Fragments of Approximate Counting

Samuel R. Buss*
Department of Mathematics
University of California, San Diego
La Jolla, CA 92093-0112, USA
sbuss@math.ucsd.edu

Leszek Aleksander Kołodziejczyk†
Institute of Mathematics
University of Warsaw
Banacha 2, 02-097 Warszawa, Poland
lak@mimuw.edu.pl

Neil Thapen‡
Institute of Mathematics
Academy of Sciences of the Czech Republic
Žitná 25, 115 67 Praha 1, Czech Republic
thapen@math.cas.cz

January 27, 2012

Abstract

We study the long-standing open problem of giving $\forall\Sigma^b_1$ separations for fragments of bounded arithmetic in the relativized setting. Rather than considering the usual fragments defined by the amount of induction they allow, we study Jeřábek's theories for approximate counting and their subtheories. We show that the $\forall\Sigma^b_1$ Herbrandized ordering principle is unprovable in a fragment of bounded arithmetic that includes the injective weak pigeonhole principle for polynomial time functions, and also in a fragment that includes the surjective weak pigeonhole principle for $\mathrm{FP}^{\mathrm{NP}}$ functions. We further give new propositional translations, in terms of random resolution refutations, for the consequences of T^1_2 augmented with the surjective weak pigeonhole principle for polynomial time functions.

*Supported in part by NSF grant DMS-1101228.

†Partially supported during the early stages of this work by Polish Ministry of Science and Higher Education grant N201 382234. Part of the work was carried out while the author was visiting the University of California, San Diego, supported by Polish Ministry of Science and Higher Education programme "Mobilność Plus".

‡Partially supported by institutional research plan AV0Z10190503 and grant IAA100190902 of GA AV ČR and by grant 1M0545 (ITI) of MŠMT. Part of this work was done on visits to the University of Warsaw and to the University of California, San Diego.

1 Introduction

In a series of papers [12, 14, 16], Jeřábek developed two theories of approximate counting in bounded arithmetic. In order to give them succinct names, we will call these theories APC_1 and APC_2. The weaker theory, APC_1, is $\mathrm{PV}_1 + \mathrm{sWPHP}(\mathrm{PV}_1)$. Here sWPHP is the surjective version of the weak pigeonhole principle stating that there is no mapping from $[x]$ onto $[x(1 + 1/|x|)]$, and PV_1 stands for both a set of function symbols used to represent polynomial time functions and a theory in which these functions are well-behaved. APC_1 can reason about the approximate size of a polynomial time subset X of an interval $[0, a)$, up to an error a polynomial fraction of the size a of the interval. This makes APC_1 suitable for developing some parts of the theory of probability and probabilistic computations [14].

The (conjecturally) stronger theory, APC_2, is $T^1_2 + \mathrm{sWPHP}(\mathrm{PV}_2)$, where PV_2 denotes the relativization of PV_1 to an NP oracle. APC_2 can reason about the approximate size of an NP subset X of an interval $[0, a)$, up to an error a polynomial fraction of the size of the set X. This allows APC_2 to carry out many arguments based on approximate counting and repeatedly subdividing a set, including the usual proofs of the finite Ramsey theorem and the tournament principle [16].

The theories APC_1 and APC_2 are subtheories of the fragment T^3_2 of bounded arithmetic (see Figure 1). It is of course an open question whether these theories are distinct, and even whether S^1_2 equals $S_2 = T_2$. However, separations are known in the relativized setting where an uninterpreted predicate symbol α (an oracle) is added to the language. In particular, as we discuss below, the three theories $S^1_2(\alpha)$, $T^1_2(\alpha)$, and $T^2_2(\alpha)$ are known to be separated by their $\forall\Sigma^b_1(\alpha)$ consequences. On the other hand, S^{i+1}_2 is $\forall\Sigma^b_i$-conservative over T^i_2 [6], and $S^1_2 + \mathrm{sWPHP}(\mathrm{PV}_1)$ is conservative over APC_1 [12]. These conservation results also hold for the relativized theories.

It is thus interesting to consider the strength of the relativized versions of the theories APC_1 and APC_2 and other related theories. We would like to compare or separate the $\forall\Sigma^b_1(\alpha)$ consequences of these theories from those of others in the bounded arithmetic hierarchy.

It is a long-standing open question whether there is some fixed k such that the fragments $T^i_2(\alpha)$ of the bounded arithmetic hierarchy can be separated by a $\forall\Sigma^b_k(\alpha)$ sentence, and, in particular, whether there is a $\forall\Sigma^b_1(\alpha)$ sentence which is provable in $T_2(\alpha)$ but not in $T^2_2(\alpha)$. We can ask a similar question about $\mathrm{APC}_2(\alpha)$:

Open Problem 1 *Is there a $\forall\Sigma^b_1(\alpha)$ sentence which is provable in full bounded arithmetic $T_2(\alpha)$, but not in $\mathrm{APC}_2(\alpha)$?*

This seems to be a hard question. It is interesting to note that we do not know any inclusions between the $\forall\Sigma^b_1$ consequences of T^2_2 and APC_2,

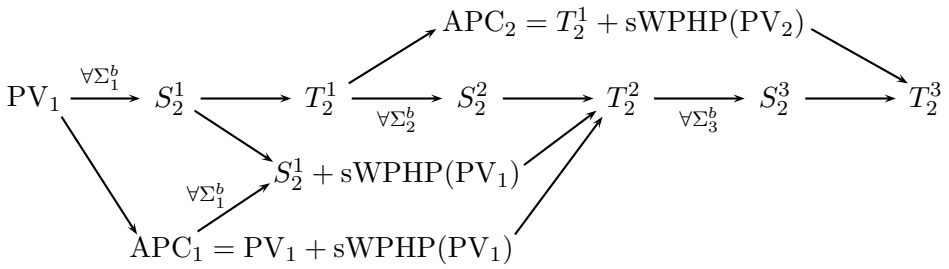

Figure 1: The prior-studied fragments of bounded arithmetic based on approximate counting. The arrows show the direction of inclusion, and the labels show conservativity. For instance, $T_2^2 \vdash \mathrm{APC}_2$ and S_2^1 is $\forall\Sigma_1^b$-conservative over PV_1.

and that the two theories T_2^2 and APC_2 live at roughly the same level in the bounded arithmetic hierarchy, namely between T_2^1 and T_2^3. Thus T_2^2 and APC_2 both lie on the boundary of where we are unable to establish relativized $\forall\Sigma_1^b$ separation results. A deeper reason for seeking lower bounds for theories such as APC_1 and APC_2 that are based on approximate counting is that the $\forall\Sigma_1^b$ principles that are known to separate weak bounded arithmetic theories tend to fall into two groups: either they are natural principles coming from finite combinatorics (such as the weak pigeonhole principle or the finite Ramsey theorem) for which we can prove lower bounds directly, or they are principles which are in some sense "complete" (such as reflection principles for propositional logic), for which unprovability follows from unprovability for some principle in the first group. The principles in the first group all happen to be provable using the amount of approximate counting available in APC_2. So approximate counting is an obstacle that we should expect to have to tackle when looking for stronger unprovability results.

More generally, the importance of approximate counting in computational complexity and finite combinatorics suggests strongly that the theory $\mathrm{APC}_2(\alpha)$ is worth studying in its own right. Indeed, if all the $\forall\Sigma_1^b(\alpha)$ consequences of $T_2(\alpha)$ do turn out to be provable in some weak fragment of $T_2(\alpha)$, in many ways a theory for approximate counting is a more natural candidate for that fragment than $T_2^2(\alpha)$ is. Thus, we feel that $\mathrm{APC}_2(\alpha)$ might be an even more natural choice for a "barrier" theory than $T_2^2(\alpha)$: namely, we might be able to separate the $\forall\Sigma_1^b(\alpha)$ consequences of theories below $\mathrm{APC}_2(\alpha)$, but not to separate $\mathrm{APC}_2(\alpha)$ from full $T_2(\alpha)$.

Before stating our results, we briefly recall the best known (relativized) separations for the theories shown in Figure 1. First, the theories $\mathrm{PV}_1(\alpha)$ and $S_2^1(\alpha)$ are $\forall\Sigma_1^b(\alpha)$-separated from $T_2^1(\alpha)$ as shown by Buss [unpublished], Krajíček [18] and Pudlák [28], and also as follows from the PLS characterization of the $\forall\Sigma_1^b$-definable functions of T_2^1 [8]. The theories $T_2^1(\alpha)$ and $S_2^2(\alpha)$ are $\forall\Sigma_1^b(\alpha)$-separated from $T_2^2(\alpha)$, via versions of the iteration princi-

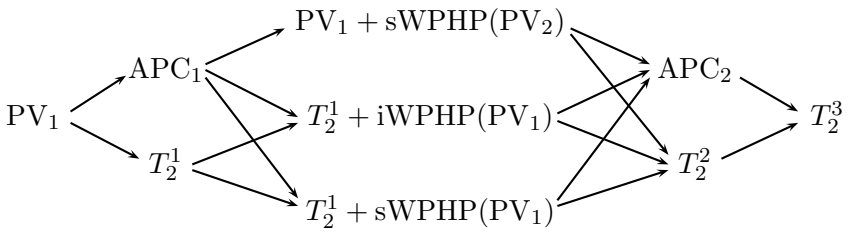

Figure 2: The present paper obtains results for relativized APC_1 and for relativized versions of theories intermediate between APC_1 and APC_2. Sections 3 and 4 show that, in the relativized setting, HOP is not a consequence of $T^1_2 + \mathrm{iWPHP}(\mathrm{PV}_1)$ or $\mathrm{PV}_1 + \mathrm{sWPHP}^{2a}_a(\mathrm{PV}_2)$. Section 5 discusses propositional translations of $\forall\Sigma^b_1$-consequences of $T^1_2 + \mathrm{sWPHP}(\mathrm{PV}_1)$.

ple, the pigeonhole principle, and the Ramsey principle [9, 10]. Finally, for $i \geq 2$, $T^i_2(\alpha)$ and $S^{i+1}_2(\alpha)$ are $\forall\Sigma^b_{i+1}(\alpha)$-separated from $T^{i+1}_2(\alpha)$ [8]; however, it is open whether they are $\forall\Sigma^b_i(\alpha)$-separated. There are also a number of characterizations of the $\forall\Sigma^b_k$-consequences of T^i_2 [33, 30, 1, 2, 17], but none of these have achieved any separations.

We are not able to prove a $\forall\Sigma^b_1(\alpha)$ separation of $\mathrm{APC}_2(\alpha)$ and $T^2_2(\alpha)$; instead, we prove results for several fragments of $\mathrm{APC}_2(\alpha)$. In particular, we study (the relativized versions of) the three theories shown in the middle of Figure 2. In the figure, sWPHP is, as above, the surjective version of the weak pigeonhole principle, whereas iWPHP is the more traditional injective version. Some care is needed when defining the theory $\mathrm{PV}_1 + \mathrm{sWPHP}(\mathrm{PV}_2)$, because PV_1 does not prove the totality of PV_2 functions. (Complete definitions are found in the next section and Section 4.) In addition, we are not able to prove our separation results for the theory $\mathrm{PV}_1 + \mathrm{sWPHP}(\mathrm{PV}_2)$, but instead work with its subtheory $\mathrm{PV}_1 + \mathrm{sPHP}^{2a}_a(\mathrm{PV}_2)$.

By a theorem of Wilkie (see [20, 35]), the $\forall\Sigma^b_1$ sentences provable in APC_1 can be witnessed in probabilistic polynomial time. It is known that this witnessing theorem can be used to show that $\mathrm{APC}_1(\alpha)$ does not prove all the $\forall\Sigma^b_1(\alpha)$ consequences of $T^2_2(\alpha)$ [35].[1] Two of our main results below state that the same holds for $T^1_2(\alpha) + \mathrm{iWPHP}(\mathrm{PV}_1(\alpha))$ and $\mathrm{PV}_1(\alpha) + \mathrm{sPHP}^{2a}_a(\mathrm{PV}_2(\alpha))$. In particular, neither theory proves the (relativized) Herbrandized ordering principle, HOP, which we define in the next section. However, we should note that the relativized HOP principle is not a candidate for separating $T^2_2(\alpha)$ from $\mathrm{APC}_2(\alpha)$, as it is provable in both of those theories.

Somewhat surprisingly, even though in many contexts sWPHP seems weaker than iWPHP, we have not been able to prove a $\forall\Sigma^b_1(\alpha)$ separation re-

[1] Actually, it can be shown in a similar way that $\mathrm{APC}_1(\alpha)$ does not prove all the $\forall\Sigma^b_1$ consequences of $T^1_2(\alpha)$.

sult for $T^1_2(\alpha)$+sWPHP($\mathrm{PV}_1(\alpha)$). On the other hand, we do prove a witnessing theorem for this theory in terms of randomized polynomial local search problems. We also discuss the notion of *random resolution refutation* and show that if a $\forall\Sigma^b_1(\alpha)$ sentence σ is provable in $T^1_2(\alpha) + \mathrm{sWPHP}(\mathrm{PV}_1(\alpha))$, then the propositional translations of the negation of σ have polylogarithmic width random resolution refutations (or, equivalently, quasipolynomial size random treelike Res(log) refutations). We conjecture that HOP and iWPHP have no such refutations, but are not able to prove this. In particular, it seems to be difficult to adapt the usual lower bound techniques for narrow resolution (or the methods for showing unprovability from $T^1_2(\alpha)$) to work in this case.

