The Proof Analysis Problem*

Noel Arteche† Albert Atserias‡ Susanna F. de Rezende§ Erfan Khaniki¶

Abstract

Atserias and Müller (*JACM*, 2020) proved that for every unsatisfiable CNF formula φ, the formula $\textsc{Ref}(\varphi)$—stating that "φ has small Resolution refutations"—does not have subexponential-size Resolution refutations. Conversely, when φ is satisfiable, Pudlák (*TCS*, 2003) showed how to construct a polynomial-size Resolution refutation of $\textsc{Ref}(\varphi)$ given a satisfying assignment of φ. A question that had remained open is: *do all short Resolution refutations of* $\textsc{Ref}(\varphi)$ *explicitly leak a satisfying assignment of* φ*?*

We answer this question affirmatively by providing a polynomial-time algorithm that extracts a satisfying assignment for φ given any short Resolution refutation of $\textsc{Ref}(\varphi)$. The algorithm follows from a new feasibly constructive proof of the Atserias–Müller lower bound, formalizable in Cook's theory PV_1 of bounded arithmetic. This implies that Extended Frege can efficiently prove (a suitable formalization of the statement) that automating Resolution is **NP**-hard.

Motivated by this algorithm, we introduce a new meta-computational problem concerning Resolution lower bounds: the *Proof Analysis Problem* (PAP). For a fixed proof system Q, the Proof Analysis Problem for Q asks, given a CNF formula φ and a Q-proof of a Resolution lower bound for φ, encoded as $\neg\textsc{Ref}(\varphi)$, whether φ is satisfiable. In contrast to the Proof Analysis Problem for Resolution, which is in **P**, we prove that PAP for Extended Frege (EF) is **NP**-complete. In particular, EF can prove Resolution lower bounds on satisfiable formulas without necessarily revealing a satisfying assignment.

Our results yield new insights into proof search and the meta-mathematics of Resolution lower bounds: (i) for every proof system that simulates EF as well as for Resolution, the system is (weakly) automatable if and only if it can be (weakly) automated exclusively on formulas stating Resolution lower bounds; (ii) we provide explicit Ref formulas that are exponentially hard for bounded-depth Frege systems; and (iii) for every strong enough theory of arithmetic T we construct explicit unsatisfiable CNF formulas that are exponentially hard for Resolution but for which T cannot prove even a quadratic Resolution lower bound. This latter result applies to arbitrarily strong theories like PA or ZFC, and does not require any complexity-theoretic assumptions.

*version: June 19, 2025

†Lund University and University of Copenhagen, noel.arteche@cs.lth.se

‡Universitat Politècnica de Catalunya and Centre de Recerca Matemàtica atserias@cs.upc.edu

§Lund University, susanna.rezende@cs.lth.se

¶University of Oxford, erfan.khaniki@cs.ox.ac.uk

Contents

1 Introduction

The most natural computational problem arising in proof complexity is that of *proof search*: what is the complexity of finding proofs? In the late 90s, the notion of *automatability*, defined by Bonet, Pitassi, and Raz [BPR00], emerged as a central concept in the theory of propositional proof complexity. A proof system Q is *automatable* if there is a deterministic algorithm that finds a Q-proof of a formula φ in time polynomial in the shortest one available. Except for Tree-like Resolution, which is automatable in quasi-polynomial time [BP96], no other non-trivial proof system is known to be automatable in polynomial or quasi-polynomial time.

Krajíček and Pudlák [KP98] and Bonet, Pitassi, and Raz [BPR00] proved that under standard worst-case number-theoretic assumptions in cryptography, strong proof systems like $\mathbf{TC}^0$-Frege and Extended Frege are not automatable. These results can be transferred to $\mathbf{AC}^0$-Frege under slightly stronger hardness assumptions [BDG+04], but it seems hard to push them further. Essentially, their proof techniques require some amount of basic number theory to be formalized in the system, something that is likely unworkable for Resolution. Since then, efforts focused on showing the hardness of automating Resolution and related weak systems [Pud03; AR08; GL10; AM11; HP11; Ats13; BPT14; MPW19], culminating in the final answer by Atserias and Müller [AM20], who proved that Resolution is not automatable unless $\mathbf{P} = \mathbf{NP}$. This is the optimal hardness assumption since $\mathbf{P} = \mathbf{NP}$ implies the automatability of any proof system.

The technique used in [AM20] relies on the insight that Resolution *cannot reason about its own lower bounds*. To every CNF formula φ, they associate a new formula $\textsc{Ref}_s(\varphi)$ that encodes the statement "there is a size-s Resolution refutation of φ". As a tautology, $\neg\textsc{Ref}_s(\varphi)$ is a natural propositional encoding of a Resolution lower bound. (We postpone to the preliminaries the details of the encoding of the Ref formula we use, where we also discuss previously studied variations.)

Pudlák [Pud03] had shown already in 2003 that whenever φ is satisfiable, the formula $\textsc{Ref}_s(\varphi)$ is easily refutable by Resolution. On the other hand, Atserias and Müller [AM20] proved that whenever φ is unsatisfiable, Resolution will require exponential size to refute $\textsc{Ref}_s(\varphi)$, for s being some fixed polynomial in the number of variables of φ, which we omit in the subscript for the rest of this introduction for the sake of clarity.

As a consequence, an automating algorithm running on formulas of the form $\textsc{Ref}(\varphi)$ can be used to decide SAT in polynomial time: if $\varphi \in \text{SAT}$, then the algorithm must find a short refutation of $\textsc{Ref}(\varphi)$ that Pudlák guarantees must exist; on the other hand, if $\varphi \notin \text{SAT}$, then there are no short refutations of $\textsc{Ref}(\varphi)$, so we can stop the automating algorithm after a polynomial number of steps and be certain that φ is unsatisfiable.

