \wedge S(y) \wedge \phi(i,x) \wedge (|y| = x+1)_J$$

and let r be some such y. Then $J \restriction |[r]_J| \vDash S^1_0$ (since it is isomorphic to an initial segment of K) so by lemma 7.1, $J \restriction [r]_J \vDash S^1_0$. So we can choose r so that it is a power of 2 in J, and the equivalence class of r is the element of J corresponding to 2^i in K. Hence 2^i is the greatest power of 2 which exists in both $K \restriction |a|^{(m-1)}$ and J (in some sense).

Define $\theta(x,y)$ as

$$\begin{aligned} & x < 2^i \wedge y <_J r \wedge S(y) \wedge \\ & \quad \forall 1 \leqslant j \leqslant i \, \exists z \, , \, S(z) \wedge \phi(j,z) \wedge (\mathrm{bit}(x,j) = 1 \leftrightarrow (\mathrm{bit}(y,z) = 1)_J). \end{aligned}$$

We claim that $\sigma : x \mapsto \{y : \theta(x,y)\}$ is an isomorphism from $K \restriction 2^i$ onto $J \restriction [r]_J$.

To show well-definedness, suppose $y, y' <_J r$ with $y \neq_J y'$. Then, since $J \restriction [r]_J \vDash S^1_0$, without loss of generality we have $(\mathrm{bit}(y,v) = 1)_J$ and $(\mathrm{bit}(y',v) \neq 1)_J$ for some v such that $J \vDash 1 \leqslant [v]_J \leqslant |[r]_J|$. Since ϕ defines an isomorphism, $\phi(j,v)$ holds for some $1 \leqslant j \leqslant i$. Hence if for some x, x' we have $[y]_J = \sigma(x)$ and $[y']_J = \sigma(x')$, we must have $\mathrm{bit}(x,j) \neq \mathrm{bit}(x',j)$, so $x \neq x'$. We show that σ is injective in a similar way.

To show that σ is defined on all of $K\lceil 2^i$, let $x < 2^i$ and let $\chi(j)$ be the formula

$$\begin{aligned}&\exists y\,, S(y) \wedge \forall 1 \leqslant k \leqslant i\, \exists z\,,\ S(z) \wedge \phi(k,z)\\ &\qquad \wedge[k \leqslant j \to (\mathrm{bit}(x,k) = 1 \leftrightarrow (\mathrm{bit}(y,z) = 1)_J)]\\ &\qquad \wedge[j < k \to (\mathrm{bit}(y,z) \neq 1)_J]\end{aligned}$$

expressing that some $[y]_J$ is the correct image of x up to its jth bit and the remaining bits are 0. Then $\chi(0)$ holds, and if for any $j < i$ we have that $\chi(j)$ holds and is witnessed by y, we can find the element of J corresponding to 2^j and, depending on $\mathrm{bit}(x, j+1)$, let y' be either y or $(y + 2^j)_J$ (this sum exists in J and is not too big, because $J\lceil [r]_J \vDash S^1_0$). Then y' witnesses that $\chi(j+1)$ holds. Hence by $\bar{\Sigma}^b_1$−LIND in K, $\chi(i)$ holds and the set $\sigma(x)$ is not empty. Similarly we use comprehension in K to show that σ is a surjection.

Finally, since we can define all our relations bitwise, σ is an isomorphism.

To extend σ to the rest of $K\lceil |a|^{(m-1)}$, notice that by our choice of i either $2^i \geqslant |a|^{(m-1)}/2$ or $J \vDash [r]_J \geqslant [t]_J/2$. So we map $x \geqslant 2^i$ to the set

$$\{y \in S : \exists z\,, \phi(x - 2^i, z) \wedge (y = r + z)_J\}$$

if this is non-empty, which it will be until we reach the top element t of J. □

Theorem 7.4 *Suppose $K \vDash S^1_0$ is of the form $[0, b)$, and $a, a^\varepsilon \in K$ where $\varepsilon > |b|^{(n)}$ for some $n \in \mathbb{N}$. Suppose $J \vDash R$ is $\bar{\Sigma}^b_1$ defined in K below a. Then there is a $\bar{\Sigma}^b_1$ isomorphism, either from all of K onto an initial segment of J, or from an initial segment of K onto all of J.*

Proof Let $\theta(i, w)$ be the formula

$$\begin{aligned}&\forall j,k,l \leqslant i\,, (w_j + w_k = w_l)_J \leftrightarrow j + k = l\\ &\qquad \wedge (w_j \cdot w_k = w_l)_J \leftrightarrow j \cdot k = l\\ &\qquad \wedge w_j <_J w_k \leftrightarrow j < k\\ &\qquad \wedge (|w_j| = w_k)_J \leftrightarrow |j| = k\\ &\qquad \wedge w_0 =_J 0_J.\end{aligned}$$

Let $i < \varepsilon$ be greatest such that $\exists w \leqslant a^i\, \theta(i, w)$ and let $t = w_i$. We must have $i \geqslant 1$, since we can set $w_0 = 0_J$ and $w_1 = 1_J$. Let $\phi(x, y)$ be the formula

$S(y) \wedge y =_J w_x$. We claim that $\sigma : x \mapsto \{y : \phi(x, y)\}$ is an isomorphism from $K \restriction [0, i]$ onto $J \restriction [0, [t]_J]$.