The question of finding a $\forall\Sigma^b_1(\alpha)$ separation of $T^1_2(\alpha)$+sWPHP($\mathrm{PV}_1(\alpha)$) from stronger fragments of bounded arithmetic thus appears to be the natural next step towards a solution of Open Problem 1 and related problems. As far as we know, $T^1_2(\alpha) + \mathrm{sWPHP}(\mathrm{PV}_1(\alpha))$ is the weakest relatively natural bounded arithmetic theory whose $\forall\Sigma^b_1(\alpha)$ consequences have not been separated from those of full $T_2(\alpha)$.

The remainder of the paper is organized as follows. Section 2 discusses definitions and background material, and presents a generalized "Student-Teacher" algorithm which can be used to witness high complexity consequences of universal theories. The Student-Teacher game is applied in Section 4 to the case of $\mathrm{PV}_1(\alpha)$. Section 3 shows that $T^1_2(\alpha)$+iWPHP($\mathrm{PV}_1(\alpha)$) does not prove the relativized HOP principle, and briefly discusses a similar result in which the weak pigeonhole principle is replaced with a weak version of the finite Ramsey theorem. Section 4 shows that $\mathrm{PV}_1(\alpha)$ + $\mathrm{sPHP}^{2a}_a(\mathrm{PV}_2(\alpha))$ also does not prove the relativized HOP principle. Section 5 studies the $\forall\Sigma^b_1(\alpha)$ consequences of $T^1_2(\alpha) + \mathrm{sWPHP}(\mathrm{PV}_1(\alpha))$ and proves our witnessing results and propositional translations based on random resolution refutations.

Acknowledgements The authors would like to thank Albert Atserias, Stefan Dantchev, Emil Jeřábek, and Massimo Lauria for discussions that helped to improve this paper.

2 Preliminaries

We assume that the reader has some familiarity with bounded arithmetic (see [5, 20, 7, 11]), but we recall the basic definitions below, and fix some notation and terminology in the process.

We take the basic language of bounded arithmetic to consist of the symbols $+, \cdot, \#, |x|, \lfloor x/2^y \rfloor, \leq, 0, 1$, where $|x|$ is $\lceil \log_2(x+1) \rceil$ (length in binary) and $x\#y$ is $2^{|x|\cdot|y|}$. (A more traditional choice would have been to use $\lfloor x/2 \rfloor$ instead of $\lfloor x/2^y \rfloor$, but the latter facilitates sequence coding and makes some theories more robust.)

A hierarchy of bounded formulas in this language is defined as follows. The class Σ^b_0, or Π^b_0, consists of formulas in which all the quantifiers are *sharply bounded*, i.e., bounded by a term of the form $|t|$. For $i \geq 1$, the class Σ^b_i is the closure of Π^b_{i-1} under $\wedge, \vee$, bounded $\exists$, and sharply bounded quantifiers, while Π^b_i is defined dually. In the standard model of arithmetic $\mathbb{N}$, the Σ^b_i formulas define exactly those sets which lie on the i-th level of the polynomial hierarchy, Σ^p_i.

The theory T^i_2 is axiomatized by a finite universal theory BASIC which fixes the meaning of the symbols, and the induction scheme for Σ^b_i formulas. In the theory S^i_2, the usual Σ^b_i induction scheme is replaced by *length induction*:

$$\theta(0) \wedge \forall x{<}|c|\,(\theta(x) \rightarrow \theta(x+1)) \rightarrow \theta(c).$$

Full bounded arithmetic, $S_2 = T_2$, is the union of T^i_2 over all i. It is well-known that

$$S^1_2 \subseteq T^1_2 \subseteq S^2_2 \subseteq T^2_2 \subseteq \cdots$$

and the fundamental problem of bounded arithmetic is whether this hierarchy of theories collapses.

The theory S^{i+1}_2 is $\forall\Sigma^b_{i+1}$-conservative over T^i_2 ([6]; for $i = 0$, this result is due to [13] and requires the presence of $\lfloor x/2^y \rfloor$, cf. [4]). Moreover, all $\forall\Sigma^b_{i+1}$ consequences of S^{i+1}_2 are witnessed by polynomial time functions with a Σ^p_i oracle [5]. Conversely, all $\mathrm{FP}^{\Sigma^p_i}$ functions have provably total Σ^b_{i+1} definitions in S^{i+1}_2 (and hence in T^i_2). In particular, the $\forall\Sigma^b_1$ consequences of S^1_2 are witnessed by polynomial time functions, and all polynomial time functions are provably total in S^1_2.

The $\forall\Sigma^b_1$ consequences of T^1_2 are witnessed by *polynomial local search* (PLS) problems [8]. A PLS problem is given by a term u and polynomial time functions C and N (the *cost* and *neighborhood* functions), all of which take an extra parameter c. The function N always maps the interval $[0, u(c))$ to itself, and a solution to the PLS problem on input c is any $s < u(c)$ such that $C(N(s)) \geq C(s)$. The PLS problem witnesses the formula $\forall x\,\exists z\,\theta(x, z)$ if some polynomial time function F takes every solution of the PLS problem on input c to a witness for $\exists z\,\theta(c, z)$.

For $i \geq 1$, S^i_2 proves a replacement principle which implies that every Σ^b_i formula is equivalent to a *strict* Σ^b_i, denoted $\hat{\Sigma}^b_i$, formula, which has the form

$$\exists x_1{<}t_1\,\forall x_2{<}t_2\,\cdots \mathrm{Q}x_i{<}t_i\,\theta$$

with θ sharply bounded. For this reason, we typically do not distinguish between Σ^b_i and $\hat{\Sigma}^b_i$ when working in a theory that contains S^i_2. However, we do use the $\hat{\Sigma}^b_i$ notation when dealing with theories which do not prove the relevant replacement principles (as in Section 4 of the present paper).

PV_1 is a universal theory in a language with symbols for all polynomial time computable functions, which may be taken to contain the basic language of bounded arithmetic. The notation PV_1 is also commonly used for this set of function symbols itself. The axioms of PV_1 include defining equations for the PV_1 functions and a form of induction for all open formulas in the language. PV_1 is often simply called PV in the literature. For $i \geq 1$, an analogous theory and set of symbols PV_{i+1} corresponding to $\mathrm{FP}^{\Sigma_i^b}$ functions can also be defined. PV_{i+1} is a conservative extension of T_2^i, and the two theories are often identified (though for a number of reasons the notation PV_1 is more common than T_2^0). In particular, all PV_1 functions are already Σ_1^b definable in T_2^0, and hence we may treat them as part of the language of bounded arithmetic without changing significantly the complexity of the formulas we are interested in. In the rest of this paper, PV_1 function symbols will be used freely in terms.

A simple but important property of bounded arithmetic, first proved in [25], is known as Parikh's theorem: if bounded arithmetic proves $\forall x\, \exists y\, \theta(x, y)$ for bounded θ, then it actually proves $\forall x\, \exists y{<}t(x)\, \theta(x, y)$ for some term t.

All bounded arithmetic theories can be relativized by adding an uninterpreted predicate, or "second order parameter", α (and more predicates, if convenient) to the language. The predicate α is allowed in atomic formulas, and Σ_i^b induction is replaced with $\Sigma_i^b(\alpha)$ induction. In the case of PV_i, the relevant changes include modifying the definition of the PV_i functions so that they are allowed to make oracle queries to α. The relativized versions of T_2^i, S_2^i and PV_{i+1} are commonly denoted $T_2^i(\alpha)$, $S_2^i(\alpha)$ and $\mathrm{PV}_{i+1}(\alpha)$.

All the provability and conservativity results mentioned above carry over to the relativized case. On the other hand, as discussed in the introduction, a number of separations are known for relativized theories. Since the present paper is primarily interested in the relativized case, and deals with theories whose names are long enough already and sometimes have multiple second-order parameters, we often suppress the "(α)" notation and treat relativization as implicit. More precisely, we adhere to the following conventions. First, our positive results, about provability and witnessing, are stated without relativization, and hold for both the relativized and unrelativized cases (whenever they make sense). Second, our negative results, about unprovability, are stated in the form "In the relativized language with the symbols $\cdots$, the theory $T \nvdash \varphi$", where it is implicitly understood that T is actually $T(\cdots)$, and that φ typically contains symbols from $\cdots$ in addition to those from the basic language of arithmetic. (Since we do not have any unprovability results for the unrelativized cases, there should be no risk of confusion.)

Other notational conventions include the following. Interval notation such as $[a, b)$ always stands for the appropriate interval in the integers. We often write $[a]$ for $[0, a)$. Whenever f is a function of more than one argument, the notation $f_e(u)$ means exactly the same thing as $f(e, u)$, and is

intended to emphasize that we treat u as the "actual argument" and e as a parameter.

2.1 Weak pigeonhole principles

For a function f, possibly with parameters, and elements $a < b$, the injective pigeonhole principle $\mathrm{iPHP}^b_a(f)$ says that f does not map b pigeons injectively into a holes:

$$\exists x{<}b\, f(x){\geq}a \;\vee\; \exists x_1{<}x_2{<}b\, f(x_1){=}f(x_2).$$

For a class of functions Γ (typical choices include $\Gamma = \mathrm{PV}_1, \mathrm{PV}_2$ etc.) and a term t such that $t(x) > x$, the scheme $\mathrm{iPHP}^{t(a)}_a(\Gamma)$ consists of the universal closures of $\mathrm{iPHP}^{t(a)}_a(f)$ for $f \in \Gamma$.

The principle is referred to as a "weak" pigeonhole principle, and denoted iWPHP, when $t(x)$ is "much bigger" than x. The weak pigeonhole principle has been the object of extensive study in both bounded arithmetic, beginning with [26], and propositional proof complexity. Traditionally, $t(x)$ "much bigger" than x meant $t(x) = x\#x$, x^2, or $2x$. We use $t(x) = x(1 + 1/|x|)$ for reasons that have more to do with the surjective variant of WPHP (see below), but in most settings of interest to us the exact choice of t is irrelevant, due to *amplification* ([26, 34, 15]). Given a polynomial time function f violating $\mathrm{iPHP}^{a(1+1/|a|)}_a(f)$ and an element $b > a$, the theory PV_1 can amplify f to give another polynomial time function g violating $\mathrm{iPHP}^b_a(g)$ (the only new parameter used by g is b, and this could be replaced by any other parameter which is not superpolynomially smaller than b). Thus, already over the base theory PV_1, the scheme $\mathrm{iWPHP}(\mathrm{PV}_1) = \mathrm{iPHP}^{a(1+1/|a|)}_a(\mathrm{PV}_1)$ is equivalent to $\mathrm{iPHP}^{a\#a}_a(\mathrm{PV}_1)$ and all the intermediate schemes.

The scheme $\mathrm{iWPHP}(\mathrm{PV}_1)$ is known to be provable in T^2_2 [24] and, in the relativized setting, unprovable in S^2_2 [31].

The surjective (or dual) pigeonhole principle $\mathrm{sPHP}^b_a(f)$ says that f does not map a arguments surjectively onto b values:

$$\exists v{<}b\, \forall u{<}a\, (f(u) \neq v).$$

Note that this is a $\hat{\Sigma}^b_2$ formula about a, b and f, unlike $\mathrm{iPHP}^b_a(f)$, which is $\hat{\Sigma}^b_1$. For a class of functions Γ, the scheme $\mathrm{sPHP}^{t(a)}_a(\Gamma)$ consists of universal closures of $\mathrm{sPHP}^{t(a)}_a(f)$ for $f \in \Gamma$.

As before, the principle is referred to as "weak", and denoted sWPHP, when $t(x)$ is "much bigger" than x. Originally less studied than iWPHP, sWPHP gained prominence when Jeřábek showed first that the theory $\mathrm{PV}_1 + \mathrm{sPHP}^{a(1+1/|a|)}_a(\mathrm{PV}_1)$ supports a basic notion of approximate cardinality of polynomial time sets [14], and later that $T^1_2 + \mathrm{sPHP}^{a(1+1/|a|)}_a(\mathrm{PV}_2)$ supports a more robust notion of approximate cardinality, powerful enough to formalize

typical combinatorial arguments using approximate counting, in particular proofs of many combinatorial principles used to separate low levels of the relativized bounded arithmetic hierarchy [16].

We let sWPHP(Γ) stand for $\mathrm{sPHP}_a^{a(1+1/|a|)}(\Gamma)$. We let APC_1 stand for $\mathrm{PV}_1 + \mathrm{sWPHP}(\mathrm{PV}_1)$ and APC_2 stand for $T_2^1 + \mathrm{sWPHP}(\mathrm{PV}_2)$.