The proof strategy behind the Resolution lower bound on Ref formulas was soon adapted to a variety of weak proof systems (those where size lower bounds are known), although the Ref-like formulas used in these spin-off results are no longer natural lower-bound statements for these systems. In general, as pointed out by Pudlák, the question of whether a proof system can prove any of its own lower bounds "is widely open, except for Resolution, and we consider it more important than automatability" [Pud20, p. 3]. It is currently open, for example, whether systems like constant-depth Frege have polynomial-size proofs of any of their own lower bounds.

The feat of the Resolution lower bound on Ref formulas, combined with the upper bound for satisfiable formulas, implies that Resolution can *only* reason about "trivial" Resolution lower bounds (i.e., lower bounds on satisfiable formulas, which do not have refutations of any size). This highlights the upper bound construction as something even more remarkable, given that Resolution cannot efficiently argue about its own soundness [AB04]. Intriguingly, the known upper bound for $\textsc{Ref}(\varphi)$ for satisfiable φ crucially relies on Resolution guessing a satisfying assignment and using it as the backbone of the refutation. It is then natural to ask whether this is necessary:

(Q_1) *Is it the case that whenever there is a short Resolution refutation of* $\textsc{Ref}(\varphi)$*, the proof must* leak *a satisfying assignment?*

By "leaking" we mean that a satisfying assignment is always readable in polynomial time from the given refutation. It is important to note that given a refutation π of $\textsc{Ref}(\varphi)$, one cannot simply restrict π in a way that corresponds to $\textsc{Ref}(\varphi_{\restriction x_1=0})$ and $\textsc{Ref}(\varphi_{\restriction x_1=1})$ to extract a satisfying assignment. This is because the variables of φ are *not* variables of $\textsc{Ref}(\varphi)$. In principle, such a self-reducibility trick seems to require access to an automating algorithm, so that one could successively look for refutations of $\textsc{Ref}(\varphi_{\restriction x_1=0})$ or $\textsc{Ref}(\varphi_{\restriction x_1=1})$, then $\textsc{Ref}(\varphi_{\restriction x_1=b_1,x_2=0})$ or $\textsc{Ref}(\varphi_{\restriction x_1=b_1,x_2=1})$, and so on for all variables. Without access to an automating algorithm, it is not at all clear whether satisfying assignments can be extracted efficiently.

Yet another way of phrasing the lower bound on Ref formulas is to see it as the correctness proof of a *lower bound analysis algorithm*. Namely, the result proves that there is an algorithm that given a Resolution refutation π of $\textsc{Ref}(\varphi)$ decides whether φ is satisfiable. The algorithm consists simply of checking whether π is correct and short enough. The correctness of this procedure requires the proof of the lower bound, and this framing naturally leads to the following second natural question regarding Ref formulas:

(Q_2) *Is there an algorithm that given an Extended Frege proof* π *of a Resolution lower bound* $\neg\textsc{Ref}(\varphi)$ *decides in polynomial time whether* φ *is satisfiable?*

If the answer were affirmative, this would settle the long-standing open problem of the **NP**-hardness of automating Extended Frege: given a CNF formula φ, construct the formula $\textsc{Ref}(\varphi)$ and run the automating algorithm to find a short Extended Frege refutation. If an algorithm as the one asked for in (Q_2) existed, then we could apply it on this refutation to *analyze* whether φ is satisfiable. This distills the main idea in [AM20], and the framing of the question in terms of algorithm design suggests that such an algorithm might well be possible *without the need for unconditional Extended Frege lower bounds*.

Overall, the two questions (Q_1) and (Q_2) above hint at the central role of meta-mathematical lower bound statements in the theory of proof search. We believe this calls for a deeper structural understanding that could lead to much-needed conceptual insights in automatability.

1.1 Contributions

Motivated by questions (Q_1) and (Q_2) above, we introduce a new meta-computational problem relating proofs and computation: the *Proof Analysis Problem*.

For every propositional proof system Q, the *Proof Analysis Problem for* Q (PAP_Q) consists in analyzing Resolution lower bounds proven by Q. More formally, given a CNF formula φ and a Q-proof of the Resolution lower bound encoded by the formula $\neg\textsc{Ref}(\varphi)$, the task is to decide whether φ is satisfiable.

The Proof Analysis Problem for Q (PAP_Q)	
Input	A CNF formula φ, a size parameter s in unary and a Q-proof π of the formula $\neg\textsc{Ref}_s(\varphi)$.
Output	Is φ satisfiable?

The problem can be seen as the computational task of distinguishing "true" Resolution lower bounds (those where φ is actually unsatisfiable) from "trivial" ones (those where the lower bound trivially holds because φ is satisfiable and there is therefore no Resolution refutation, of any size). For those proof systems for which $\text{PAP}_Q \in \mathbf{P}$, we say that Q is *analyzable*. We remark that the Ref formula in the definition of PAP_Q is always referring to Resolution refutations, while the proof system Q where $\textsc{Ref}(\varphi)$ is being derived can be arbitrarily strong.

For the case of Resolution itself, the problem $\mathrm{PAP}_{\mathrm{Res}}$ is easy to compute thanks to the lower bound on $\textsc{Ref}$ formulas [AM20]: if π is a correct refutation of $\textsc{Ref}(\varphi)$ and it is small, then φ must be satisfiable. Until now, however, to the best of our knowledge this was the extent of what could be said about PAP-like problems. In particular, we are not aware of any other upper or lower bounds on this problem for proofs systems other than Resolution.

In the language of PAP, questions (Q_1) and (Q_2) can be neatly rephrased as follows:

(Q_1) Does $\mathrm{PAP}_{\mathrm{Res}}$ admit a search-to-decision reduction?

(Q_2) Is Extended Frege analyzable? Namely, is $\mathrm{PAP}_{\mathsf{EF}}$ in $\mathbf{P}$?

In this work we kick-start the systematic study of these Proof Analysis Problems and settle questions (Q_1) and (Q_2) above. This in turn yields a series of interesting consequences for the meta-mathematics of proof complexity lower bounds as well as proof search. We outline our results next.