It is sufficient to show that σ is surjective. Suppose it is not, and for some $s \in S$ we have $s <_J t$ and $\forall j \leqslant i \neg\phi(j, s)$. Let $j \leqslant i$ be greatest such that $w_j <_J s$. Then $j < i$ and $w_{j+1} \geqslant_J s$. But $J \vDash [w_{j+1}]_J = [w_j]_J + 1$, so $J \vDash [w_{j+1}]_J = [s_j]_J$ and hence $w_{j+1} =_J s$, a contradiction.

If $i < \varepsilon - 1$, then t must be the $<_J$ greatest element of J (otherwise we could add an extra element to w). Hence we have constructed an isomorphism from an initial segment of K onto all of J.

If $i = \varepsilon - 1$, then we have an isomorphism from $K \restriction |b|^{(n)}$ onto an initial segment of J and can use lemma 7.3. □

The next lemma has the consequence that if our defined structure J is isomorphic to an initial segment of K, then that initial segment is unique.

Lemma 7.5 *Suppose $K \vDash S^1_0$ and there is a $\bar{\Sigma}^b_1$ isomorphism σ between $K \restriction a$ and $K \restriction b$. Then σ is the identity function and in particular $a = b$.*

Proof First notice that σ must be the identity at least up to $|a|$, since otherwise there would be a least $i \leqslant |a|$ for which $\sigma(i) \neq i$, which is impossible. Now suppose $x < a$ and $\sigma(x) = y$. Then $|y| \leqslant |b| = \sigma(|a|) = |a|$, and for each $i < |a|$,

$$(K \restriction a \vDash \mathrm{bit}(x, i) = 1) \leftrightarrow (K \restriction b \vDash \mathrm{bit}(y, \sigma(i)) = 1)$$

since σ is an isomorphism. But $\sigma(i) = i$ for all such i. Hence $x = y$. □

7.2 Corollaries

Corollary 7.6 *Suppose $K \vDash S^1_0$ is of the form $[0, a^\varepsilon)$, for $\varepsilon = |a|^{(n)}$ some $n \in \mathbb{N}$, and that K is isomorphic to a structure J that is $\bar{\Sigma}^b_1$ defined in K below a. Suppose further that K is not isomorphic to any proper initial segment of K. Then there is a $\bar{\Sigma}^b_1$ formula ϕ giving an isomorphism $J \cong K$. Hence (multifunction)* WPHP *fails in K. In particular, if S (the set on which J is defined) is all of $K \restriction a$, then ϕ gives a surjection $a \twoheadrightarrow a^\varepsilon$; if each $=_J$ equivalence class has precisely one member, then ϕ gives an injection $a^\varepsilon \hookrightarrow a$.*

We cannot do without the condition "K is not isomorphic to any proper initial segment of K" because otherwise we have the following counterexample. Let M be any countable nonstandard model of PA. By Friedman's theorem, there exist $a, b \in M$ with $M \restriction a^{|a|} \cong M \restriction b^{|b|}$ and $a^{|a|} < b$. Hence the structure defined inside $M \restriction b$ on the set $[0, a^{|a|})$ by the normal relations is isomorphic to $M \restriction b^{|b|}$; but the weak pigeonhole principle does not fail in M.

Corollary 7.7 *If $K \vDash \mathrm{PA}^{\mathrm{top}}$ is of the form $[0,a)$ and is not isomorphic to any proper initial segment of itself, then for all $n \in \mathbb{N}$ no end-extension of K to a model of $\mathrm{PA}^{\mathrm{top}}$ of the form $[0, a^{|a|^{(n)}})$ is definable in K.*

Proof All the relevant results above go through if we use formulas of unrestricted quantifier complexity in the place of Σ_1^b formulas. Then use the fact that $\mathrm{I}\Delta_0 + (a^{|a|^{(n)}}\text{exists})$ proves $\mathrm{PHP}_a^{a^2}(\Delta_0)$. $\square$

We can interpret theorem 7.4 and lemma 7.5 as saying that a $\bar{\Sigma}_1^b$ set in a model of S_0^1 is either bigger than the model or has a unique precise size in the model, provided of course that this set comes with lots of structure and that we take counting statements to be about the existence of isomorphisms, rather than just bijections. In some ways this is a natural step, similar to moving from cardinal to ordinal numbers by adding an ordering relation.

We can use the weak pigeonhole principle to make sure that our definable structures do not get too big, and in particular to keep them inside a model of S_2^1. We summarise this as: in a model of S_2^1 satisfying WPHP we can precisely count structured sets. We make this precise below, using the injective WPHP. There are similar results for surjective or multifunction WPHP.

We will say that a Σ_1^b set S is *structured* if it is bounded and there are relations $<_S, |\;|_S, +_S, \cdot_S$ that are Δ_1^b in S such that $\langle S, <_S, |\;|_S, +_S, \cdot_S\rangle \vDash R$.