The situation with amplifying sPHP is more complicated than with iPHP. In particular, the relativized schemes $\mathrm{sWPHP}(\mathrm{PV}_1)$, $\mathrm{sPHP}_a^{2a}(\mathrm{PV}_1)$ and $\mathrm{sPHP}_a^{a^2}(\mathrm{PV}_1)$ are all distinct over PV_1 [15]. On the other hand, they are all equivalent over S_2^1. Moreover, already over PV_1 there is a conservativity result: $\mathrm{PV}_1 + \mathrm{sWPHP}(\mathrm{PV}_1)$ is $\forall\Sigma_1^b$-conservative over $\mathrm{PV}_1 + \mathrm{sPHP}_a^{a\#a}(\mathrm{PV}_1)$ (noted in [15] as a corollary of earlier results).

Similarly to $\mathrm{iWPHP}(\mathrm{PV}_1)$, $\mathrm{sWPHP}(\mathrm{PV}_1)$ is provable in T_2^2 but its relativized version is unprovable in S_2^2. It follows from the provability result that APC_1 is a subtheory of T_2^2, and APC_2 is a subtheory of T_2^3.

The known relationships between the relativized versions of $\mathrm{iWPHP}(\mathrm{PV}_1)$ and $\mathrm{sWPHP}(\mathrm{PV}_1)$ are as follows:

- $S_2^2(\alpha) + \mathrm{iWPHP}(\mathrm{PV}_1(\alpha))$ does not prove $\mathrm{sWPHP}(\mathrm{PV}_1(\alpha))$ [15].
- On the other hand, already $\mathrm{PV}_1 + \mathrm{iWPHP}(\mathrm{PV}_1)$ proves all the $\forall\Sigma_1^b$ consequences of $\mathrm{PV}_1 + \mathrm{sWPHP}(\mathrm{PV}_1)$ [12].
- $\mathrm{PV}_1(\alpha) + \mathrm{sWPHP}(\mathrm{PV}_1(\alpha))$ does not prove $\mathrm{iWPHP}(\mathrm{PV}_1(\alpha))$ [35].
- The provability of iWPHP(PV) in $S_2^2 + \mathrm{sWPHP}(\mathrm{PV}_1)$, or equivalently $T_2^1 + \mathrm{sWPHP}(\mathrm{PV}_1)$, is an open problem. In general, the strength of (the relativized version of) $T_2^1 + \mathrm{sWPHP}(\mathrm{PV}_1)$ is not very well understood (see Section 5 below).

Both $\mathrm{iPHP}_a^{t(a)}(\mathrm{PV}_1)$ and $\mathrm{sPHP}_a^{t(a)}(\mathrm{PV}_1)$ are officially infinite schemes. However, since every violation of a pigeonhole principle takes place on a bounded interval, each scheme can be replaced by a single sentence involving the universal polynomial time machine with parameters.

Normally, WPHP(Γ), in whatever version, is studied in a context where functions from Γ are represented by function symbols or at least provably total definitions. However, in Section 4 of this paper we study sWPHP for PV_2 functions over the theory PV_1, too weak to prove the totality of $\mathrm{FP}^{\mathrm{NP}}$ functions. We defer discussing the issues this raises until Section 4.

2.2 Ordering principles

As our sentence that separates various fragments of APC_2 from T_2^2, we will use a weak, Herbrandized version of the following principle.

Definition 2 *The* ordering principle *is the universal closure of a Σ_2^b (in fact, $\hat{\Sigma}_2^b$) formula with a size parameter c and a second order parameter for*

a binary relation $\preceq$ on $[c]$. It asserts that if $\preceq$ is a partial ordering, then it has a minimal element.

We will write $x \prec y$ for $x \preceq y \wedge x \neq y$.

Proposition 3 *The ordering principle is provable in T^2_2.*

Proof Induction on c. □

Theorem 4 *The ordering principle is provable in* APC_2.

Proof If we weaken the principle to "if $\preceq$ is a *total* ordering, then it has a minimal element", then it follows easily from the tournament principle, which is provable in APC_2 [16]. This is because the total ordering directly defines a tournament on $[c]$ by the rule "x beats y if and only if $x \prec y$". By the tournament principle, this tournament has a logarithmic size dominating set. We can find the minimal element of this set by a brute force search, and this element must then be minimal for the whole interval.

For the full principle, we use a more sophisticated argument due to Jeřábek. Say that a partial ordering is *directed* if every two elements have a common lower bound. We can find a minimal element in a directed order in much the same way as in a total one: define a tournament by "x beats y if and only if either x is $\preceq$-comparable to y and $x \prec y$, or x is not $\preceq$-comparable to y and $x < y$". (We could use an arbitrary polynomial time antisymmetric relation instead of $<$.) Let S be the logarithmic size dominating set. Since $\preceq$ is directed, by brute force using Σ^b_1 length induction we can find a point q which is a $\preceq$-lower bound for S. Then q must be a $\preceq$-minimal element of $[c]$. Otherwise there would exist a point $r \prec q$, which would therefore beat every member of S, but that is impossible.

It remains to show that the case of general $\preceq$ reduces to that of directed $\preceq$. For $p < c$, write L_p for the set $\{x < c : x \preceq p\}$. We first find a point p for which the ordering $\preceq$ restricted to L_p is directed. Informally, we do this by finding p for which the size $|L_p|$ is "approximately minimal", meaning that there is no point q with $|L_q| < |L_p|/2$. Once we have such a p, it must be the case that $\preceq$ is a directed ordering on L_p, since otherwise there would exist two points $q_1, q_2 \prec p$ with no common lower bound, implying that L_{q_1} and L_{q_2} are disjoint, and hence that one of them must have size less than $|L_p|/2$.

This argument can be formalized using the machinery of [16]. We write $X \trianglelefteq_\varepsilon s$ for Jeřábek's relation "X is ε-approximately smaller than s". Fix $\varepsilon = 1/10$. Let $\theta(p, s)$ express $L_p \trianglelefteq_\varepsilon s$. By Σ^b_2-length minimization, which we can use by Corollary 1.15 of [12], find a pair $\langle p, s\rangle$ for which $\theta(p, s)$ holds and $|s^2|$ is minimal. Suppose there are $q_1, q_2 \prec p$ with L_{q_1} and L_{q_2} disjoint. By Theorem 3.17 of [16], if we let $r = \lfloor s/2 \rfloor$, then since $L_{q_1} \cup L_{q_2} \trianglelefteq_\varepsilon r + r + 1$ we must have either $L_{q_1} \trianglelefteq_\varepsilon r(1 + 2\varepsilon)$ or $L_{q_2} \trianglelefteq_\varepsilon r(1 + 2\varepsilon)$. Now

$(r(1+2\varepsilon))^2 \leq (s^2/4)(12/10)^2 < s^2/2$, so $|(r(1+2\varepsilon))^2| < |s^2|$ and in either case we contradict the minimality of $|s^2|$.

Change $\preceq$ by totally ordering the set $[c] \setminus L_p$ according to the standard $<$ ordering and putting its elements above all elements of L_p, which remain ordered by the old $\preceq$. Then the new $\preceq$ is directed, but since L_p was nonempty, we have not introduced any new minimal elements. □

Definition 5 *The* Herbrandized ordering principle HOP *is the universal closure of a Σ^b_1 (in fact, $\hat{\Sigma}^b_1$) formula with a size parameter c and two second order parameters, one for a binary relation $\preceq$ on $[c]$ and one for a unary function $h : [c] \to [c]$. The formula asserts that the following cannot all be true:*

1. *$\preceq$ is a total ordering on $[c]$;*
2. *for all $x < c$, $h(x) \prec x$;*
3. *for all $x, y < c$, it is not the case that $h(x) \prec y \prec x$.*

In other words: in a finite total ordering, it is not possible for every element to have an immediate predecessor.

As defined, HOP is a relativized principle and uses two second order parameters, $\preceq$ and h. These could be coded by a single unary predicate α by letting α encode both the binary predicate $\preceq$ and the bit graph of h. HOP can also be used as an unrelativized principle, by letting $\preceq$ and h be replaced by arbitrary predicates and functions from PV_1.

Proposition 6 HOP *is provable in T^2_2 and in* APC_2. □

We also note the following:

Proposition 7 HOP *is provable in* $\mathrm{PV}_1 + \mathrm{PHP}(\mathrm{PV}_1)$.

Proof Suppose HOP fails. Then define an injection $f : [c] \to [c] \setminus \{0\}$ by

$$f : x \mapsto \begin{cases} x & \text{if } x \succ 0 \\ h(x) & \text{if } x \preceq 0. \end{cases}$$

□

Similar sentences, called *generalized iteration principles*, were considered in [9] and shown to separate $T^1_2(\alpha)$ from $T^2_2(\alpha)$. Versions of the ordering principle are also used in propositional proof complexity, under the name *graph ordering principle* or *graph tautology*. In particular a natural bounded-width DNF (analogous to $\hat{\Sigma}^b_1$) version arises by fixing a bounded-degree expander graph and considering orderings defined only on edges in the graph [32].

2.3 A generalized Student-Teacher game

The Student-Teacher game variant of Herbrand's theorem was first introduced by Krajíček, Pudlák and Takeuti [22], and has subsequently been used for a number of other similar applications in bounded arithmetic. Pudlák [29] presented a generalized version of the Student-Teacher game. We present below a somewhat simplified version of the Student-Teacher game in which the Student and Teacher establish the truth of a prenex formula by taking turns writing formulas on a blackboard. The Student uses terms to instantiate existential quantifiers, and the Teacher replies with values for universal quantifiers.

Our application below (Theorem 17 in Section 4) will use the theory PV_1 and the standard model $\mathbb{N}$. However, for the sake of generality, we define the Student-Teacher game for an arbitrary model M of an arbitrary universal theory T. Let $\Phi(z)$ be a formula

$$\exists x_1 \forall y_1 \cdots \exists x_n \forall y_n \ \phi(z, x_1, y_1, \ldots, x_n, y_n)$$

where ϕ is quantifier-free.[2] Let $m \in M$. The game is played by the two players, the Student and the Teacher, who construct a sequence of formulas. We think of these formulas being written one after another on a blackboard, with no formula ever erased. These formulas will all be substitution instances of subformulas of Φ, where we allow certain elements of M to appear in the formulas. The Student wins the game when (and if) a quantifier-free formula which is true in M is written on the blackboard.

To begin the game, the formula

$$\exists x_1 \forall y_1 \cdots \exists x_n \forall y_n \ \phi(m, x_1, y_1, \ldots, x_n, y_n)$$

is written on the blackboard.

In the i-th round of the game, the Student chooses one of the non-quantifier-free formulas on the blackboard (possibly one he has chosen before). This formula has the form

$$\exists x_j \forall y_j \cdots \forall y_n \ \phi(m, s_1, n_1, \ldots, s_{j-1}, n_{j-1}, x_j, y_j, x_{j+1}, y_{j+1}, \ldots, x_n, y_n),$$

where $s_1, \ldots, s_{j-1}$ are terms that were selected by the Student in earlier rounds, and $n_1, \ldots, n_{j-1}$ are elements of M that were selected by the Teacher in earlier rounds. The Student chooses a term t_i as a value for the existentially quantified x_j. As explained below, the term t_i may involve the element m of M and any elements of M played by the Teacher in earlier rounds. The Teacher now chooses an element m_i of M as a value for the universally quantified y_j. Then, the formula

$$\exists x_{j+1} \forall y_{j+1} \cdots \forall y_n \ \phi(m, s_1, n_1, \ldots, s_{j-1}, n_{j-1}, t_i, m_i, x_{j+1}, y_{j+1}, \ldots, x_n, y_n)$$

[2] For simplicity, we let treat the variables z, x_i, y_i as single variables, but the construction extends readily to the case where they are vectors of variables.

is written on the blackboard.

A precise way to define the restriction on what elements of M may appear in the Student's term t_i chosen in round i is as follows. Let L_0 be the first order language of T with an extra constant symbol denoting the element m. Define L_i to be the language L_{i-1} plus a constant symbol denoting the element m_i played by the Teacher in round i. Then the term t_i must be an L_{i-1}-term. Equivalently, we have $t_i = t_i(z, z_1, \ldots, z_{i-1})$ an L_0-term with i free variables (not all of which necessarily appear in t_i), and the term played by the Student in round i is $t_i(m, m_1, \ldots, m_{i-1})$.

A *strategy* for the Student is a sequence of pairs $\langle t_1, j_1\rangle$, $\langle t_2, j_2\rangle$,..., $\langle t_k, j_k\rangle$. The strategy indicates that, in the i-th round, the Student chooses the j_i-th formula written on the blackboard and uses the term t_i for his move. Accordingly, each term t_i in the strategy is an L_0-term of the form $t_i(z, z_1, \ldots, z_{i-1})$.

A *winning strategy* is a strategy, of some finite length k, such that for any $M \vDash T$ and for any choice of m in M and any choices for elements m_i played by the Teacher, the Student wins by the end of the k-th round.