1.1.1 An algorithm for assignment extraction

On the topic of question (Q_1), our main result is that the search version of $\mathrm{PAP}_{\mathrm{Res}}$ can be solved deterministically in polynomial time.

Theorem 1.1 (Assignment extraction algorithm, informal)**.** *The search version of the Proof Analysis Problem for Resolution can be solved in deterministic polynomial time whenever the size parameter s is at least n^3. That is, there is an algorithm that, given a CNF formula φ over n variables and $\mathrm{poly}(n)$ clauses and a Resolution refutation π of $\textsc{Ref}_s(\varphi)$ with $s \geq n^3$, extracts a satisfying assignment for φ in time polynomial in n, s and the size $|\pi|$ of π, whenever φ is satisfiable.*

The question can be stated more formally in terms of *Levin reductions*. A Levin reduction between search problems R_1 and R_2 is a Karp-style many-one reduction that maps instances of R_1 to instances of R_2, with the additional property that it also maps solutions of R_1 to solutions of R_2, and back. The reduction $\varphi \mapsto \textsc{Ref}(\varphi)$ showing that SAT reduces to the Proof Size Problem for Resolution with an exponential gap is clearly Levin in one direction: given a satisfying assignment of φ, Pudlák's construction can craft a refutation of $\textsc{Ref}(\varphi)$. However, it had remained open whether this Levin reduction could be made two-way: given a refutation π of $\textsc{Ref}(\varphi)$, can one always extract a satisfying assignment to φ in polynomial time?

For most if not all natural **NP**-complete languages, the corresponding search problems tend to be complete under Levin reductions. However, the same decision problem could admit different search problems associated to it, and it is known that if $\mathbf{P} \neq \mathbf{NP} \cap \mathbf{coNP}$, then there are **NP** search problems that do not reduce to each other under Levin reductions, while their decision versions are **NP**-complete (and hence do reduce to each other) under Karp reductions (see, for example, [KM00; FFNR03]). To the best of our knowledge, until now the only natural examples of candidates to be **NP**-hard search problems without Levin reductions were precisely certain problems arising in the context of meta-complexity. One is the Minimum Circuit Size Problem (MCSP), for which Mazor and Pass [MP24] recently proved that a certain gap version is *not* **NP**-complete under Levin reductions, assuming the existence of indistinguishability obfuscation (iO). The other candidate was precisely the reduction from SAT to the Proof Size Problem for Resolution. Theorem 1.1 settles this, giving a two-way Levin reduction.

The existence of the extraction algorithm answers question (Q_1) in the affirmative: Resolution refutations of $\textsc{Ref}_s(\varphi)$ *must leak a satisfying assignment.* This has a certain information-theoretic flavor: the fact that satisfying assignments can *always* be efficiently extracted implies that the most succinct description of a refutation of $\textsc{Ref}_s(\varphi)$ must include the description of a satisfying assignment for φ. We can make this precise in the language of Kolmogorov complexity using the framework of *information efficiency* of Krajíček [Kra22], who studied the minimum time-bounded Kolmogorov complexity (Kt) of propositional proofs.

Theorem 1.2 (Assignment extraction as information efficiency, informal)**.** *For every satisfiable CNF formula φ over n variables and* $\mathrm{poly}(n)$ *clauses,*

$$\mathsf{info}_{\mathsf{Res}}(\neg \textsc{Ref}(\varphi)) \approx \min\{\mathrm{Kt}(\alpha \mid \varphi) \mid \varphi(\alpha) = 1\},$$

where $\mathsf{info}_Q(\psi) := \min\{\mathrm{Kt}(\pi \mid \psi) \mid \pi : Q \vdash \psi\}$ *is Krajíček's information efficiency function.*

To the best of our knowledge, this is one of the first applications of Krajíček's framework.

1.1.2 The Proof Analysis Problem for strong proof systems

Motivated by (Q_2), we ask whether PAP is in **P** for strong proof systems. We conclude that the answer is likely negative by proving optimal conditional lower bounds in the form of **NP**-hardness for every proof system that p-simulates Extended Frege (EF).

Theorem 1.3 (**NP**-hardness of $\mathrm{PAP}_{\mathsf{EF}}$, informal)**.** *For every propositional proof system S that p-simulates Extended Frege, the Proof Analysis Problem for S is* **NP***-complete.*

This means that, unlike Resolution, strong proof systems are seemingly able to prove "trivial" Resolution lower bounds on satisfiable formulas without having to first prove that the underlying formula is satisfiable. In particular, this means Extended Frege is strong enough to *obfuscate* the satisfying assignments. As a consequence, for strong proof systems like Extended Frege, one cannot hope to prove they are **NP**-hard to automate following a strategy similar to that of [AM20].

1.1.3 Formalization of the Atserias–Müller lower bound in PV_1

The inspiration for why the extraction algorithm in Theorem 1.1 might exist in the first place comes from *witnessing theorems* in bounded arithmetic. We work here with Cook's theory PV_1 and Buss's S^1_2, which are first-order theories of arithmetic formalizing polynomial-time reasoning. In these theories, if a statement of the form $\forall x \exists y \varphi(x, y)$ with a low-complexity $\varphi(x, y)$ is provable in the theory, then there exists a polynomial-time algorithm that *witnesses* y given x. This implies, in particular, that if a problem is proven **NP**-hard in one of these theories, then the reduction will be a Levin reduction.

The key observation for us is that the statement of the lower bound is itself of this form, a $\forall\Sigma^b_1$ sentence:

> "*for every* formula φ and *every* Resolution refutation π of $\textsc{Ref}(\varphi)$,
> *there exists* a satisfying assignment for φ, or else π is large."

Thus, if the previous statement were provable in PV_1, we would get a polynomial-time function extracting satisfying assignments given φ and π.

While the extraction algorithm presented in Theorem 1.1 is given directly in natural language, it is still worth formalizing the lower bound in bounded arithmetic to obtain a variety of applications.