Corollary 7.8 *Let $M \vDash S_2^1 + \forall x\, \mathrm{PHP}_x^{x^2}(\Sigma_1^b)$ and suppose S is a structured Σ_1^b subset of M, with relations $<_S, |\;|_S, +_S, \cdot_S$. Then there exists a unique $b \in M$ for which there is a Σ_1^b function $f : \langle S, <_S, |\;|_S, +_S, \cdot_S\rangle \cong M \restriction b$.*

Proof Suppose S is bounded by a. Notice that we are using Σ_1^b sets here, where theorem 7.4 applies to $\bar{\Sigma}_1^b$ sets. However, since we are only interested in subsets of $[0, a)$ and the quantifiers in a Σ_1^b formula are bounded by terms, we can find b in M such that all of the sets considered are $\bar{\Sigma}_1^b$ definable inside $M \upharpoonright b$. Let c be greater than both b and $a^{|a|}$. Let $K = M \upharpoonright c$, and apply theorem 7.4. If there is a $\bar{\Sigma}_1^b$ isomorphism from S onto an initial segment of K, then we are done. If not, then there is a $\bar{\Sigma}_1^b$ isomorphism from K onto an initial segment of S, and hence there is a Σ_1^b injection $c \hookrightarrow a$, violating WPHP. $\Box$

These results (except for corollary 7.7, where we have to be very careful about the classes of formulas for which induction holds in our end-extension) also hold in the relativized case. For example, we have

Theorem 7.9 *Suppose $\langle K, \alpha\rangle \vDash S_0^1(\alpha)$ is of the form $[0, b)$, and $a, a^\varepsilon \in K$ where $\varepsilon > |b|^{(n)}$ for some $n \in \mathbb{N}$. Suppose $J \vDash R$ is $\bar{\Sigma}_1^b(\alpha)$ defined in $\langle K, \alpha\rangle$ below a. Then there is a $\bar{\Sigma}_1^b(\alpha)$ isomorphism, either from all of K (without the α) onto an initial segment of J, or from an initial segment of K onto all of J.*

Corollary 7.10 *Let α be a set $\{+^*, \cdot^*, <^*, |\ |^*, 0^*, 1^*, 2^*\}$ of relation and constant symbols of the same form as but disjoint from our normal language of arithmetic. Let R^* and S_0^{1*} be our normal theories re-written in this language. Then "every finite model of R is a model of S_0^1", or*

$$\forall a\, ,(\langle [0, a), \alpha\rangle \vDash R^*) \to (\langle [0, a), \alpha\rangle \vDash S_0^{1*}),$$

is provable in $S_2^1(\alpha) + \forall a\, \mathrm{PHP}_a^{a^2}(\Sigma_1^b(\alpha))$ but not in $S_2^2(\alpha)$.

Proof The independence from $S_2^2(\alpha)$ follows from theorem 5.3 since there is an infinite model of R that is not a model of S_0^1.

Now suppose that in a model of $S_2^1(\alpha) + \forall a\, \mathrm{PHP}_a^{a^2}(\Sigma_1^b(\alpha))$ the structure $J = \langle [0, a), \alpha\rangle$ is a model of R^*. Then by the relativized version of corollary 7.6 J is definably isomorphic to an initial segment of M (in the normal language, without α). Hence J is a model of S_0^{1*}. $\Box$

This highlights one reason why it is difficult to find relativized independence results for theories as strong as or stronger than $S_2^3(\alpha)$. $S_2^3(\alpha)$ proves $\mathrm{WPHP}(\Sigma_1^b(\alpha))$, so there is a class of sentences in the language of α (namely those of the form $\alpha \vDash R \to \Phi(\alpha))$ that are as hard to prove independent of $S_2^3(\alpha)$ as their unrelativized versions are of S_2^3, since if such a sentence is false in the structure given by α it will also be false in an initial segment of a model without α.

Open Problem 7.11 *Find the weakest theory that will work in the place of R to give the results of this chapter. For example, is it sufficient if our structure J is only assumed to be a model of a universally axiomatized theory of "discretely ordered abelian groups with a greatest element" in the language $\{0, 1, e, <, +, -, \lfloor \frac{x}{2} \rfloor, parity\}$ (with some sort of modulo addition)? (In this case multiplication should be definable, in the model $K \vDash S_2^1$ in which J is defined, in terms of repeated doubling.)*

The techniques from chapter 5 may be of some help in proving a lower bound, for the relativized case.

8 Generalizing WPHP

We consider the consequences for a structure of the presence or absence of a definable surjection from a subset P onto the whole structure. This is a generalized version of the surjective WPHP. We use some tools from abstract model theory, but most of the interesting applications are to models of $\mathrm{PA}^{\mathrm{top}}$ and hence, indirectly, to $\mathrm{I}\Delta_0$. In the first section we look for converses to the "unique end-extension" part of theorem 6.6. This works for models of $\mathrm{PA}^{\mathrm{top}}$ (corollary 8.3) and we obtain a partial converse for models of S^1_0 (corollary 8.7). In the second section we characterize WPHP in terms of the possible cardinalities of initial segments of a model (corollary 8.13) and construct an uncountable model of S_2 in which the polynomial size sets are precisely the countable sets (corollary 8.14).