Theorem 8 *Let T be a universal theory, and Φ be as above. Suppose $T \vdash \Phi$. Then the Student has a winning strategy.*

Note that since the same winning strategy works for all models M of T, it follows that T proves the correctness of the winning strategy.

When T is PV_1, and M is the standard model $\mathbb{N}$ of the integers, we may assume that the Teacher gives her answers m_i in the form of their binary representations. In this setting, a winning strategy for the Student is polynomial time computable. In addition, we may assume that the Student outputs explicit integer values (in their binary representations) instead of terms. This is because any fixed PV_1-term can be evaluated in polynomial time. With these conventions, Theorem 8 becomes:

Corollary 9 *If Φ is provable in PV_1, then there is a constant $k \in \mathbb{N}$ and a polynomial time strategy for the Student, using which the Student always wins the game within k rounds, provably in PV_1.*

Theorem 8 follows readily from the usual Herbrand theorem. Nonetheless, for sake of completeness, we give a sketch of a model-theoretic proof. Enumerate, for $i = 1, 2, \ldots$, all pairs $\langle t_i, j_i\rangle$, where t_i is an L_0-term and $j_i \in \mathbb{N}$ such that $j_i \leq i$ and such that each term t_i involves only (at most) the free variables $z, z_1, \ldots, z_{i-1}$. This enumeration constitutes an infinitely long strategy for the Student. Consider the formulas that are written on the board while playing as follows. Let $c, c_1, c_2, \ldots$ be new constant symbols. Initially write $\Phi(c)$ on the blackboard. For the i-th round, if the j_i-th formula on the blackboard, Φ_{j_i}, is not quantifier-free, substitute the term

$t_i(c, c_1, \ldots, c_{i-1})$ for the outermost (existential) quantifier of Φ_{j_i} and substitute the constant symbol c_i for the outermost universal quantifier. Write the resulting formula on the blackboard as the next formula Φ_i. If, however, Φ_{j_i} is quantifier-free, just write it again on the blackboard.

Let Γ be the set of quantifier-free formulas that are written on the blackboard after playing the game infinitely many steps. Let Δ contain the negations of the formulas in Γ, so $\Delta = \{\neg\phi : \phi \in \Gamma\}$. If Δ is inconsistent with T, then by compactness, some finite subset Δ_0 is inconsistent with T. In particular, for any model M of T and choice of interpretations for the constants $c, c_1, c_2, \ldots$ in M, some member of Δ_0 is false in M. Thus, in any model M, some member of the corresponding finite subset Γ_0 of Γ is true in M. Thus, choosing k large enough so that all members of Γ_0 have been written on the blackboard after k rounds implies that the strategy, truncated to k rounds, is a finite winning strategy for the Student.

Finally, we claim that Δ must be inconsistent with T. Otherwise, since T is universal, we can construct a Henkin model M of $T \cup \Delta$ which has as domain exactly the closed L_0-terms that use also the constant symbols $c, c_1, c_2, \ldots$. By the definition of Δ, every quantifier-free formula in Γ is false in M. It is easy to show, by induction on quantifier complexity, that each Φ_i written on the blackboard is false in M. This means that Φ itself is false in M, contradicting the assumption that $T \vdash \Phi$. □

3 $T_2^1 + \mathrm{iWPHP}(\mathrm{PV}_1)$

Theorem 10 *In the relativized language with second order parameters $\preceq$ and h, $T_2^1 + \mathrm{iWPHP}(\mathrm{PV}_1) \nvdash \mathrm{HOP}$.*

Corollary 11 *The theory $T_2^1(\alpha) + \mathrm{iWPHP}(\mathrm{PV}_1(\alpha))$ is separated by $\forall\Sigma_1^b(\alpha)$ consequences from $T_2^2(\alpha)$ and $\mathrm{APC}_2(\alpha)$.*

The corollary follows from (the relativized version of) Proposition 6, and the fact that α can encode both $\preceq$ and the bit graph of h.

Proof (of Theorem 10) By amplification, we can use the a^2-to-a version of iWPHP. Suppose

$$T_2^1 + \forall a\, \forall e\, \mathrm{iPHP}_a^{a^2}(f_e) \vdash \exists z\, \theta(c, z),$$

where f_e is the universal function with a parameter e, and $\exists z\, \theta(c, z)$ stands for the Σ_1^b part of HOP. (We are now using the notation T_2^1 to mean the relativized form; the function f_e has oracle access to $\preceq$ and h.)

Again by amplification, we may without loss of generality assume that the parameter e is the same as the size parameter a, that the output of f is always smaller than a, and that we only need iWPHP for values of a of size

at least c. Hence writing iWPHP out in full, negating it, and moving it to the right hand side, and using Parikh's theorem to give a term t bounding a, we get

$$T_2^1 \vdash \exists a{\in}[c,t)\,(\forall x_1{<}x_2{<}a^2\, f_a(x_1){\neq}f_a(x_2)) \vee \exists z\,\theta(c,z).$$

We would like to have a Σ_1^b formula on the right so that we can use the PLS witnessing theorem for T_2^1. So we introduce new function symbols r_1 and r_2 as Skolem functions to get rid of the universal quantifiers, giving

$$T_2^1 \vdash \big[\exists a \in [c,t)\,(r_1(a){<}r_2(a){<}a^2 \to f_a(r_1(a)){\neq}f_a(r_2(a)))\big] \vee \exists z\,\theta(c,z).$$

The theory T_2^1 is now understood to be relativized with the four symbols $\preceq$, h, r_1, and r_2.

The Σ_1^b formula can now be witnessed by a PLS problem with oracles $\preceq$, h, r_1 and r_2. That is, there is a term u (in c), a cost function C, a neighborhood function $N : [u] \to [u]$ and an oracle-free reduction function F such that for any solution $s < u$ with $C(N(s)) \geq C(s)$, we have that $F(s)$ is a pair $\langle a, z\rangle$ witnessing the right hand side. Furthermore, we may bound the running times of C, N, and F by some number k, polynomial in $|c|$. Let l be a similar bound on the running time of f_a over all $a < t$. Note that C and N can make oracle queries to $\preceq$, h, r_1, and r_2, but that f can only query $\preceq$ and h.

Choose c large enough that $c > 12l^2$ and $c/3 > 16kl$.

The key to obtaining a contradiction is to show that we can partially define $\preceq$ and h in such a way that we never introduce a witness to HOP, but we are still able to find a collision in the function f whenever we are asked for one. For this, we need the following definition and claim.

A *good partial structure* consists of:

1. A set of points $P \subseteq [c]$;
2. A set of numbers $A \subseteq [c,t)$;
3. A relation $\preceq$ which defines a total ordering on P and is undefined elsewhere;
4. A partial function h, undefined outside P and undefined on the $\preceq$-least element of P, but giving the immediate $\preceq$-predecessor of any other element of P;
5. Functions r_1, r_2 with domain A satisfying that for all $a \in A$, $r_1(a) < r_2(a) < a^2$ and $f_a(r_1(a))$ and $f_a(r_2(a))$ are defined and equal, where "defined" means that all oracle queries to $\preceq$ and h made when computing f_a are defined in the structure.

Define the size $|H|$ of a structure H to be $|P|$.

Claim 1 Given any good partial structure H which is not too big (in particular, $|P| < c/3$), and any query of the form "$p \preceq q$?", "$h(p) = ?$" or "$r_1(a) = ?$ and $r_2(a) = ?$", there is a good partial structure H' extending H in which an answer to this query is defined.

Given Claim 1, we may complete the proof by using a standard lower bound argument for PLS witnessing. We will first need a bound on how much of the structure needs to be specified to define a computation.

Claim 2 Suppose M is a deterministic Turing machine which can ask queries of the form "$p \preceq q$?", "$h(p) = ?$" or "$r_1(a) = ?$ and $r_2(a) = ?$" and which runs for exactly n steps on input $s < u$. Then given a good partial structure H in which the computation of M on input s is defined, we can define a good partial structure H' of size at most $4nl$ in which the computation of M on input s is defined and is identical to the computation in H.

Proof of Claim 2 Say that a point p is *touched* in a computation if, for any point q, the computation makes an oracle query of the form "$p \preceq q$?" or "$h(p) = ?$", or makes a query of the form "$h(q) = ?$" and gets p as a reply. Now suppose that H is a good partial structure and the computation of $M^H(s)$ is defined. Let the "substructure" H' of H be defined as follows. Let A' be the set of numbers a in A such that the computation queries $r_1(a)$ and $r_2(a)$. Let P' consist of every point in P that is touched directly in the computation of $M^H(s)$, or is touched in the computation of $f_a^H(r_1(a))$ and $f_a^H(r_2(a))$ for any a in A'. Let r_1', r_2' and $\preceq'$ be induced from H, and let h' be the predecessor function arising from $\preceq'$ on P' (this may disagree with h on some points, which is why it is not entirely correct to call H' a substructure of H). Then the computation of $M^{H'}(s)$ is identical to the computation of $M^H(s)$, but the size of H' is at most $4nl$, since in the worst case M makes n queries of the form "$r_1(a) = ?$ and $r_2(a) = ?$" and each computation of $f_a(r_1(a))$ or $f_a(r_2(a))$ makes l queries of the form "$p \preceq q$?".

We now explain how the theorem follows from Claims 1 and 2, and then complete the argument by proving Claim 1.

Consider the set of pairs (s, H) where $s < u$ and H is a good partial structure of size $\leq 12kl$ in which $C^H(s)$, the cost of s in H, is defined. It follows from the two claims that this set is nonempty: by repeatedly invoking Claim 1 to answer successive queries made by C and Claim 2 to control the size of the arising structures, we can extend the empty structure to a suitable good partial structure H of size $\leq 4kl$. Let (s, H) be a pair in this set for which the cost $C^H(s)$ is minimal. We may assume that $|H| \leq 4kl$.

Extend H to H' in which the neighbour $s' := N^{H'}(s)$ of s is defined and has a defined cost $C^{H'}(s')$. We may assume that $|H'| \leq 12kl$.

By the choice of (s, H), we must have $C^{H'}(s') \geq C^H(s) = C^{H'}(s)$.

Thus s is a solution to the PLS problem for any oracle extending H'. Let $F(s) = \langle a, z \rangle$. Using Claim 1 again, we may extend H' to a good partial structure H'' in which $r_1(a), r_2(a)$ are defined, as are the queries needed to determine whether z codes a witness to HOP. By the definition of a good partial structure, in any oracle extending H'' we have $r_1(a) < r_2(a) < a^2 \wedge f_a(r_1(a)) = f_a(r_2(a))$, but z does not code a witness to HOP. This is a contradiction.

Proof of Claim 1 Queries of the form "$p \preceq q$?" and "$h(p) = ?$" are easy to handle by adding new points to the bottom of P as necessary. So consider the last form of query, and suppose we have a good structure H and are queried "$r_1(a) = ?$ and $r_2(a) = ?$" for some $c \leq a < t$.

Extend H to a "total" good partial structure H_0 by extending the ordering $\preceq$ to all points in $[c]$, putting all the new ones below P, arbitrarily ordered, and extending h accordingly, so that it is defined everywhere except for the $\preceq$-minimal point. Call this minimal point p_0. Given any pigeon $x \in [a^2]$, we can simulate a computation of $f_a(x)$ running with H_0 as an oracle, up until the first time "$h(p_0) = ?$" is queried (if this ever happens). At this point we will simply abandon the simulation.

Let X_0 be the set of pigeons $x \in [a^2]$ such that $f_a(x)$, when run with H_0 as oracle, never queries "$h(p_0) = ?$". If $|X_0| > a$ then we are done, because by the strong PHP there must exist x_1, x_2 in X_0 with $x_1 < x_2$ such that $f_a(x_1)$ and $f_a(x_2)$ are defined and equal, when run with H_0 as oracle, so we can define $r_1(a) = x_1$ and $r_2(a) = x_2$.

If $|X_0| \leq a$, let Y_0 be the set of pigeons which do query "$h(p_0) = ?$" when run with H_0. That is, $Y_0 = [a^2] \backslash X_0$. As before, say that a pigeon $x \in Y_0$ *touches* a point p under H_0 if, in the course of computing $f_a(x)$, any query of the form "$p \preceq q?$ ", "$q \preceq p?$" or "$h(p) = ?$" occurs, or some query "$h(q) = ?$" gets p as a reply (for any q). Let $S_0 := P \cup \{p_0\}$ be the set of *settled* points, whose ordering we will not change through the construction. Call all other points *tentative*.

Each pigeon in Y_0 touches (under H_0) at most $2l$ of the $c - |P| - 1$ tentative points. Hence by an averaging argument there must exist some tentative point p_1 which is touched by no more than $|Y_0| \cdot 2l/(c - |P| - 1) \leq a^2 \cdot 4l/c < a^2/3l$ pigeons in Y_0. Let Z_0 be the set of pigeons in Y_0 which do not touch p_1 under H_0. Note that $|Z_0| \geq a^2 - a - a^2/3l \geq a^2 - a^2/2l$.