Theorem 1.4 (Atserias–Müller lower bound [AM20] in PV_1, informal)**.** *The theory* PV_1 *proves the statement that for every CNF formula φ over n variables and every size parameter $s \in \mathbb{N}$, if φ is unsatisfiable and π is a correct Resolution refutation of* $\textsc{Ref}_s(\varphi)$*, then* $|\pi| \geq 2^{\Omega(s/n^2)}$.

Formalizations in bounded arithmetic tend to be particularly interesting when they lead to new proofs of known statements. This has been the case, for example, with Razborov's formalizations of circuit lower bounds leading to the now-famous proof of Håstad's switching lemma via a simpler counting argument [Raz95]. Remarkably, the method introduced by Razborov to formalize the switching lemma is recognized for enabling proofs to at least two major conjectures in combinatorics [ALWZ21; PP24]. Another example

is a recent new proof of the Schwartz-Zippel lemma [AT24], proven via a hybrid argument formalizable in S^1_2. Our formalization of the lower bound on Ref formulas also relies on a new proof. We elaborate on this in the technical overview.

We remark that our bound of the form $2^{\Omega(s/n^2)}$ is slightly worse than the original one, which we state here for convenience.

Theorem 1.5 (Atserias–Müller lower bound [AM20])**.** *For every CNF formula φ over n variables and every size parameter $s \in \mathbb{N}$, if φ is unsatisfiable and π is a correct Resolution refutation of $\textsc{Ref}_s(\varphi)$, then $|\pi| \geq 2^{\Omega(s/n)}$.*

The difference in the bound means that we can only show that PV_1 proves hardness of $\textsc{Ref}_s(\varphi)$ for $s \geq n^3$. We leave it open whether PV_1 can achieve the original $2^{\Omega(s/n)}$ bound via a different argument. In any case, this is not particularly important for our applications. We comment on this further in the technical overview.

1.1.4 Formalization of Pudlák's upper bound in Resolution

We complement the formalization of the lower bound with a formalization of the upper bound [Pud03], showing that there are short refutations of $\textsc{Ref}(\varphi)$ whenever φ is satisfiable. This construction can be carried out by a constant-depth circuit and could be formalized in S^1_2, but certainly also in much weaker theories. We prove the somewhat surprising fact that the construction can be proven correct in Resolution itself.

Theorem 1.6 (Pudlák's upper bound [Pud03] in Resolution, informal)**.** *There is a polynomial-size depth-2 Boolean circuit $P(\alpha, \varphi, s)$ of fan-in 2 that given a CNF formula φ, a satisfying assignment α, and $s \in \mathbb{N}$, outputs a Resolution refutation π of $\textsc{Ref}_s(\varphi)$. Furthermore, the correctness of this circuit P has polynomial-size proofs in Resolution.*

That is, not only Resolution has short refutations of $\textsc{Ref}(\varphi)$ when φ is satisfiable: Resolution can show that the circuits generating these refutations from satisfying assignments are correct. This, again, is in striking contrast with the fact that Resolution does not have small proofs of its own soundness [AB04].

1.1.5 Propositional fragments of Atserias–Müller: automatability in terms of Ref formulas

The main consequence of the extraction algorithm together with its formalization in bounded arithmetic is the following precise characterization theorem relating the provability of a formula $\neg\varphi$ to the provability of the formula $\neg\textsc{Ref}(\textsc{Ref}(\varphi))$. (For the sake of clarity, we ignore for now the exact size parameters of the Ref formulas, which are always some fixed polynomials; in general, when we write $S \vdash_{\text{poly}} \varphi$ we mean that S has polynomial-size proofs of φ, and by "reasonable proof system" we mean essentially that the system is closed under *modus ponens*.)

Theorem 1.7 (Propositional fragments of Atserias–Müller, informal)**.** *Let S be a reasonable propositional proof system that simulates Extended Frege. Then, for every sequence $\{\varphi_n\}_{n\in\mathbb{N}}$ of unsatisfiable CNF formulas,*

$$S \vdash_{\text{poly}} \neg\varphi_n \qquad \textit{if and only if} \qquad S \vdash_{\text{poly}} \neg\textsc{Ref}(\textsc{Ref}(\varphi_n)).$$

The lower bound on Ref formulas says that for every unsatisfiable φ, the corresponding $\textsc{Ref}(\varphi)$ is hard for Resolution, making $\textsc{Ref}(\textsc{Ref}(\varphi))$ unsatisfiable. The latter encodes the statement "$\textsc{Ref}(\varphi)$ is hard for Resolution", and our theorem shows that when restricted to the reasoning power of a specific proof system S, such a lower bound has small proofs if, and only if, S has short proofs of the unsatisfiability of φ in the first place. That is, the *fragment* of the Atserias–Müller lower bound that has short proofs in S is precisely the one corresponding to the formulas that S can prove unsatisfiable with short proofs.

This characterization is surprisingly tight and has consequences for automatability and proof search. Since we can relate the proof size of φ in S to the proof size of $\textsc{Ref}(\textsc{Ref}(\varphi))$, this means that looking for proofs of $\textsc{Ref}(\textsc{Ref}(\varphi))$ can be a proxy for searching for proofs of φ.

Theorem 1.8 (Automatability in terms of Ref formulas, informal)**.** *For every reasonable proof system S that simulates Extended Frege as well as for Resolution itself,*

(i) S is automatable if, and only if, S is automatable exclusively on Ref *formulas;*

(ii) S is weakly automatable if, and only if, S is weakly automatable exclusively on Ref *formulas.*

We remark again that these Ref formulas are always talking about Resolution, not about S. That is, for every strong enough proof system, *efficient proof search over all tautologies is equivalent to efficient proof search over Resolution lower bounds.*

Until now no such general structural result was known that related proof search generally to proof search for a particular class of formulas. This goes in line with a question of Pich and Santhanam [PS22], who asked whether automating a proof system on truth-table tautologies (i.e., formulas stating circuit lower bounds) implies the automatability of the system on all tautologies. We have proved that this is the case for the class of formulas stating Resolution lower bounds in place of truth-table tautologies.