8.1 Categoricity

Lemma 8.1 *Let M be a recursively saturated structure with a definable subset P containing at least two elements 0, 1 which are named in the language. For any formula $\phi(\bar{y})$, if for all $k \in \mathbb{N}$ there is no parameter free definable surjection from P^k onto $\phi(M) \cup \{0\}$, then there is $\bar{c} \in \phi(M)$ such that $\bar{c}$ is not definable with parameters from P. The converse also holds.*

Proof Recursively enumerate all parameter free formulas as $\psi_1(\bar{x}, \bar{y})$, $\psi_2(\bar{x}, \bar{y}), \ldots$ where we assume $\bar{x}$ has arity at most i in ψ_i. Let $\Gamma(\bar{y})$ be the type

$$\{\phi(\bar{y})\} \cup \{\bigwedge_{i \leqslant m} \forall \bar{x} \subseteq P \text{ "}\bar{y} \text{ is not unique such that } \psi_i(\bar{x}, \bar{y})\text{"} : m \in \mathbb{N}\}$$

Suppose Γ is not finitely satisfiable in M. Then there is a finite sequence of formulas $\psi_1, \ldots, \psi_m$ (where we now assume that each $\bar{x}$ has arity m) such that for each $\bar{c}$ satisfying ϕ, there exist $i \leqslant m$ and $\bar{d} \subseteq P$ for which $\bar{c}$ is unique such that $\psi_i(\bar{d}, \bar{c})$.

Define a surjection $f : P^{2m} \to \phi(M) \cup \{0\}$ as follows: given $(a_1, \ldots, a_m, d_1, \ldots, d_m) \in P^{2m}$, if for some $i \leqslant m$, $a_1 = \ldots = a_i = 0$, $a_{i+1} = \ldots = a_m = 1$ and $M \vDash \exists! \bar{y}\, \phi(\bar{y}) \wedge \psi_i(\bar{d}, \bar{y})$ then map $(\bar{a}, \bar{d})$ to that unique $\bar{y}$; otherwise map it to 0. This contradicts the assumption that there is no such surjection.

Hence Γ, since it is recursive, is realized in M. Clearly any element realizing it is not definable from P.

The converse direction is trivial. $\square$

Corollary 8.2 *Let M be a recursively saturated model of* $\mathrm{PA}^{\mathrm{top}}$ *of the form* $[0, b)$ *and let $a \in M$ be such that $a^k < b$ for all $k \in \mathbb{N}$. Suppose $M \models$* $\mathrm{PHP}^{a^k}_b(\Delta_0)$*, for every $k \in \mathbb{N}$. Then M is not relatively categorical over* $[0, a)$ *with respect to $Th(M)$.*

Proof Let $K = K([0, a); M)$, the definable closure of $M \restriction a$ in M, which is elementarily equivalent to M but omits the type "y is not definable from $[0, a)$". This type is realized in M, by lemma 8.1. So M and K are both end-extensions of $M \restriction a$, but are not isomorphic. $\square$

This is how the pigeonhole principle is typically used in the model theory of arithmetic, see for example [10] or chapter IV of [8]. We can now write down a characterization of the provability of WPHP in $\mathrm{I}\Delta_0$:

Corollary 8.3 $\mathrm{I}\Delta_0 \vdash \forall x\, \mathrm{PHP}^{x}_{x^2}(\Delta_0)$ *if and only if for every recursively saturated model M of* $\mathrm{PA}^{\mathrm{top}}$ *of the form* $[0, b)$ *and every $a \in M$ with $a^{\mathbb{N}} < b$, there is more than one end-extension of $M \restriction a$ to a model of* PA *of the form* $[0, b)$ *(we assume b is definable from parameters in* $[0, a)$*).*

Hence for $\mathrm{I}\Delta_0$ we have a neat model-theoretic characterization of WPHP, in terms of relative categoricity.

With the ultimate goal of extending this to weaker theories and finding a converse of the part of theorem 6.6 that showed that failure of WPHP in a model of S^1_2 implies relative categoricity, we give a proof of Gaifman's coordinatization theorem, that (assuming WPHP) to get two different models with the same restriction to P, we do not need definable Skolem functions, but only rigidity over P. In normal arithmetical situations we will always have this, in a very strong sense.

Lemma 8.4 *Suppose in $K \models$* BASIC' *we can define a parameter-free function* $\mathrm{bit}(x, i)$ *and prove that no two numbers in K encode the same sequence of bits. Then for any $b \in K$, no two elements of K smaller than b share the* *same type over $[0,|b|)$. Hence $K\restriction b$ is rigid over $K\restriction|b|$, and by repeating the argument $K\restriction b$ is rigid over $K\restriction|\ldots|b|\ldots|$, for any nesting of $|\ |$s (see Kaye [10]).*

We prove a simple version of the coordinatization theorem that makes direct use of this property that an element is uniquely given by its type.