Now construct a new "total" good partial structure $H_1 \supseteq H$ from H_0 by moving p_1 to the bottom of the ordering, directly below p_0. Change h accordingly, so that $h(p_0) = p_1$, and $h(p_1)$ is undefined, and if q and q' in H_0 were such that $h(q) = p_1$ and $h(p_1) = q'$, then $h(q) = q'$ in H_1.

Consider any pigeon $x \in Z_0$. Under H_0, the computation of $f_a(x)$ proceeded until it queried "$h(p_0) = ?$", at which point we abandoned the simulation, and at no step did it touch p_1. Therefore, under H_1, the first steps of the computation up to and including the query to $h(p_0)$ will be the same,

but the computation can now continue after this point, since we can reply to the query with p_1. What we have gained is that, under H_1, every pigeon in Z_0 wastes at least one step in querying a pigeon which is not the least element.

We can now repeat the construction, but replacing $[a^2]$ with the smaller domain Z_0 of pigeons and adding p_1 to the set of settled points. That is: we define X_1 to be the set of pigeons in Z_0 which do not query "$h(p_1) = ?$" under H_1. If $|X_1| > a$ we are done; otherwise define Y_1 to be $Z_0 \setminus X_1$. Find a tentative point p_2 which is touched by no more than $a^2/3l$ members of Y_1, define Z_1 to be the set of pigeons in Y_1 which do not touch p_2, and define H_2 to be H_1 with the point p_2 moved to be $\preceq$-minimal. Now every point in Z_1 queries h of both p_0 and p_1 under H_2, and $|Z_1| > a^2 - 2 \cdot a^2/2l$.

We repeat the construction l times, stopping if at any stage $i < l$ we have a set X_i which contains a collision that does not query h of the current least point. If there is no such stage i, then we will have Z_{l-1}, H_l and $p_0, \ldots, p_l$, such that $|Z_{l-1}| \geq a^2 - l \cdot a^2/2l = a^2/2$, so Z_{l-1} must contain a collision. But every pigeon in Z_{l-1} queries h of every point $p_0, \ldots, p_{l-1}$, hence is unable to query the least point p_l of H_l, since it has run out of time. Either way, we are done. □

We conclude this section by observing that the proof of the theorem above relies on the fact that the injective WPHP is very over-determined, in the sense that even relatively small subsets of the a^2 pigeons must already contain a collision. A similar phenomenon occurs if we consider a weak form of another principle studied in proof complexity, the finite Ramsey theorem.

The usual Ramsey principle RAM states that $a \to (|a|/2)^2_2$, that is, that any graph on a vertices contains a clique or independent set of size $|a|/2$. This was shown to be provable in bounded arithmetic in [27]. It was later shown that the standard proof of the principle can be formalized in APC_2 [16], and it is conceivable that $T^1_2 + \mathrm{RAM}$ already proves all of the $\forall\Sigma^b_1$ consequences of APC_2.

We show that, in the relativized setting, HOP does not follow from T^1_2 extended by a weak version of RAM. Natural versions to consider would be those saying something like $a \to (|a|/4)^2_2$ or, say, $a \to (|a|/10)^2_2$. Unfortunately, there are difficulties related to the fact that although colouring a relatively small *subgraph* of a graph of size a does give a homogeneous set of size $|a|/10$, it is not true that colouring a relatively small number of *edges* already gives a homogeneous set of that size. For this reason, we are only able to prove independence of HOP from a weaker Ramsey principle (which, however, is still unprovable in relativized S^2_2).

Definition 12 *The weak finite Ramsey theorem* WRAM *takes a first order size parameter a, a first order parameter b, and a ternary second order parameter E for the edge relation of a graph parameterized by an extra argument. It states that the graph E_b induced on $[a]$ by E with parameter b*

contains a homogeneous subset of size at least $|a|^{1/f(a)}$*, where* f *is some non-decreasing function with* $f(a) \geq 2$*. We take* f *here to be* $\log^*(a)$*. (Since our* f *is non-constant, we should perhaps be calling this the "very weak" Ramsey theorem.)*

Theorem 13 *In the relativized language with second order parameters* $\preceq$ *and* h*,* T^1_2 *plus* WRAM *for* PV_1 *graphs does not prove* HOP*.*

Proof (sketch) The proof is similar in outline to that of Theorem 10, and we only describe the main differences. We have to derive a contradiction from the existence of a PLS problem whose solutions produce either witnesses to HOP or pairs $\langle a, b\rangle$ below some $t(c)$ such that $r(a, b)$ is not a $|a|^{1/\log^* a}$-sized E_b-homogeneous subset of $[a]$. Here E is a PV_1 relation with oracle access to $\preceq$ and h, while r is a new function symbol which plays the role of r_1, r_2 from the previous proof; note that we cannot identify the parameter b with the size parameter a.

Once again, the proof comes down to formulating the right notion of good partial structure and proving an analogue of Claim 1 on extending good partial structures.

In the definition of good structure, parts 1, 3 and 4 remain unchanged. Part 2 now speaks about a set A of pairs $\langle a, b\rangle$ rather than single numbers a. In part 5, we set $d := c^{1/3}$ and for all $\langle a, b\rangle \in A$ require $r(a, b)$ to be an E_b-homogeneous subset of $[\min(a, d)]$ of size $|\min(a, d)|/2$. Note that w.l.o.g. $|\min(a, d)|/2$ exceeds $|a|^{1/\log^* a}$.

Our version of Claim 1 again says that good partial structures of size $< c/3$ can be extended to answer a query. The proof splits into two entirely symmetrical cases of $a \geq d$ and $d \geq a$; we sketch the former. The basic construction is as before, except that there are more stages and at each stage we now branch depending on whether there is any pair $\{x, x'\}$ of points below d for which $E_b(x, x')$ queries h of the current $\preceq$-least point. If not (case 1), we are done. Otherwise (case 2), we can choose a new least point which is not touched by *any* of the $\binom{d}{2}$ pairs $\{x, x'\}$; in this way, we ensure that at least one pair progresses at least one more step in its computation without querying h of the least point. After at most $2l\binom{d}{2}$ stages, we have to end up in case 1. □

4 $\mathrm{PV}_1 + \mathrm{sWPHP}(\mathrm{PV}_2)$

The aim of this section is to show, in the relativized setting, the unprovability of HOP in PV_1 extended by the surjective weak pigeonhole principle for PV_2 functions. However, as earlier noted, with only PV_1 as the base theory even making the notion of "surjective weak pigeonhole principle for PV_2 functions" precise requires some care.

The first difficulty encountered is that PV_1 is too weak to prove the totality of PV_2 functions. For $f \in \mathrm{PV}_2$, possibly with a side parameter, we want $\mathrm{sPHP}_a^{t(a)}(f)$ to state that

$$\exists v{<}t(a)\, \forall u{<}a\, f(u){\neq}v, \tag{1}$$

but "$f(u) = v$" can be formulated as either a $\hat{\Pi}_2^b$ formula or a $\hat{\Sigma}_2^b$ formula, and the meaning of (1) depends on the choice. There are reasons to prefer the $\hat{\Sigma}_2^b$ formulation. For instance, the $\hat{\Sigma}_2^b$ formulation allows (1) to be expressed as a $\hat{\Sigma}_3^b$ formula, and this seems preferable to the alternative of using a $\hat{\Sigma}_4^b$ formula. More importantly, a statement of the weak pigeonhole principle using the $\hat{\Pi}_2^b$ formulation of "$f(u) = v$" would imply that f must be total, and this would push the strength of the theory well up towards PV_2, which seems undesirably strong.

Thus, we formalize the surjective WPHP for PV_2 functions using the $\hat{\Sigma}_2^b$ formulation of "$f(u) = v$". To do this, we think of f as computed by a polynomial time machine which can make oracle queries to some NP set Ω. (Note that, for simplicity, if the language itself includes symbols for oracles we assume that f accesses those oracles via queries to Ω.) We say that w *is a computation of* $f_e(u)$ provided w encodes in some natural way

(i) a correct computation of the polynomial time machine on input $\langle e, u\rangle$,

(ii) the string of oracle answers obtained during the computation,

(iii) a sequence of witnesses showing correctness of the "Yes" answers to queries to Ω,

and also has the property that

(iv) all the "No" answers to queries to Ω are correct.

It is important to note that even though w includes witnesses for "Yes" oracle answers, the machine computing f does not get to see these witnesses, but only sees the binary Yes/No answers to the oracle queries. It follows that, provably in PV_1, there is at most one possible output of a computation of $f_e(u)$. This lets us use the expression "w is a computation of $f_e(u) = v$" to mean "w is a computation of $f_e(u)$ with output v".

The property "w is a computation of $f_e(u)$" is expressed naturally as a $\hat{\Pi}_1^b$ formula, where the universal quantifier is needed for condition (iv). The principle $\mathrm{sPHP}_a^{t(a)}(f)$ is then defined as the universal closure of

$$\exists v{<}t(a)\, \forall u{<}a\, \forall w\, [\text{“}w \text{ is not a computation of } f_e(u) = v\text{”}].$$

Note that the quantifier $\forall w$ is implicitly bounded, since $|w|$ can be polynomially bounded in terms of $|e|$ and $|u|$. The notation $\mathrm{sPHP}_a^{t(a)}(\mathrm{PV}_2)$ stands for the scheme $\{\mathrm{sPHP}_a^{t(a)}(f) : f \in \mathrm{PV}_2\}$.

The second decision to make when formalizing sWPHP for PV_2 functions in PV_1 is the choice of the size function $t(a)$. As noted in Section 2, PV_1 cannot use iteration to amplify failures of sWPHP even for PV_1 functions, so there is no reason to expect $\mathrm{sPHP}_a^{t(a)}(\mathrm{PV}_2)$ for various choices of t to be equivalent over PV_1. Moreover, in contrast to the case of PV_1 functions, there are no known $\forall\Sigma_1^b$-conservativity results between the different theories $\mathrm{PV}_1 + \mathrm{sPHP}_a^{t(a)}(\mathrm{PV}_2)$.

We keep the notation $\mathrm{sWPHP}(\mathrm{PV}_2)$ for $\mathrm{sPHP}_a^{a(1+1/|a|)}(\mathrm{PV}_2)$, because of the importance of that particular choice of t in formalizing approximate counting. However, we are unable to prove a separation for $\mathrm{PV}_1 + \mathrm{sWPHP}(\mathrm{PV}_2)$, and work with the possibly weaker variant $\mathrm{PV}_1{+}\mathrm{sPHP}_a^{2a}(\mathrm{PV}_2)$ instead. The power of this theory is not very well understood, but it is at least strong enough to prove $\mathrm{iWPHP}(\mathrm{PV}_1)$ in the natural way. Indeed a slightly stronger result holds:

Proposition 14 $\mathrm{PV}_1 + \mathrm{sPHP}_a^{a\#a}(\mathrm{PV}_2) \vdash \mathrm{iWPHP}(\mathrm{PV}_1)$.

Proof It suffices to show that $\mathrm{PV}_1{+}\mathrm{sPHP}_a^{a\#a}(\mathrm{PV}_2)$ proves $\mathrm{iPHP}_a^{a\#a}(\mathrm{PV}_1)$, since amplification will then give $\mathrm{iPHP}_a^{a(1+1/|a|)}(\mathrm{PV}_1)$ over PV_1. We argue inside PV_1, and suppose that $f : [a\#a] \to [a]$ is an injective PV_1 function. We define a new mapping g, essentially equal to the inverse of f. The function g takes as input a value $u < a$ and seeks a value $v < a\#a$ such that $f(v) = u$ by using NP queries to determine bit-by-bit the value of v if such a value v exists. If v exists, $g(u)$ outputs v. Otherwise, $g(u)$ outputs 0. Clearly, g is a PV_2 function. Moreover, given $v < a\#a$ we can use a PV_1 procedure to find $u < a$ and a computation of $g(u) = v$: namely, u is $f(v)$, and a query in the computation of the form "$\exists y{<}a\#a\,(\mathrm{bit}(y,i) = 1 \wedge f(y) = u)$?" gets the answer "Yes" (with v as witness) whenever the i-th bit of v is 1, and the answer "No" otherwise. Thus g is a counterexample to $\mathrm{sPHP}_a^{a\#a}(\mathrm{PV}_2)$. □

Although PV_1 is not strong enough to polynomially amplify PV_2 functions violating the pigeonhole principle, it can compose PV_2 functions and define them by cases. This allows us to prove the following lemma, which will simplify the proof of Theorem 17.

Lemma 15 *Let f and g be PV_2 functions.*

1. $\mathrm{PV}_1 + \neg\mathrm{sPHP}_a^{2a}(f) \vdash \neg\mathrm{sPHP}_a^{4a}(k)$, *for some* PV_2 *function* $k(x) = k(x,a)$.
2. $\mathrm{PV}_1 + \neg\mathrm{sPHP}_{a_1}^{t_1}(f) + \neg\mathrm{sPHP}_{a_2}^{t_2}(g) \vdash \neg\mathrm{sPHP}_{a_1+a_2}^{t_1+t_2}(k)$ *for some* PV_2 *function* $k(x) = k(x, a_1, t_1)$.
3. $\mathrm{PV}_1 + \neg\mathrm{sPHP}_a^{2a}(f) \vdash c > 0 \to \neg\mathrm{sPHP}_{ca}^{2ca}(k)$ *for some* PV_2 *function* $k(x) = k(x, a, c)$.