1.1.6 Unprovability of Resolution lower bounds

Theorem 1.6, together with Theorem 1.5, further imply that true Resolution lower bounds can be essentially arbitrarily hard to prove. Namely, if S is a propositional proof system where $\{\varphi_n\}_{n\in\mathbb{N}}$ is a sequence of formulas that S cannot refute in polynomial size, then S cannot refute $\{\textsc{Ref}(\textsc{Ref}(\varphi_n))\}_{n\in\mathbb{N}}$ either.

Theorem 1.9 (Propositional unprovability of Resolution lower bounds, informal)**.** *Let Q be a reasonable propositional proof system that simulates Resolution. If $\{\varphi_n\}_{n\in\mathbb{N}}$ is a sequence of hard unsatisfiable CNF formulas for Q, where φ_n has n variables and size $|\varphi_n| = \mathrm{poly}(n)$, then*

(i) the formulas $\textsc{Ref}_{n^2}(\varphi_n)$ over $N = \mathrm{poly}(n)$ variables are all unsatisfiable and require size $2^{N^{\Omega(1)}}$ to be refuted in Resolution;

(ii) yet, Q does not have polynomial-size refutations of the formulas $\textsc{Ref}_{N^2}(\textsc{Ref}_{n^2}(\varphi_n))$ stating quadratic lower bounds on $\textsc{Ref}_{n^2}(\varphi_n)$.

There is nothing special about quadratic lower bounds being unprovable—one can get arbitrarily small polynomial lower bounds by tweaking the encoding. See the discussion after Theorem 1.11.

We note that Iwama showed in 1997 that the Proof Size Problem for Resolution is **NP**-complete [Iwa97]. This means, in particular, that its complement in **coNP**-complete and hence, unless **NP** = **coNP**, no propositional proof system can efficiently derive *all* tautological Ref formulas (i.e., all true Resolution lower bounds), or else there would be a polynomially bounded proof system. While this has a similar flavor to our result, our theorem is different in at least two aspects. First, from an explicit family of hard tautologies we obtain an explicit family of hard Ref formulas for the system, in a generic way. Second, the parameters are essentially optimal: we identify a sequence of unsatisfiable formulas for which an exponential (and hence maximal) Resolution lower bound holds —while the system Q in question cannot even prove a quadratic lower bound.

As a corollary of Theorem 1.9, for example, we get the first explicit lower bounds for Ref formulas in bounded-depth Frege systems.

Corollary 1.10 (Hard Ref formulas for bounded-depth Frege, informal)**.** *For every $d \le O(\log n/\log\log n)$, the formulas $\textsc{Ref}(\textsc{Ref}(\mathrm{PHP}_n))$ are all unsatisfiable but require size $\exp\big(\Omega(n^{1/(2d+1)})\big)$ to be refuted in depth-d Frege systems.*

This corollary contrasts again with the open question of Pudlák [Pud20] about whether constant-depth Frege proves any of its own lower bounds.

Finally, using similar ideas, we obtain unconditional independence results for first-order theories of arithmetic.

Theorem 1.11 (First-order unprovability of Resolution lower bounds, informal)**.** *Let T be a consistent first-order theory extending Robinson Arithmetic by a set of polynomial-time recognizable axioms. Then, there exits a sequence of unsatisfiable propositional formulas $\{\psi_N\}_{N\in\mathbb{N}}$ described uniformly by a polynomial-time algorithm, where ψ_N has N variables, such that*

(i) Resolution refutations of the formula ψ_N require size $2^{N^{\Omega(1)}}$;

(ii) there exists $c > 0$ such that the theory T cannot prove $\Omega(N^c)$ lower bounds on the Resolution size of these refutations; that is, there is $N_0 \in \mathbb{N}$ such that the lower bound expressed by the first-order sentence $\forall N \forall \pi\, (N > N_0 \wedge \mathrm{Ref}_{\mathrm{Res}}(\psi_N, \pi) \rightarrow |\pi| > N^c)$ is unprovable in T.

We remark that the theory T in this theorem can be arbitrary strong. This implies that, unconditionally, theories like Peano Arithmetic (PA) cannot prove all true Resolution lower bounds. The same ideas apply to Zermelo-Fraenkel Set Theory (ZFC) and similarly powerful formal systems.

We also note that the constant $c > 0$ in the exponent of the unprovable lower bound depends on the definition of ψ_N. In general, one can alter ψ_N to get an unprovable lower bound of the form $\Omega(N^c)$ for any fixed constant $c > 0$.

1.2 Technical overview

Next we provide a technical overview of the main proof ideas and how these are combined to yield our main results and corollaries.

1.2.1 Assignment extraction: techniques

We obtain the extraction algorithm in Theorem 1.1 by derandomizing the proof of the Resolution lower bound for the Ref formulas. The original proof revolves around the concept of *block-width* (called *index-width* in [AM20]) in Resolution refutations of $\mathrm{Ref}_s(\varphi)$. The variables of the formula are arranged into s *blocks*, each encoding a clause in the purported refutation of size s. The block-width of a refutation π is then the largest number of blocks mentioned in a clause of the refutation π. The proof proceeded in two steps:

1. derive a *block-width lower bound*, showing that if $\varphi \notin$ SAT, then the block-width of any refutation of $\mathrm{Ref}_s(\varphi)$ must be large;

2. by a *random restriction argument*, argue that if the refutation π is small, there exists a restriction that makes the block-width of the restricted refutation small, contradicting the previous point.