Lemma 8.5 *Suppose M is a structure with a definable subset P such that no two elements of M have the same type over P. Then the principal types over P are realized in M by precisely the elements of M that are definable from P.*

Proof Suppose $p(x) = tp_M(c;P)$ has a principal formula $\phi(x)$ with parameters from P. Then we must have $\exists! x\, \phi(x)$, or two elements of M would have the same type over P. Hence c is definable from P. $\square$

Theorem 8.6 (Gaifman [9]) *Suppose M is a countable recursively saturated structure in a language with no function symbols and with a definable subset P which contains all the elements named by constants. Suppose that P contains at least two elements 0, 1 named in the language, that there is no parameter-free definable surjection from any standard power of P onto M and that no two elements of M have the same type over P. Then there is $N \equiv M$ such that $N\restriction P \cong M\restriction P$ but this isomorphism cannot be extended to an isomorphism $N \cong M$.*

Proof List the elements of $P(M)$ as $\bar{r}$. By lemma 8.1 there is $c \in M$ not definable from $\bar{r}$, and by lemma 8.5 the type $p(x) = tp_M(c;\bar{r})$ is not principal. The type

$$q(y) = \{P(y)\} \cup \{y \neq r : r \in \bar{r}\}$$

is not principal either, since it is not realized in M. Hence there is a structure $(N,\bar{r}) \equiv (M,\bar{r})$ omitting both p and q. Since $(N,\bar{r})$ omits q, $N\restriction P$ is isomorphic to $M\restriction P$. Since $(N,\bar{r})$ omits p, this isomorphism cannot be extended to an isomorphism $(N,\bar{r}) \cong (M,\bar{r})$. $\square$

Corollary 8.7 *Let M be a countable recursively saturated model of S^1_0 of the form $[0,b)$, and let $a \in M$ be such that $|b| < a$, $a^{\mathbb{N}} < b$ and M is relatively categorical over $[0,a)$ with respect to the complete theory of M. Then $M \vDash \neg \mathrm{PHP}^{a^k}_b(f)$ for some $k \in \mathbb{N}$ and some definable function f.*

This is not a very good converse to theorem 6.6, since it uses the complete theory of M and we cannot limit the quantifier complexity of f. The ideal result would be something like: relative categoricity with respect to S^1_0 implies failure of surjective WPHP for a Σ^b_1 function. However it is not clear whether this is attainable. It would mean that in S^1_2 surjective WPHP(Σ^b_1) implies injective WPHP(Σ^b_1), and we showed in chapter 3 that if this were true for PV function symbols (rather than for Σ^b_1 formulas) then we could crack RSA.

8.2 Cardinality

There are many combinatorial principles in arithmetic which are normally proved by counting arguments, but which turn out only to need approximate rather than precise counting. There have been some successes in proving these in S_2 using the weak pigeonhole principle, which could be taken to say that, as far as definable functions are concerned, n^2 is bigger than n. See for example Pudlák's proof of the Ramsey theorem [24] or the proof that there are infinitely many primes [21]. It would be nice to be able to characterize the approximate counting available in bounded arithmetic and to give a uniform way of dealing with combinatorial proofs that make use of it.

If there is no definable map from a onto b, one would sometimes like to say that the "definable cardinality" of b is bigger than that of a; Krajíček has suggested developing this idea into a theory of the definable combinatorics of a structure [14], [16].

We present a simple application of Vaught's two-cardinal theorem (see [9]) to give a result in this direction, that in a countable, recursively saturated model of a theory with Skolem functions (such as PA), we can choose any definable set P and extend the model to one in which P is unchanged, hence still countable, but every other definable set has greater cardinality than P if and only if it has greater definable cardinality than any standard power of

P. This may be of some use in formalizing approximate counting arguments in S_2, and leads to an interesting characterization of the polynomial size sets in models of S_2.

In our setting we can sharpen the two-cardinal theorem slightly, using resplendence. First, however, we will use the normal version to prove a second model-theoretic statement equivalent to the provability of WPHP in $\mathrm{I}\Delta_0$.

Theorem 8.8 $\mathrm{I}\Delta_0(\alpha) \vdash \forall x\, \mathrm{PHP}^{x}_{x^2}(\Delta_0(\alpha))$ *if and only if for every countable model K of $\mathrm{PA}^{\mathrm{top}}(\alpha)$ of the form $[0, b)$ and every $a \in K$ for which $a^{\mathbb{N}} < b$ there is some uncountable $J \succeq K$ in which $J \restriction a$ is countable.*

Proof For the forwards implication, extend K to a recursively saturated structure K', and let I be the definable closure of $K' \restriction a$ in K'. Then as in the previous section, $I \preceq K'$, $I \restriction a = K' \restriction a$ but by WPHP, $I \subset K'$ so we can apply the two-cardinal theorem to get J. For the other direction, if there is a definable surjection $a \twoheadrightarrow a^2$ then we can amplify it as in the proof of lemma 3.6 to a surjection $a \twoheadrightarrow b$. So J and $J \restriction a$ must have the same cardinality. $\square$

Lemma 8.9 (Resplendence [11]) *Suppose M is a countable recursively saturated L-structure in a recursive language L, the language L' is a recursive extension of L and T is a recursively axiomatized L'-theory. Then, if $Th(M) + T$ is consistent, there is an expansion of M to L' satisfying T.*