4. $\mathrm{PV}_1 + \neg\mathrm{sPHP}_a^{2a}(f) \vdash a \leq b \rightarrow \neg\mathrm{sPHP}_b^{2b}(k)$ *for some* PV_2 *function* $k(x) = k(x, a, b)$.

If f and g have extra side parameters, then k also has these parameters.

Proof Note that the PV_2 functions are closed under composition and definition by cases, provably in PV_1. To prove item 1, define k as

$$k(x) \;=\; \begin{cases} f(f(x)) & \text{if } f(x) < a \\ f(f(x) - a) + 2a & \text{otherwise,} \end{cases}$$

where we have suppressed all side parameters of f and k in the notation.

To prove item 2, define k by

$$k(x) \;=\; \begin{cases} f(x) & \text{if } x < a_1 \\ g(x - a_1) + t_1 & \text{otherwise.} \end{cases}$$

Item 3 is proved similarly. To prove item 4, argue in PV_1 to find a value c such that $ca \leq b < 2b \leq 4ca$, then use 1 and 3 to get a surjective mapping from ca to $4ca$. This immediately gives a surjective mapping from b to $2b$. □

We now turn to the unprovability result. In the proof, we will need to run many computations in parallel on one partially defined ordering, and limit the number of computations that find its least element. For this we will use the following lemma.

Lemma 16 *Let X be any set of Turing machines (with parameters) querying oracles $\preceq$ and h on a domain R, with each machine running in time p, where $|R| \geq 8p^2 + 4p + 1$. Let R have some distinguished element d. Furthermore let each machine in X have some positive real number assigned as its weight. Then there exists a set $Y \subseteq X$, a set $S \subseteq R$, a total ordering $\preceq$ on S, and the associated predecessor function h defined on all but the $\preceq$-minimal element of S, such that: Y contains at least weighted fraction $1/2$ of the machines in X; $R \setminus S$ has size at least $|R|/8p$; d is the $\preceq$-maximal element of S; and each machine in Y has its computation completely defined in the partial structure $(\preceq, h)$, and in particular does not query the oracle h on the $\preceq$-minimal element of S.*

Proof Consider the uniform distribution over all pairs $(\preceq', h')$, where $\preceq'$ is a total ordering of R with d as the greatest element and h' is the associated predecessor function. Letting b stand for the $\preceq'$-minimal element, we claim that for a fixed machine A in X, the probability over $(\preceq', h')$ that A queries "$h(b) = ?$" is at most $1/4$.

To see this, argue as follows. Say that $A \in X$ *touches* an element $s \in R$ if, when run on $(\preceq', h')$, it makes a query of the form "$s \preceq s'?$", "$s' \preceq s?$" or "$h(s) = ?$", or makes a query "$h(s^*) = ?$" where s^* is the $\preceq'$-successor

of s. If A ever queries "$h(b) = ?$", then there exists some $i = 1, \ldots, p$ such that the i-th query made by A touches one of the $p - i$ bottom elements of $\preceq'$. By case analysis of the various possible forms of the i-th query, the probability that the i-th query is the *first* query for which this occurs is bounded by $(2p - 2i + 1)/(|R| - 2i - 1)$, which is clearly no more than $(2p+1)/(|R|-1)$. Since A makes at most p queries, the union bound implies that the probability that A queries "$h(b) = ?$" is at most $(2p^2+p)/(|R|-1)$, and thus at most $1/4$ by our assumptions on p and $|R|$.

Therefore, there exists a choice for $\preceq'$ and h' such that the *weighted* fraction of machines in X which query "$h(b) = ?$" is less than $1/4$. Fix some such choice for $\preceq'$ and h', and let X' be the set of machines in X which do not query "$h(b) = ?$".

Each machine can touch at most $2p$ many elements. Hence the average (over $s \in R$) weighted fraction of machines in X' touching an element s is at most $2p/|R|$. Now let S^- be the subset of R of size $|S^-| = |R|/8p$ consisting of those $s \in R$ that are touched by the smallest weighted fraction of the machines in X'. Then the total weighted fraction of machines in X' touching *any* $s \in S^-$ is at most $(2p/|R|)(|R|/8p) = 1/4$.

Take Y to be X' without all the machines A that touch any point in S^-. Take S to be $(R \setminus S^-) \cup \{d\}$, take $\preceq$ to be $\preceq'$ restricted to S, and take h to be the $\preceq$-predecessor function on S without its $\preceq'$-minimal element (note that $h(s)$ will equal $h'(s)$ only if the $\preceq'$-predecessor of s is in S, which will be the case if "$h(s) = ?$" is queried by any machine in Y). $\square$

Before proving the main result of this section, Theorem 17, we outline a proof of a simpler special case to illustrate the main ideas. Suppose for a contradiction that for some particular PV_1 function f and term t,

$$\mathrm{PV}_1 + \mathrm{sPHP}_t^{2t}(f) \vdash \mathrm{HOP}.$$

(Here t and f have access to the size parameter c of HOP, but for tidiness we will usually not write c. The function f has oracle access to both $\preceq$ and h.) The difference between this and the full sPHP_a^{2a} is that we have taken away the universally quantified parameters a and e.

Rearranging, we get that

$$\mathrm{PV}_1 \vdash \forall v{<}2t\, \exists u{<}t\, \exists w\, (w \text{ is a computation of } f(u) = v) \vee \exists z\, \theta(z)$$

where $\exists z\, \theta(z)$ is the sentence HOP for oracles $\preceq$ and h on the domain $[c]$. That is, $\theta(z)$ expresses that z gives a value or tuple of values which falsify one of the conditions 1-3 of Definition 5 of the HOP principle.

In the Student-Teacher game for this formula, the Student is first given a number v. Then in each round the Student either (i) specifies values for u and w, claiming that w is a computation of $f(u) = v$ or (ii) specifies a value for z, claiming that $\theta(z)$. In case (i), the Teacher must reply with a number y

which witnesses that w is not such a computation; otherwise, the Student wins. In case (ii), the Student wins if $\theta(z)$ is true. By Corollary 9, there is a fixed $k \in \mathbb{N}$ and a polynomial p such that Student has a winning strategy, computable in time $p(|c|)$ in each round, which allows him to always win within k rounds against arbitrary Teacher moves.

We may assume, without loss of generality, that the Student algorithm is designed never to output a value z (case (ii) above) without first checking that it satisfies $\theta(z)$. This assumption will allow us to ignore case (ii) from now on, because in our construction we will make sure that the Student never knows any correct value of z. We also assume w.l.o.g. that, before making a move of type (i), the Student is required to check the syntactic correctness of w and the correctness of the witnesses for the "Yes" answers to the NP oracle queries.

Our goal is to find a number $v < 2t$ and definitions of $\preceq$ and h such that the Teacher is able to survive in the game on v for k rounds, meaning that in each round the Teacher finds a witness showing that the Student's computation w is incorrect. We do this by playing $2t$ games in parallel, one for each possible value of $v < 2t$, and constructing $\preceq$ and h in stages. In each stage some subset of the games will advance by at most one round (in which we will make sure that the Student does not win). At the end of the construction we will show that one of the games must have advanced by k rounds, giving a contradiction.

We now formally describe what happens in a stage. At the beginning of each stage, the function h and predicate $\preceq$ are defined on a set $S \subseteq [c]$, so that $\preceq$ is a total ordering of S, and h is the predecessor function on S. Thus, the domain of h is the set S except for its $\preceq$-minimum element. In each of the parallel games, either it is the Student's turn to move next, in which case we say the game is in state (WS) ("Waiting for Student"), or it is the Teacher's turn, which we call (WT) ("Waiting for Teacher"). At the beginning of the construction S is empty and all games are in state (WS).

Each stage proceeds in two steps. In step one, we take the set X of all games in state (WS) and consider the $p(|c|)$-time algorithm computing the Student's strategy for the next round in those games. We use Lemma 16 to extend S in such a way that, for at least half of the games in X, the Student can compute his strategy without finding a witness to HOP and without touching any point outside S. For these games we then actually run the Student's strategy, play the output (u, w) as the Student's move and mark the game as (WT).

In step two, we consider each game that is (now) in state (WT). Let (u, w) be the Student's move in the current round. The Teacher considers whether there is any NP query in the computation w which receives the answer "No" in w, but for which there exists a witness y that shows the answer is "Yes", using only the currently defined values of $\preceq$ and h (on the set S as extended in step one). If there is no such y, then the Teacher does

nothing and the game remains in state (WT). If there is such a y, then the Teacher answers with some such y (say, the least one) and the game advances to state (WS).

The key observation is that at the end of the stage, at most t games can be in state (WT). To see this, for each such game, consider the last move (u, w) made by the Student. We will show that each $u < t$ appears in at most one of these games. For suppose that there exist v, v' in which (u, w) is the Student's last move in the game on v, and (u, w') is the Student's last move in the game on v'. If $v \neq v'$, then the computations w and w' cannot be the same, since w has output v and w' has output v'. So there must be some first NP query which gets different Yes/No answers in w and w'. Since the witness provided by one of them for the "Yes" answer has already been verified by Student using only defined values of $\preceq$ and h, it can be used by the Teacher as a valid counterexample y to the other computation's "No" answer. Thus, if both games remain in the form (WT) at the end of the stage, it must be the case that $v = v'$.

It follows that at the start of each stage, at least half of the games are in state (WS). Therefore, in step one of each stage, at least a quarter of the games advance from state (WS) to state (WT). Hence after $4(k+1)$ stages have been completed, at least one of games must have advanced from state (WS) to state (WT) at least $k+1$ many times. Hence in that game the Teacher was able to survive for k many rounds and we are done.

We now prove the general result.

Theorem 17 *In the relativized language with second order parameters $\preceq$ and h, $\mathrm{PV}_1 + \mathrm{sPHP}_a^{2a}(\mathrm{PV}_2) \nvdash \mathrm{HOP}$.*

Corollary 18 *The theory $\mathrm{PV}_1(\alpha)$+$\mathrm{sPHP}_a^{2a}(\mathrm{PV}_2(\alpha))$ is separated by $\forall\Sigma_1^b(\alpha)$ consequences from $T_2^2(\alpha)$ and $\mathrm{APC}_2(\alpha)$.*

Proof (of Theorem 17) Suppose for a contradiction that PV_1+$\mathrm{sPHP}_a^{2a}(\mathrm{PV}_2)$ does prove HOP. As in the case of more typical versions of WPHP, even though $\mathrm{sPHP}_a^{2a}(\mathrm{PV}_2)$ is technically an infinite family of axioms, it is actually equivalent (already over PV_1) to sWPHP for a single universal $f \in \mathrm{PV}_2$. Therefore, PV_1 proves

$$[\exists a\, \exists e\, \forall v{<}2a\, \exists u{<}a\, \exists w\, (w \text{ is a computation of } f_e(u) = v)] \vee \exists z\, \theta(z),$$

where $\exists z\, \theta(z)$ is the HOP principle. As mentioned above, the existential quantifier $\exists w$ is bounded in terms of a and e; hence by Parikh's theorem, the values of a and e can be bounded in terms of the free variable c. Therefore, using the construction of part 4 of Lemma 15, f can be amplified so as to always have $a = t$ for some fixed term $t(c)$. Thus PV_1 proves

$$[\exists e\, \forall v{<}2t\, \exists u{<}t\, \exists w\, (w \text{ is a computation of } f_e(u) = v)] \vee \exists z\, \theta(z).$$

The generalized Student-Teacher game for this formula consists of the Student in each round specifying either

(i) a value for e,

(ii) values for u and w, together with which pair of earlier e and v values they are associated with, or

(iii) a value for z.

In case (i), the Teacher then specifies a value $v < 2t$ associated with that e. In case (ii), the Teacher must specify a value y that falsifies one of the "No" answers in the computation w, otherwise the Student wins. In case (iii), if $\theta(z)$ is true then the Student wins. By Corollary 9, the Student has a strategy computable in time $p(|c|)$ in each round, which will always win within k rounds against arbitrary Teacher moves, where k is a fixed constant and p is a polynomial.