Our algorithm works by following these steps in reverse. First, given a refutation π from which we want to extract a satisfying assignment, instead of sampling a restriction at random from a specific distribution, we construct a restriction *deterministically* in a greedy fashion, tailored to the specifics of π. This is reminiscent of the kind of greedy deterministic restrictions used by Cook and Pitassi [CP90] to formalize Haken's lower bound for the pigeonhole principle in bounded arithmetic, and more broadly in the style of Beame and Pitassi [BP96], Clegg, Edmonds, and Impagliazzo [CEI96] and Ben-Sasson and Wigderson [BW01]. Our algorithm runs in deterministic polynomial time and always succeeds in finding a restriction that reduces the block-width to $O(\sqrt{s \log |\pi|})$.

In the second step, we look at the proof of the block-width lower bound and interpret it as a Prover-Delayer game in the style of Atserias and Dalmau [AD08]. The Prover issues queries about the values of

the variables of $\textsc{Ref}_s(\varphi)$, or forgets previously recorded such values, and the Delayer replies following a concrete strategy that allows them to keep playing until a large number of blocks appear queried. Our algorithm traverses the Resolution refutation guided by the Delayer's strategy in the Prover-Delayer game. We can then prove that this Delayer's strategy will, after a polynomial number of steps, reach either

(a) a clause of *high block-width*, or

(b) a clause encoding a *satisfying assignment* to φ.

Since the greedy deterministic restriction in the first step made sure the block-width is small, the Delayer will be guaranteed to find a satisfying assignment.

We note that our deterministic restriction only achieves a reduction of block-width to $O(\sqrt{s \log |\pi|})$, while using a random restriction one could achieve up to $O(\log |\pi|)$. In fact, if one allows randomness in the extraction algorithm, then an argument similar to the random restriction of [AM20] yields a zero-error probabilistic polynomial-time extraction algorithm that works even when the refutation π being analyzed is for the formula $\textsc{Ref}_{n^2}(\varphi)$. In contrast, the price to pay for determinism is that the size parameter s should be at least n^3.

1.2.2 NP-hardness of PAP_{EF}

The idea behind the hardness proof in Theorem 1.3 is best explained as a reduction from the Minimum Circuit Size Problem (MCSP) —although given that MCSP is not known to be **NP**-hard, the actual proof in the main text is a bit more technical and goes instead via a reduction from Vertex Cover. In 2004, Razborov [Raz04] proved that Resolution cannot efficiently prove circuit size lower bounds. This statement is captured by the the well-known *truth-table tautologies* $\textsc{tt}(f, s)$ stating that a truth-table f can be computed by a circuit of size s. Then, the formula $\textsc{Ref}_t(\textsc{tt}(f, s))$ states that there exists a size-t Resolution refutation of $\textsc{tt}(f, s)$. By Razborov's lower bound there are no such Resolution refutations, meaning that $\textsc{Ref}_t(\textsc{tt}(f, s))$ is unsatisfiable for values of t polynomial in $|\textsc{tt}(f, s)|$.

An interesting feature of Razborov's lower bound is that it is agnostic about whether f is actually hard for circuits of size s. Namely, even if f was computable by size-s circuits, making $\textsc{tt}(f, s)$ satisfiable and hence $\textsc{Ref}_t(\textsc{tt}(f, s))$ unsatisfiable for a trivial reason, Razborov's argument can still prove the unsatisfiabilty of this Ref formula without exhibiting a satisfying assignment for $\textsc{tt}(f, s)$.

Now, suppose that Razborov's lower bound was formalizable in, say, Extended Frege[1] in a uniform manner. That is, suppose there is a polynomial-time algorithm that given a truth-table f and size parameters s and t outputs an EF proof π such that

$$\pi : \mathsf{EF} \vdash \neg \textsc{Ref}_t(\textsc{tt}(f, s)).$$

Then, if PAP_{EF} happened to be in **P**, there would be a polynomial-time algorithm that given π would decide whether $\textsc{tt}(f, s) \in \text{SAT}$, which is the same as deciding MCSP. Hence, PAP_{EF} (or PAP for whatever system capable of formalizing Razborov's proof) would be at least as hard as MCSP.

For our proof we do not formalize Razborov's lower bounds, and instead instantiate this idea for a specific propositional encoding of Vertex Cover for which Resolution lower bounds follow from Haken's lower bound for the pigeonhole principle. Here we leverage the formalization of Cook and Pitassi, who showed that Haken's lower bound is provable in EF [CP90].

[1] We note that it is not known nor clear at all that Razborov's argument goes through in EF. The assumption is only for the sake of exposition.

1.2.3 Formalization of the upper and lower bounds

A large part of our technical contribution consists in formalizing the proofs leading to the **NP**-hardness of automating Resolution. We summarize below some of the challenges encountered, and the solutions devised.

The Atserias–Müller lower bound in PV_1. Different proofs of the lower bound exist in the literature, but none of them seem directly formalizable in PV_1. The original proof has the caveat of the random restriction argument, which might be formalizable in Jeřábek's theory APC_1, but likely not in PV_1. In addition, the block-width lower bound is proven by relating small refutations to the canonical exponential-size tree-like refutation of any formula, which could be hard to reason about in bounded theories.

In the work of de Rezende, Göös, Nordström, Pitassi, Robere, and Sokolov [dRGN+21] two alternative proofs were presented. The first proof consists of a random restriction followed by a block-width lower bound proven via a reduction to the retraction weak pigeonhole principle. The random restriction presents the same formalization issues as the original proof, and the block-width reduction is equally problematic: the decision tree reduction they use has low depth, which is necessary to transfer the lower bounds on (block-)width, but the size of the decision tree itself seems to be superpolynomial, and hence cannot be reasoned about in PV_1. The second proof of de Rezende *et al.* uses this same block-width lower bound followed by the size-width trade-offs of Ben-Sasson and Wigderson [BW01]. The block-width lower bound is still problematic, of course, but in addition to this, the statement of Ben-Sasson and Wigderson—"for every small Resolution refutation, there exists another Resolution refutation in small width"—is itself impossible to formalize in bounded arithmetic. The reason for this is that, as demonstrated by Thapen, in general these narrow proofs can require superpolynomial size [Tha16], and therefore the statement of Ben-Sasson and Wigderson cannot possibly be a bounded formula.