Lemma 8.10 *Let L be a recursive language, $\phi(x)$, $\psi(x)$ be parameter free L-formulas and M, N be countable L-structures such that M is recursively saturated, $N \preceq M$, $\phi(N) = \phi(M)$ and $\psi(N) \subset \psi(M)$. Then there is $M' \succeq M$ such that $\phi(M) = \phi(M')$, $\psi(M) \subset \psi(M')$ and $M \cong M'$.*

Proof Let $L^+ = L \cup \{H, f\}$ where f is a one-place function and H is a one-place predicate. Writing χ^H for the relativization of χ to H, let T be the following set of sentences:

1. H is the range of f;
2. $\forall \bar{x} \subseteq H(\chi^H(\bar{x}) \leftrightarrow \chi(\bar{x}))$ for each L-formula χ;

3. $\forall \bar{x}\,(\theta(\bar{x}) \leftrightarrow \theta(f(\bar{x})))$ for each atomic L-formula θ;

4. $\forall x\,(\phi(x) \rightarrow H(x))$;

5. $\exists x\,(\psi(x) \wedge \neg H(x))$.

By the proof of Vaught's two-cardinal theorem (in [9]), there are structures U, V with $M \preceq V$ such that $U \preceq V$, $\phi(U) = \phi(V)$, $\psi(U) \subset \psi(V)$ and there is an isomorphism $V \cong U$. So we may expand V to an L^+ structure satisfying T by interpreting H as membership of U and f as the isomorphism $V \cong U$.

T is a recursive theory, and we have shown that $Th(M) \cup T$ is consistent. Hence by lemma 8.9 we can expand M to an L^+ structure satisfying T.

So M is isomorphic to an elementary submodel M^- of itself, with $\phi(M^-) = \phi(M)$ and $\psi(M^-) \subset \psi(M)$. By identifying M with M^-, we can find an elementary extension M' of M with the properties required. $\square$

Lemma 8.11 *The union of a countable elementary chain $\{M_\gamma : \gamma < \delta\}$ of countable, recursively saturated structures isomorphic to M_0 is a countable, recursively saturated structure isomorphic to M_0.*

Proof The union is recursively saturated and realizes the same types as M_0. $\square$

Theorem 8.12 *Let M be a countable, recursively saturated structure with definable Skolem functions in a recursive language. Let P be a definable subset of M containing at least two elements. Then we can find $N \succeq M$ such that $P(N) = P(M)$ but the countable definable subsets $\phi(N)$ of N (with parameters from N) are precisely those for which there is a definable surjection (with parameters from N) from some standard power of $P(N)$ onto $\phi(N)$. Every other definable subset is uncountable.*

Proof By the existence of Skolem functions there are two definable elements of P; add names 0, 1 to the language for these. We will construct an elementary chain $\{M_\beta : \beta < \omega_1\}$, with $M_0 = M$, of pairwise isomorphic structures such that for all $\beta < \omega_1$, $P(M_\beta) = P(M_0)$ and for any formula $\phi(x)$ with parameters from M_β, if there is no definable surjection in M_β from any standard power of P onto $\phi(M_\beta)$ then $\phi(M_{\beta+1}) \supset \phi(M_\beta)$. By lemma 8.11 we can put $M_\delta = \bigcup_{\beta<\delta} M_\beta$ for δ a limit.

For the successor step, enumerate as $\phi_1(x), \phi_2(x), \ldots$ the formulas with parameters from M_β for which there is no surjection (with parameters from M_β) from any standard power of $P(M_\beta)$ onto $\phi_i(M_\beta)$, and for which $\phi_i(M_\beta)$ is non-empty. Let $\bar{m}_i \subseteq M_\beta$ be the tuple of parameters appearing in ϕ_i. Writing P for $P(M_\beta) = P(M_0)$, let $K_1 := K(P \cup \bar{m}_1; M_\beta)$ be the definable closure of $P \cup \bar{m}_1$ in M_β. There is no surjection with parameter $\bar{m}_1$ from any standard power of P onto $\phi_1(M_\beta)$, so there is certainly no surjection onto $\phi_1(M_\beta) \cup \{0\}$. Thus, temporarily adding $\bar{m}_1$ to the language, by lemma 8.1 there is $c \in \phi_1(M_\beta)$ not definable from P with parameter $\bar{m}_1$; so $c \notin K_1$.

Now $K_1 \preceq M_\beta$, $P(K_1) = P(M_\beta)$ and $\phi_1(M_\beta) \supset \phi_1(K_1)$ so by lemma 8.10 there is $M_\beta^1 \succeq M_\beta$ with $M_\beta^1 \cong M_\beta$, $P(M_\beta^1) = P(M_\beta)$ and $\phi_1(M_\beta^1) \supset \phi_1(M_\beta)$. Similarly, if we let $K_2 = K(P \cup \bar{m}_2; M_\beta)$ then $\phi_2(M_\beta^1) \supset \phi_2(K_2)$ so we can find $M_\beta^2 \succeq M_\beta^1$ with $M_\beta^2 \cong M_\beta^1$, $P(M_\beta^2) = P(M_\beta)$ and $\phi_2(M_\beta^2) \supset \phi_2(M_\beta^1)$. Repeating this step for $\phi_3, \phi_4, \ldots$ gives an elementary chain $M_\beta \preceq M_\beta^1 \preceq M_\beta^2 \preceq \ldots$ and taking the union of the chain gives us, by lemma 8.11, $M_{\beta+1} \cong M_\beta$ with the properties required.