As before, we may assume that the Student never plays a move of type (iii) without first checking that $\theta(z)$ holds, and as a result our construction will ensure that no move of this type ever occurs. We also assume that, before making a move of type (ii), the Student checks the syntactic correctness of w and the correctness of the witnesses for the "Yes" answers to the NP oracle queries; and that the Student algorithm only queries the oracles h or $\preceq$ when making a move of type (ii), but not when making a move of type (i). Finally, we assume for simplicity that the Student alternates between moves of type (i) and type (ii). So the Student's and Teacher's moves in any partial play of the Student-Teacher game will follow the pattern

$$e_1, v_1, \langle j_1, u_1, w_1\rangle, y_1,\ e_2, v_2, \langle j_2, u_2, w_2\rangle, y_2,\ e_3, v_3, \langle j_3, u_3, w_3\rangle, y_3, \ldots$$

where the values e_i and the triples $\langle j_i, u_i, w_i\rangle$ are the Student's moves, and the values v_i and y_i are the Teacher's moves, and we call each such sequence of four moves a *round*. The triple $\langle j_i, u_i, w_i\rangle$ has $1 \leq j_i \leq i$ and indicates the Student is asserting that w_i is a correct computation of $f(e_{j_i}, u_i) = v_{j_i}$.

All these assumptions can be made without loss of generality by at most doubling the number of Student moves.

We shall prove that no such Student algorithm exists. For this, we choose $c \in \mathbb{N}$ sufficiently large, and construct the function h and the binary predicate $\preceq$ in stages while simultaneously keeping track of many possible plays of the Student-Teacher game. In the earlier proof, we used one function f and ran $2t$ copies of the game in parallel, one for each candidate v for a number outside the range of f. This time we will need to consider $2t$ many values v for *every* parameter e for f proposed by the Student. Hence we will construct a $2t$-branching tree T of possible plays, all working with the same partially defined h and $\preceq$. To get a contradiction it will suffice to show that

the tree has a branch of length $k+1$, since this will describe a play in which the Teacher survives for k rounds.

We now formally describe our construction. The root of the tree T is labeled with the empty string. All other nodes of T will be labeled with certain partial plays of the game. Namely, nodes at depth i correspond to the i-th round of a Student-Teacher game, and are labeled with sequences either of the form

$$e_1, v_1, \langle j_1, u_1, w_1\rangle, y_1,\ e_2, v_2, \langle j_2, u_2, w_2\rangle, y_2, \ldots,\ e_i, v_i \tag{WS}$$

or of the form

$$e_1, v_1, \langle j_1, u_1, w_1\rangle, y_1,\ e_2, v_2, \langle j_2, u_2, w_2\rangle, y_2, \ldots,\ e_i, v_i, \langle j_i, u_i, w_i\rangle \tag{WT}$$

Leaves of T labelled with sequences of type (WS) correspond to the situation where the Student must calculate a value for $\langle j_i, u_i, w_i\rangle$ as his next move. Leaves labelled with sequences of type (WT) correspond to the situation where the Teacher should provide a counterexample to the correctness of the computation w_i as her next move. Internal nodes of T will always have labels of type (WT). As usual, the sequence labeling a node is an initial subsequence of the labels of its children.

The *weight* of a node at depth i is defined to equal $1/(2t)^i$. Consequently, the total weight of the leaves of T always equals 1.

Initially h and $\preceq$ are completely undefined. To initialize the tree T, run the Student algorithm on input c and let it produce a value e_1. (This is possible since the Student does not query h or $\preceq$ when computing e_i values.) Then let T be the tree of height 1 in which the root has $2t$ many children, each labelled with a sequence e_1, v_1 — one for each value $v_1 < 2t$. In each later stage of the construction we will extend some subset of the leaves of T and the height of T will increase by at most 1.

At each stage, the function h and predicate $\preceq$ will be defined on a set $S \subseteq [c]$, so that $\preceq$ is a total ordering of S, and h is the predecessor function on S without its $\preceq$-minimal element. Let R be the complement of S, namely $R = [c] \setminus S$. Initially, R is all of $[c]$. After the ℓ-th stage of the construction, the size of R will be at least $c/(8p(|c|))^\ell$, where $p(|c|)$ is the running time of the Student's strategy. The construction will run for at most $(k+1)2^{k+1}$ stages, so since k is a fixed constant, $|R| \gg 0$ always holds (for c chosen sufficiently large).

Each stage proceeds in two steps. In step one, we take the set X of all leaves of type (WS) and consider the $p(|c|)$-time algorithm computing the Student's strategy for his next move in the corresponding games. Take all elements of R to be $\preceq$-smaller than all elements of S, let d be the $\preceq$-minimal point of S, and apply Lemma 16 with $R \cup \{d\}$ in the role of R and d as the distinguished element. The lemma allows us to extend S in such a way that, for a set Y of at least half the leaves in X by weight, the Student

can compute his strategy without witnessing HOP and without touching any point outside S. For each leaf in Y, we then actually run the Student's strategy and add the output $\langle j_i, u_i, w_i \rangle$ to the label as the Student's move, so that the leaf is now (WT).

In step two, we consider each leaf m that is in state (WT) (either as a result of step one, or left over from some earlier stage). Let $\langle j_i, u_i, w_i \rangle$ be the Student's last move in the current round. The Teacher considers whether there is any NP query in the computation w_i which receives the answer "No" in w_i, but for which there exists a witness y that shows the answer is "Yes", using only the currently defined values of $\preceq$ and h (on the set S as extended in step one). If there is no such y, then the Teacher does nothing and m remains in state (WT). If there is such a y, then the following happens. The Teacher picks the least such y, and plays it in the game as y_i. We run the Student algorithm to compute a value e_{i+1} for the Student's next move. Then for each possible value $v_{i+1} < 2t$ for the Teacher's next move, we add a new leaf to T as a child of m and label it with the label of m followed by y_i, e_{i+1}, v_{i+1}, so that it is in state (WS).

For the next two claims, consider the tree at the beginning of an arbitrary stage ℓ in the construction.

Claim 1 Let m be any internal node. Let i be the depth of m in the tree. Consider any pair n, n' of children of m, where the nodes n and n' are at depth $i+1$, and were given labels ending with e_{i+1}, v_{i+1} and e_{i+1}, v'_{i+1}, respectively, when they were created. Suppose that n is, or has as a descendant, a (WT) leaf node q with a label ending with $\langle i+1, u, w \rangle$ and that n' is, or has as a descendant, a (WT) leaf node q' with a label ending with $\langle i+1, u, w' \rangle$. Then $n = n'$.

To prove Claim 1, observe that the label of q asserts that w is a computation of $f_{e_{i+1}}(u) = v_{i+1}$, whereas the label of q' asserts that w' is a computation of $f_{e_{i+1}}(u) = v'_{i+1}$. As in the simple case, since both labels are (WT), this must mean that $v_{i+1} = v'_{i+1}$, as otherwise the Teacher would have been able to give a counterexample to one of these computations. Hence $n = n'$.

Claim 2 Let d be the height of the tree T. Then at most weighted fraction $1 - 1/2^d$ of the leaves of T have the form (WT) .

To prove Claim 2, define a node n at depth $i+1$ in T to be "internally (WT) with respect to u" if n is, or has as a descendant, a (WT) leaf node q with a label ending with $\langle i+1, u, w \rangle$, for some w. Let m be the parent of n. By Claim 1, for each $u < t$ the node m has at most one child that is internally (WT) with respect to u. Hence at most half of the children of m can be internally (WT) (with respect to any u). An easy induction on j now shows that at most weighted fraction $1 - 1/2^j$ of the tree is covered by internally (WT) nodes lying at depth $\leq j$. Since T has height d, and since every (WT) leaf must be covered by some internally (WT) node, Claim 2

holds.

By Claim 2 at least weighted fraction $1/2^d$ leaves of T are (WS) at the beginning of stage ℓ. Therefore in step one of stage ℓ at least weighted fraction $1/2^{d+1}$ of leaves advance from state to (WS) to state (WT).

Now let C_ℓ be the weighted sum over the leaves of the number of advancements from state (WS) to state (WT) recorded in the label of each leaf. As long as T has height $d \leq k$, we have shown that C_ℓ increases by at least $1/2^{k+1}$ in each stage. Therefore after $(k+1)2^{k+1}$ stages, C_ℓ is at least $k+1$. Thus, at least one leaf records a game which advanced from (WS) to (WT) at least $k+1$ times. Hence this branch of T has height at least $k+1$, and we have found a play where the Teacher survives for k rounds. □

We remark that this argument could be simplified in some respects if we were working with the weaker principle $\mathrm{sPHP}_a^{a^2}$ rather than sPHP_a^{2a}. On the other hand, we have been unable to make it work for the stronger principle $\mathrm{sPHP}_a^{a(1+1/|a|)}$. Roughly speaking, the difficulty is that the construction in this case would need too many stages and eventually the set R where $\preceq$ is undefined would become empty.

Given the prominent role played by $\mathrm{sPHP}_a^{a(1+1/|a|)}$ in [14, 16], it is an interesting open problem whether $\mathrm{PV}_1 + \mathrm{sPHP}_a^{a(1+1/|a|)}(\mathrm{PV}_2)$ proves HOP.

5 $T_2^1 + \mathrm{sWPHP}(\mathrm{PV}_1)$

This section describes three closely related sufficient conditions for proving a separation from $T_2^1 + \mathrm{sWPHP}(\mathrm{PV}_1)$; firstly in terms of what we might call "random PLS problems", and then in terms of "random refutations" in narrow resolution and in treelike Res(log).

Throughout, let c be a first order parameter and let S be a tuple of function and relation symbols which we will interpret as living on the domain $[c]$. Fix a $\hat{\Sigma}_1^b$ formula $\exists x{<}q\, \theta(c,x)$, where θ is a PV_1 predicate with oracles for S.

Lemma 19 *Suppose $T_2^1 + \mathrm{sWPHP}(\mathrm{PV}_1) \vdash \exists x{<}q\, \theta(c,x)$. Let r be any term (in c). Then there is a term t and a function g such that*

$$T_2^1 \vdash \forall v{<}rt\, \exists u{<}t\, \exists x{<}q\, [g(u){=}v \vee \theta(c,x)].$$

Proof We are given that

$$T_2^1 + \forall a\, \exists v{<}a^2\, \forall u{<}a\, f_a(u){\neq}v \vdash \exists x{<}q\, \theta(c,x).$$

Moving the instance of sWPHP to the right hand side and applying Parikh's theorem, this gives

$$T_2^1 \vdash \exists a{<}t'\, \forall v{<}a^2\, \exists u{<}a\, \exists x{<}q\, [f_a(u){=}v \vee \theta(c,x)]$$

for some term t'. The lemma then follows by amplification and elimination of parameters for sWPHP (see e.g. Lemma 2.3 of [34]). □

5.1 Random PLS problems

Recall that a PLS problem (C, N) witnesses $\exists x{<}q\,\theta(c, x)$ if there is some oracle-free polynomial time function F (which can, however, take c as a parameter) such that, if s is any solution of the PLS problem (C, N), then $F(s) < q$ and $\theta(c, F(s))$ holds. The next lemma follows directly from Lemma 19 and the PLS witnessing theorem for T^1_2 [8].

Lemma 20 *Suppose $T^1_2 + \mathrm{sWPHP}(\mathrm{PV}_1) \vdash \exists x{<}q\,\theta(c, x)$. Let r be any term. Then there is a term t and a parametrized family of PLS problems (C_v, N_v) (which also take c as a hidden parameter) such that, for all but t of the possible choices of parameter v from the interval $[rt]$, the PLS problem (C_v, N_v) witnesses $\exists x{<}q\,\theta(c, x)$.* □

The following is obtained by an averaging argument.

Corollary 21 *Suppose that, for each c, A_c is a probability distribution of oracles in the language S on the domain $[c]$. Let r be any term. If $T^1_2 + \mathrm{sWPHP}(\mathrm{PV}_1) \vdash \exists x{<}q\,\theta(c, x)$ then there is a (non-uniform in c) PLS problem which witnesses $\exists x{<}q\,\theta(c, x)$ for a random oracle $\alpha \in A_c$ with probability $1 - 1/r$.* □

Recall that it is straightforward to show that certain sentences cannot be witnessed by a PLS problem that works for *every* oracle. We give an example of the method in the case where S contains a single function α (represented by a relation coding its bit-graph), and $\exists x{<}q\,\theta(c, x)$ is the injective WPHP, asserting that α is not an injection from $[c^2]$ to $[c]$. Another example, for the HOP principle, is implicit in our proof of Theorem 10.

Given our PLS problem (C, N), look at the set of pairs (β, s) where β is a small partial injection from $[c^2]$ to $[c]$ such that the computation of $C^\beta(s)$ is defined, and choose a pair for which $C^\beta(s)$ is minimal. Then we may extend β to a small partial injection β' such that $C^{\beta'}(N^{\beta'}(s))$ is defined, and then arbitrarily to a total function α, and it must be the case that s is a solution to our PLS problem for α. But on the other hand, s does not contain any information about a witness to the injective WPHP.

This contradicts the claim that the PLS problem works for every oracle. Note, however, that it does not contradict the claim that it works for *most* oracles, if "most" means "for all but a fraction 2^{-k} of oracles" for k polynomial in $|c|$, as it does in Corollary 21 above. This is because (depending on the distribution chosen) typically only a fraction 2^{-l} of oracles extend the partial injection β', for some l polynomial in $|c|$, and it may be that $l > k$ and so we have no guarantee that the claim says anything about any oracle α extending β'.