Finally, Garlík [Gar19] has proven lower bounds on the Ref formulas for the so-called non-relativized encoding. Unfortunately for us, his proofs encounter the same barrier: they rely on random restriction arguments, which are in any case more involved than the original ones.

We resolve these issues by coming up with new proof, inspired by the extraction algorithm, that modifies both ingredients in the original proof, yielding Theorem 1.4. The random restriction argument is replaced by a greedy deterministic restriction just like the one used in the extraction algorithm. For the block-width lower bound, we show that the argument can be completely described by a Prover-Delayer game without referring to the exponential-size canonical tree-like refutation, making the entire proof formalizable in PV_1.

We remark that moving from the random restriction to the deterministic one comes at the cost of a slightly worse lower bound. The original size bound on Ref formulas is of the form $2^{\Omega(s/n)}$ and hence yields $2^{\Omega(n)}$ Resolution size lower bounds for all Ref_s formulas with $s \geq n^2$. Our deterministic restriction, in line with the parameters of the extraction algorithm, achieves a lower bound of $2^{\Omega(s/n^2)}$ which is exponential in $\Omega(n)$ for $s \geq n^3$. It seems reasonable that the original proof with the random restriction can be formalized in APC_1, but we have not carried out this formalization.

Pudlák's upper bound in Resolution. For the upper bound in Resolution (Theorem 1.6), our proof is based on a careful analysis of the construction that makes it possible to describe the construction by a low-depth circuit. Carrying out the proof of the correctness of this circuit in Resolution is tedious, but ultimately clear once the right description of the circuit is provided.

An interesting technical ingredient is the fact that the correctness statement itself ("if α is a satisfying assignment, then the circuit outputs a correct refutation") is an implication that cannot be immediately expressed as a CNF formula that Resolution can handle. To deal with this, we devise a construction using extension variables that simulates negations of CNF formulas in Resolution, which we name *pseudo-negations*. We show that, with the aid of this pseudo-negation operators, Resolution can in fact carry out

modus ponens inferences.

1.2.4 Consequences

Characterization of the propositional fragments of Atserias–Müller. Our characterization theorem (Theorem 1.7) is a consequence of the formalization of the lower bound in PV_1 (Theorem 1.4) and the upper bound in Resolution itself (Theorem 1.6). Those results, in the propositional setting, imply that

1. for the extraction algorithm E, we have $\mathsf{EF} \vdash \textsc{Ref}(\textsc{Ref}(\varphi), \pi) \to \textsc{Sat}(\varphi, E(\varphi, \pi))$; and
2. for Pudlák's algorithm P, we have $\mathsf{Res} \vdash \textsc{Sat}(\varphi, \alpha) \to \textsc{Ref}(\textsc{Ref}(\varphi), P(\varphi, \alpha))$.

Here, we highlight the variables of the Ref formula encoding a refutation π as the second argument of Ref. In particular, the Resolution proof of the correctness of P is also possible in EF. Then, simple use of contraposition allows us to go from $\neg\varphi$ to $\neg\textsc{Ref}(\textsc{Ref}(\varphi))$, and vice versa. That is, if EF can prove $\neg\varphi$, then it can also prove it in the encoding $\neg\textsc{Sat}(\varphi, \alpha)$, and when substituting $E(\varphi, \pi)$ for α, where π are the only free variables, contraposition on item (1) gives us that EF derives $\neg\textsc{Ref}(\textsc{Ref}(\varphi), \pi)$. The other direction is analogous.

Automatability in terms of Ref formulas. For the characterization of automatability in Theorem 1.8 to go through we build on Theorem 1.7 and additionally show that the characterization given there is not only in terms of proof size, but it is actually constructive. Given a proof of $\neg\varphi$ in S we can efficiently construct a proof of $\neg\textsc{Ref}(\textsc{Ref}(\varphi))$, and vice versa. In this way, searching for proofs of $\textsc{Ref}(\textsc{Ref}(\varphi))$ is a proxy for the proofs of φ.

Remarkably, our proof techniques fail for proof systems strictly between Extended Frege and Resolution. The upper bound in Resolution does imply that from a refutation of $\textsc{Ref}(\textsc{Ref}(\varphi))$ we can obtain a refutation of φ. Unfortunately, it is our extraction algorithm (Theorem 1.1) what guaranteed that if φ has a refutation of size t, then $\textsc{Ref}(\textsc{Ref}(\varphi))$ has a refutation of size $\mathrm{poly}(t)$. In Extended Frege this is true thanks to the extraction algorithm, but it seems conceivable that weaker systems might be able to easily prove $\neg\varphi$ without being able to prove $\neg\textsc{Ref}(\textsc{Ref}(\varphi))$ efficiently. (For Resolution itself this result does go through, for the more ad-hoc reason that Ref formulas talk about Resolution itself).

Unprovability of Resolution lower bounds. For Theorem 1.9 we exploit Theorem 1.6: if there is a short refutation of $\textsc{Ref}(\textsc{Ref}(\varphi))$, then there is a short refutation of φ. Since we formalized the upper bound construction in Resolution, the result applies to any proof system that contains Resolution (and behaves naturally in the sense that is closed under *modus ponens*). Then, if φ is a hard formula for Q and Q simulates Resolution, we have that $\textsc{Ref}(\varphi)$ is unsatisfiable. By the lower bound on Ref formulas (Theorem 1.5), this formula is exponentially hard for Resolution, making $\textsc{Ref}(\textsc{Ref}(\varphi))$ unsatisfiable as well —but hard to refute for Q.