Let $N = \bigcup_{\beta<\omega_1} M_\beta$. Suppose $\phi(x)$ is a formula with parameters $\bar{n} \subseteq N$ such that there is no surjection definable with parameters from N from any standard power of P onto $\phi(N)$. Suppose $\bar{n} \subseteq M_\beta$ for some $\beta < \omega_1$. Then for each $\beta \leqslant \gamma < \omega_1$, there is no surjection with parameters from M_γ from any standard power of P onto $\phi(M_\gamma)$, by elementariness. So by construction $\phi(M_{\gamma+1}) \supset \phi(M_\gamma)$. Hence $\phi(N)$ is uncountable.

Conversely, if there is a surjection from P^k onto $\phi(N)$, for $k \in \mathbb{N}$, then $\phi(N)$ must be countable because P^k is. $\square$

Corollary 8.13 *If K is a countable, recursively saturated model of* $\mathrm{PA}^{\mathrm{top}}(\alpha)$ *of the form $[0, b)$ containing an element a such that $K \models \mathrm{PHP}_b^{a^k}(\Delta_0(\alpha))$ for every $k \in \mathbb{N}$, then there is $J \succeq K$ with $J \restriction a = K \restriction a$ but with $J \restriction c$ uncountable for every $c > a^{\mathbb{N}}$.*

There are similar, rather stronger results for full Peano arithmetic in Paris and Mills [19], but these make heavy use of precise counting.

One cannot in general repeat this increase in size more than once. For suppose we have a countable recursively saturated structure $K \vDash \mathrm{PA}^{\mathrm{top}} + \forall x\, \mathrm{PHP}^{x}_{x^2}(\Delta_0)$ of the form $[0,b)$, with $|b|^{\mathbb{N}} < b$. We can find $J \succeq K$ of cardinality $\aleph_1$ with $J \restriction |b| = K \restriction |b|$. If we could go on to find an elementary extension I of J with cardinality $\aleph_2$ and with $I \restriction |b| = K \restriction |b|$, then this would imply a violation of the continuum hypothesis, since the function that takes an element of I to its set of non-zero bits is an injection from $[0,b)$ into the power set of $[0,|b|)$.

So if we could find a way of adding a predicate α to a model of $\mathrm{PA}^{\mathrm{top}}$ which ensured either that we could not increase the cardinality of part of the model in this way, or that, whenever we could increase the cardinality, we could do so more than once, we would have gone some way towards showing that $\mathrm{WPHP}(\alpha)$ is independent of $\mathrm{I}\Delta_0(\alpha)$.

We give one more application of the two-cardinal theorem.

Corollary 8.14 *Suppose M is a countable model of S_2, $a \in M$ and $\#a$ is cofinal in M. Then there exists an uncountable $N \succeq_{\Pi_1} M$ in which the coded sets are precisely the countable bounded Δ_0 sets.*

Proof Let M' be a recursively saturated extension of M, so $b > \#a$ for some $b \in M'$. Let $B = M' \restriction b$, so $B \vDash \mathrm{PA}^{\mathrm{top}}$ and B is recursively saturated. Let $C \succeq B$ be given by theorem 8.12, taking P to be the definable set "$x < |a|$". Let $N = C \restriction \#a$.

Note that for each $k \in \mathbb{N}$, $M \restriction 2^{|a|^k} \preceq N \restriction 2^{|a|^k}$. Hence $M \preceq_{\Delta_0} N$ and if $N \vDash \exists x\, \theta(\bar{m}, x)$ for θ a Δ_0 formula and $\bar{m} \subseteq M$, we must have $N \vDash \exists x < 2^{|a|^k}\, \theta(\bar{m}, x)$ for some $k \in \mathbb{N}$; so $M \vDash \exists x < 2^{|a|^k}\, \theta(\bar{m}, x)$. This shows that $M \preceq_{\Pi_1} N$.

Now suppose S is a subset of N coded as a sequence $(\sigma)_1, \ldots, (\sigma)_l$ for some $\sigma \in N$. Then $l < |a|^k$ for some $k \in \mathbb{N}$, and $N \restriction |a|^k$ is countable, so S must be countable.

Conversely, suppose that $S \subseteq N \restriction 2^{|a|^k}$ is countable and is defined by a Δ_0 formula $\phi(x)$. Then S is also definable by ϕ in C, so by the construction of C there exist $l \in \mathbb{N}$ and a definable function f such that f is a surjection from $|a|^l$ onto S.

Since S is bounded by $2^{|a|^k}$, we have that $C \vDash \forall i < |a|^l\, f(i) < 2^{|a|^k}$. So by comprehension in C, there is some $\sigma < 2^{|a|^{k+l}}$ in C with $C \vDash \forall i < |a|^l\, f(i) = (\sigma)_i$. Thus $C \vDash \forall x < 2^{|a|^k},\ \phi(x) \leftrightarrow \exists i < |a|^l\, (\sigma)_i = x$. This is a Δ_0 formula, so is also true in N. Hence S is coded in N, by σ. $\square$

References

[1] J. Avigad. Saturated models of universal theories. To appear in Annals of Pure and Applied Logic, 2001.