5.2 Random resolution refutations

In this subsection, we make the (inessential) assumption that S contains no function symbols. We also assume some familiarity with resolution and related propositional proof (or rather refutation) systems. All size descriptions such as "quasipolynomial" and "polylogarithmic" are with respect to the size parameter c.

By the *width* of a resolution refutation we will mean the number of literals in the largest clause. We will say that a refutation is *narrow* if it has polylogarithmic width, and that a CNF is narrow if every clause in it has polylogarithmic width. Note that by collapsing together identical clauses, we may make any narrow resolution refutation into one of quasipolynomial size.

Definition 22 *Given a value for the parameter c, the propositional translation $\langle \forall x{<}q\, \neg\theta(c,x)\rangle$ of the negation of $\exists x{<}q\, \theta(c,x)$ is constructed as follows: $\theta(c,x)$ can be computed by a polylogarithmic depth decision tree T_x which queries a relation in S at each internal node, and is labelled with "accept" or "reject" at each leaf. For each branch b of T_x, let C_b be the conjunction of the oracle replies along that branch. Let A_x be the set of accepting branches of T_x. Then $\langle \forall x{<}q\, \neg\theta(c,x)\rangle$ is defined to be $\bigwedge_{x<q} \bigwedge_{b\in A_x} \neg C_b$.*

Notice that $\langle \forall x{<}q\, \neg\theta(c,x)\rangle$ is a narrow CNF.

Proposition 23 *Suppose that $T^1_2 \vdash \exists x{<}q\, \theta(c,x)$. Then $\langle \forall x{<}q\, \neg\theta(c,x)\rangle$ has a a quasipolynomial size, treelike* Res(log) *refutation, and also a narrow resolution refutation.*

Proof The translation into treelike Res(log) is due to Krajíček [21].

The translation into narrow resolution can be derived straightforwardly from the PLS witnessing theorem for T^1_2 as follows: consider a game played between a Prover and an Adversary, where the Adversary claims to know an oracle falsifying $\exists x{<}q\, \theta(c,x)$ and the Prover tries to force the Adversary into a contradiction by making oracle queries. A PLS problem witnessing $\exists x{<}q\, \theta(c,x)$ can be made into a winning strategy for the Prover in which the Prover only needs to remember polylogarithmically many (i.e. polynomially many in $|c|$) bits of the oracle at any given time; and such a strategy is exactly the dual of a narrow resolution refutation of $\langle \forall x{<}q\, \neg\theta(c,x)\rangle$.

Alternatively, standard arguments in propositional proof complexity about manipulating treelike proofs (see, for example, [3] or [19]) can be carefully applied to show that a narrow CNF has a quasipolynomial size, treelike Res(log) refutation if and only if it has a narrow resolution refutation [23]. □

The next two definitions are essentially due to Stefan Dantchev [personal communication]. We state them for resolution, but they also work for treelike Res(log).

Definition 24 *An* ε-random resolution refutation distribution *for a narrow CNF* A *is a probability distribution of pairs* (B, Π)*, where*

1. B *is a narrow CNF in the same propositional variables as* A*,*
2. *for every truth assignment, a randomly chosen* B *is true with probability at least* $1 - \varepsilon$*;*
3. Π *is a resolution refutation of* $A \wedge B$*.*

We will call the formulas B auxiliary formulas*. The size of a random resolution refutation is defined to be the maximum of the sizes of the refutations* Π *in the distribution; similarly for the width.*

One way to understand this definition intuitively is as follows. Given a formula A and a truth assignment α, a resolution refutation Π of A can be used as a tool to find a false clause in A, by starting at the final, empty clause in Π and following a path upwards through false clauses, since if the conclusion of a resolution rule was false then one of the premises must be false. With a random refutation distribution, this procedure will work (that is, find a false clause in A) with high probability over the choice of (B, Π), since with high probability all the clauses in B are true in α, which means that our path upwards through false clauses will necessarily end with some clause from A.

If a CNF has an ε-random refutation distribution in the sense of Definition 24, then by an averaging argument it will also have an "ε-random refutation", with the same bounds on size and width, in the sense of the following definition:

Definition 25 *An* ε-random resolution refutation *of a narrow CNF* A *is a resolution refutation of* $A \wedge B$ *where* B *is a narrow CNF in the same propositional variables as* A *with the property that under a random truth assignment,* B *is true with probability at least* $1 - \varepsilon$*.*

Notice that such a refutation is not necessarily sound, in that there exist CNFs with a small number of satisfying assignments that nevertheless have random resolution refutations.

Theorem 26 *Suppose* $T^1_2 + \mathrm{sWPHP}(\mathrm{PV}_1) \vdash \exists x{<}q\, \theta(c, x)$*. Let* r *be any term. Then* $\langle \forall x{<}q\, \neg\theta(c, x)\rangle$ *has a* $1/r$*-random narrow resolution refutation distribution, and hence a* $1/r$*-random narrow resolution refutation. Equivalently, it has a quasipolynomial size* $1/r$*-random treelike* Res(log) *refutation distribution, and hence a quasipolynomial size* $1/r$*-random resolution refutation.*

Proof By Lemma 19, there is a term t and a function g such that

$$T_2^1 \vdash \forall v{<}rt\,[\exists u{<}t\, g(u){=}v \vee \exists x{<}q\, \theta(c,x)].$$

Hence for each $v < rt$, the negation of the bracketed formula has a narrow resolution refutation. Furthermore, if we fix any oracle and choose such a $v < rt$ uniformly at random, then with probability at least $1 - 1/r$ the formula $\exists u{<}t\, g(u){=}v$ is false. So we can take our auxiliary formulas B_v in Definition 24 to be $\langle \forall u{<}t\, g(u){\neq}v\rangle$ with v chosen at random in $[rt]$. □

We remark that, using essentially the same proof, we can get a little more information by letting r be a free variable rather than a fixed term. Then from the same assumption as the theorem, we get that for all c and r the formula $\langle \forall x{<}q\, \neg\theta(c,x)\rangle$ has a $1/r$-random resolution refutation of width polynomial in $|c|$ and $|r|$, or a $1/r$-random treelike Res(log) refutation of size quasipolynomial in c and r.

Theorem 26 motivates the following open problem.

Open Problem 27 *Find a family of narrow CNFs which have quasipolynomial size constant depth Frege refutations, but which do not have* $1/r$*-random narrow resolution refutations, for some quasipolynomial term* r*.*

Naturally, the problem could be stated in terms of quasipolynomial size treelike Res(log) refutations instead of narrow resolution. We are able to prove a lower bound for a weak version of this question, for treelike resolution rather than treelike Res(log), by the following easy argument. Let A be the bit-graph version of the injective WPHP from $2c$ pigeons to c holes, where $c = 2^k$. That is, for each pigeon i there are k variables $p_1^i, \ldots, p_k^i$ expressing, in binary, the number of the hole that pigeon i goes to; and F consists simply of $\binom{2c}{2}c$ clauses $C^j_{i_1 i_2}$, each of size $2k$, asserting that pigeons i_1 and i_2 do not both go to hole j. Suppose that B is a narrow auxiliary CNF in these variables that is true with probability greater than $3/4$ under the uniform random assignment and suppose that we have a treelike resolution refutation Π of $A \wedge B$.

Choose a small $\delta > 0$. Starting from the final, empty clause of Π, choose a path up through Π by, at each resolution step, setting the resolved variable at random and moving to the premise that is falsified by the assignment that is being built. Continue until you have either set kc^δ variables or reached (and falsified) one of the clauses in $A \wedge B$.

Falsifying a clause from A requires setting all of the bits of two pigeons in a way that makes them collide. But the procedure above can set the bits of no more than c^δ pigeons, so the probability of a collision is very small, certainly less than $1/4$ for suitable δ. We also know that the probability of falsifying a clause from B is less than $1/4$.

Hence with probability at least 1/2 the procedure runs for the full kc^δ steps. Therefore there are at least $2^{kc^\delta}/2$ distinct paths up through Π from the empty clause. Hence $|\Pi| > 2^{kc^\delta - 1}$.

References

[1] A. Beckmann and S. Buss. Polynomial local search in the polynomial hierarchy and witnessing in fragments of bounded arithmetic. *Journal of Mathematical Logic*, 9(1):103–138, 2009.

[2] A. Beckmann and S. Buss. *Characterization of Definable Search Problems in Bounded Arithmetic via Proof Notations*, pages 65–134. Ontos Verlag, 2010.

[3] E. Ben-Sasson and A. Wigderson. Short proofs are narrow - resolution made simple. *Journal of the ACM*, 48(2):149–169, 2001.

[4] S. Boughattas and L. A. Kołodziejczyk. The strength of sharply bounded induction requires MSP. *Annals of Pure and Applied Logic*, 161:504–510, 2010.

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

[6] S. Buss. Axiomatizations and conservation results for fragments of bounded arithmetic. In *Logic and Computation, Proceedings of a Workshop held at Carnegie Mellon University*, pages 57–84. AMS, 1990.

[7] S. Buss. First-order proof theory of arithmetic. In S. Buss, editor, *Handbook of Proof Theory*. Elsevier, 1998.

[8] S. Buss and J. Krajíček. An application of Boolean complexity to separation problems in bounded arithmetic. *Proceedings of the London Mathematical Society*, 69:1–21, 1994.

[9] M. Chiari and J. Krajíček. Witnessing functions in bounded arithmetic and search problems. *Journal of Symbolic Logic*, 63(3):1095–1115, 1998.

[10] M. Chiari and J. Krajíček. Lifting independence results in bounded arithmetic. *Archive for Mathematical Logic*, 38(2):123–138, 1999.

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

[12] E. Jeřábek. Dual weak pigeonhole principle, Boolean complexity, and derandomization. *Annals of Pure and Applied Logic*, 129:1–37, 2004.

[13] E. Jeřábek. The strength of sharply bounded induction. *Mathematical Logic Quarterly*, 52(6):613–624, 2006.

[14] E. Jeřábek. Approximate counting in bounded arithmetic. *Journal of Symbolic Logic*, 72(3):959–993, 2007.

[15] E. Jeřábek. On independence of variants of the weak pigeonhole principle. *Journal of Logic and Computation*, 17(3):587–604, 2007.

[16] E. Jeřábek. Approximate counting by hashing in bounded arithmetic. *Journal of Symbolic Logic*, 74(3):829–860, 2009.

[17] L. Kołodziejczyk, P. Nguyen, and N. Thapen. The provably total NP search problems of weak second-order bounded arithmetic. *Annals of Pure and Applied Logic*, 162(2), 2011.

[18] J. Krajíček. No counter-example interpretation and interactive computation. In Y.N. Moschovakis, editor, *Logic from Computer Science*, volume 21 of *Mathematical Sciences Research Institute Publ.*, pages 287–293. Springer, 1992.

[19] J. Krajíček. Lower bounds to the size of constant-depth propositional proofs. *Journal of Symbolic Logic*, 59(1):73–86, 1994.

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

[21] J. Krajíček. On the weak pigeonhole principle. *Fundamenta Mathematicae*, 170(1-3):123–140, 2001.

[22] J. Krajíček, P. Pudlák, and G. Takeuti. Bounded arithmetic and the polynomial hierarchy. *Annals of Pure and Applied Logic*, 52:143–153, 1991.

[23] M. Lauria. Short Res*(polylog) refutations if and only if narrow Res refutations. Typeset manuscript, available at www.dsi.uniroma1.it/~lauria/, 2011.

[24] 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.

[25] R. Parikh. Existence and feasibility in arithmetic. *Journal of Symbolic Logic*, 36(3):494–508, 1971.

[26] 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.

[27] P. Pudlák. 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, pages 308–317. Springer, 1991.

[28] P. Pudlák. Some relations between subsystems of arithmetic and the complexity of computations. In *Logic From Computer Science: Proceedings of a Workshop held November 13-17, 1989, Mathematical Sciences Research Institute Publication #21*, pages 499–519. Springer-Verlag, 1992.

[29] P. Pudlák. Consistency and games - in search of new combinatorial principles. In V. Stoltenberg-Hansen and J. Väänänen, editors, *Logic Colloquium '03*, number 24 in Lecture Notes in Logic, pages 244–281. ASL, 2006.

[30] P. Pudlák and N. Thapen. Alternating minima and maxima, Nash equilibria and bounded arithmetic. Typeset manuscript, available at www.math.cas.cz/~thapen/, 2009.

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

[32] N. Segerlind, S. Buss, and R. Impagliazzo. A switching lemma for small restrictions and lower bounds for k-DNF resolution. *SIAM Journal on Computing*, 33(5):1171–1200, 2004.

[33] A. Skelley and N. Thapen. The provably total search problems of bounded arithmetic. *Proceedings of the London Mathematical Society*, 103(1):106–138, 2011.

[34] N. Thapen. A model-theoretic characterization of the weak pigeonhole principle. *Annals of Pure and Applied Logic*, 118:175–195, 2002.

[35] N. Thapen. Structures interpretable in models of bounded arithmetic. *Annals of Pure and Applied Logic*, 136(3):247–266, 2005.