[2] D. Bovet, P. Crescenzi, and R. Silvestri. A uniform approach to define complexity classes. *Theoretical Computer Science*, 104:263–283, 1992.

[3] S. Buss. *Bounded Arithmetic.* Bibliopolis, 1986.

[4] S. Buss. Axiomatizations and conservation results for fragments of bounded arithmetic. pages 57–84, 1987.

[5] S. Buss. Relating the bounded arithmetic and polynomial time hierarchies. *Annals of Pure and Applied Logic*, 75(1–2):67–77, 1995.

[6] S. Cook. Feasibly constructive proofs and the propositional calculus. *Proceedings of the 7th Annual ACM Symposium on Theory of computing*, pages 83–97, 1975.

[7] O. Goldreich. *Foundations of Cryptography (Fragments of a Book).* Unpublished, 1995.

[8] P. Hájek and P. Pudlák. *The Metamathematics of First Order Arithmetic.* Springer, 1993.

[9] W. Hodges. *Model Theory.* Cambridge University Press, 1993.

[10] R. Kaye. A galois correspondence for countable recursively saturated models of peano arithmetic. In R. Kaye and D. Macpherson, editors, *Automorphisms of First-Order Structures*, pages 293–312. Clarendon Press, 1994.

[11] R. Kaye and D. Macpherson. Recursive saturation. In R. Kaye and D. Macpherson, editors, *Automorphisms of First-Order Structures*, pages 243–256. Clarendon Press, 1994.

[12] J. Köbler and J. Messner. Complete problems for promise classes by optimal proof systems for test sets. In *Proceedings of the 13th IEEE Conference on Computational Complexity*, pages 132–140, 1998.

[13] J. Krajíček. *Bounded Arithmetic, Propositional Logic and Computational Complexity*. Cambridge University Press, 1995.

[14] J. Krajíček. Uniform families of polynomial equations over a finite field and structures admitting an Euler characteristic of definable sets. *Proceedings of the London Mathematical Society*, 81(2):257–284, 2000.

[15] J. Krajíček and P. Pudlák. Some consequences of cryptographical conjectures for S_2^1 and EF. *Information and Computation*, 140(1):82–89, 1998.

[16] J. Krajíček and T. Scanlon. Combinatorics with definable sets: Euler characteristics and Grothendieck rings. *Bulletin of Symbolic Logic*, 3(3):311–330, 2000.

[17] A. Maciel, T. Pitassi, and A. Woods. A new proof of the weak pigeonhole principle. In *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing*, pages 368–377, 2000.

[18] J. Paris and C. Dimitracopoulos. Truth definitions for Δ_0 formulae. In *Logic and Algorithmic*, number 30 in Monographies de l'Enseignement Mathématique, pages 317–329. Université de Genève, 1982.

[19] J. Paris and G. Mills. Closure properties of countable non-standard integers. *Fundamenta Mathematica*, 103:205–215, 1979.

[20] J. Paris and A. Wilkie. Counting problems in bounded arithmetic. In *Methods in Mathematical Logic*, number 1130 in Lecture Notes in Mathematics, pages 317–340. Springer, 1985.

[21] J. Paris, A. Wilkie, and A. Woods. Provability of the pigeonhole principle and the existence of infinitely many primes. *Journal of Symbolic Logic*, 53(4):1235–1244, 1988.

[22] A. Pillay. $\aleph_0$-categoricity over a predicate. *Notre Dame Journal of Formal Logic*, 24(4):527–536, 1983.

[23] A. Pillay and S. Shelah. Classification theory over a predicate I. *Notre Dame Journal of Formal Logic*, 26(4):361–376, 1985.

[24] P. Pudlak. Ramsey's theorem in bounded arithmetic. In E. Börger, H. Kleine Büning, M. Richter, and W. Schönfeld, editors, *Computer Science Logic: Proceedings of the 4th Workshop, CSL '90*. 1991.

[25] A. Shamir R. Rivest and L. Adleman. A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM*, 21:120–126, 1978.

[26] S. Riis. Making infinite structures finite in models of second order bounded arithmetic. In P. Clote and J. Krajicek, editors, *Arithmetic, Proof Theory, and Computational Complexity*, pages 289–319. Oxford University Press, 1993.

[27] V. Shoup. Lower bounds for discrete logarithms and related problems. *Proceedings of Eurocrypt '97*, pages 246–266, 1997.

[28] N. Thapen. A model-theoretic characterization of the weak pigeonhole principle. To appear in Annals of Pure and Applied Logic, 2002.

[29] A. Wilkie and J. Paris. On the scheme of induction for bounded arithmetic formulas. *Annals of Pure and Applied Logic*, 35:261–302, 1987.

[30] D. Zambella. Notes on polynomially bounded arithmetic. *Journal of Symbolic Logic*, 61(3):942–966, 1996.