In the first-order setting, Theorem 1.11 relies again on the formalization of the upper bound on Ref formulas. This time, instead of starting from a sequence of hard propositional formulas, we can leverage Gödel's second incompleteness theorem to start from a sentence (the consistency of T) that is unconditionally unprovable in T. From this follows that T cannot prove the soundness of a certain propositional system based on T (the so-called *strong proof system of* T [Pud20]). We then consider the $\textsc{Ref}(\cdot)$ formula around these soundness statements. We conclude that if T could derive the lower bound on the $\textsc{Ref}(\cdot)$ formulas in question, it would also be able to prove the soundness of the strong proof system of T and, as a consequence, T would derive its own consistency. Since Gödel's incompleteness theorem gives us sentences that are unconditionally independent of T, the corresponding Resolution lower bounds are also unconditionally unprovable in T. This works essentially for any theory of arithmetic subject to Gödel's incompleteness phenomenon, and does not rely on any complexity-theoretic assumptions.

1.3 Related work

Our work fits into a trend in complexity theory concerned with the meta-mathematics of computational complexity, which has gained remarkable momentum in recent years. Most of this work has been primarily concerned with the formalization of cornerstone results of computational complexity in bounded arithmetic and establishing unprovability and logical independence as barrier results. The literature is too vast to review here, so refer the reader to the recent survey of Oliveira [Oli25]. In parallel, there has been a growing body of work deploying tools and ideas from mathematical logic to prove complexity-theoretic statements (see, e.g., [Gay21; Mül21; Kha22a; Kha22b; PS22; Nar22; PS23; Kra23; Gay23; Kha24; ACG24; AKPS24; Kra24; Nar24; Gay24]). Our work continues in this direction.

Two recent works conceptually related to our investigations on Ref formulas merit further discussion. Santhanam and Tzameret [ST21] initiated a general study of Ref formulas for arbitrarily strong proof systems. In particular, they studied *iterations* of these formulas, which are reminiscent of the nested $\text{Ref}(\text{Ref}(\varphi))$ formulas that feature in our work. Their Ref formulas are not limited to Resolution, and they consider the iterated version of Ref^Q when Ref^Q talks about an arbitrarily strong proof system Q. While we are unable to connect our work on analyzability to their results, our characterization of proof size in terms of Ref formulas (Theorem 1.7) has conceptual ties to their Iterated Lower Bounds Hypothesis.

The other relevant work is the research of Li, Li, and Ren [LLR24], who studied the provability of Resolution lower bounds in relativized theories of bounded arithmetic in the context of **TFNP**. Until their work, the only formalization of proof complexity lower bounds that we are aware of is that of Cook and Pitassi [CP90]. Li, Li, and Ren studied so-called *refuter problems* in proof complexity: given a purported Resolution refutation of, say, PHP_n, that is smaller than the known lower bounds, find a mistake in the proof (which must certainly exist, due to these very lower bounds). They connect the provability of lower bounds to the complexity of solving these refuter problems in subclasses of **TFNP**. While their results yield formalizations of some proof complexity lower bounds, our results are essentially incomparable. First, their provability results are for *relativized* theories of bounded arithmetic, where the given Resolution refutation is accessed through an oracle, while our proofs are in the non-relativized theories, where we can quantify over the objects in question. Second, they consider the provability of lower bounds for explicit families of tautologies like the pigeonhole principle or the Tseitin formulas. In contrast, the lower bound we are concerned is a sort of *meta lower bound*: it tells us that the $\text{Ref}(\varphi)$ formulas are hard whenever φ is unsatisfiable. We believe, however, that the **TFNP** perspective on analyzability might shed light on some of our open questions.

1.4 Open problems

Analyzability of constant-depth Frege and other weak proof systems. Similar techniques to those of [AM20] have been employed to prove the **NP**-hardness of automating other weak proof systems like Regular and Ordered Resolution [Bel20; BY24], k-DNF Resolution [Gar24], Cutting Planes [GKMP20], Nullstellensatz and Polynomial Calculus [dRGN+21], the OBDD proof system [IR22] and, more recently, even $\mathbf{AC}^0$-Frege [Pap24]. All proof systems weaker than Resolution are analyzable just because Resolution is (i.e, their corresponding PAP problems are in **P**), since analyzability is downwards closed under simulations. For the stronger systems, the question remains open. Are these systems analyzable? What about their search versions?

We highlight the analyzability of constant-depth Frege as a particularly interesting problem. While we have proven some unconditional lower bounds on Ref formulas here, it is open whether $\mathbf{AC}^0$-Frege can prove any true Resolution lower bounds at all. It has been conjectured in the past that the PHP lower bound might be formalizable in these systems, at least in quasi-polynomial size. If this was possible, the **NP**-hardness of PAP_{EF} in Theorem 1.3 could be improved all the way to these systems.

FP-completeness of the search version of $\mathrm{PAP}_{\mathrm{Res}}$. While we have shown that assignment extraction can be performed in polynomial time, our algorithm does not seem to be possible anywhere below **P**. The algorithm seems hard to parallelize, which raises the question of whether the search version of $\mathrm{PAP}_{\mathrm{Res}}$ is in **NC** or even below. This is related to the question of whether the formalization of the lower bound on Ref formulas is provable in theories weaker than PV_1. If the statement was provable in, say, VNC^1, witnessing theorems would give us a extraction in $\mathbf{FNC}^1$. We conjecture that this improvement is in fact impossible, and that the search problem of $\mathrm{PAP}_{\mathrm{Res}}$ is complete for **FP** under $\mathbf{AC}^0$-reductions, but we are unable to prove it. The reduction, if true, likely requires some new technical idea. This would imply, amongst other things, that V^0 does not prove the lower bound on Ref formulas, unconditionally.

On the weak automatability of Resolution. Recall that a proof system is *weakly automatable* if there exists a proof system that p-simulates it and is automatable. By our Theorem 1.8, the weak automatability of Resolution is equivalent to a proof system Q simulating Resolution and being automatable on Ref formulas. If $Q \geq \mathsf{EF}$, then our theorem would imply that Q itself would be automatable on all formulas, hitting cryptographic hardness results [KP98; BPR00; BDG+04; ACG24]. However, if Q is strictly weaker than EF, our statement does not apply and the automatability of Q on Ref formulas does not imply automatability on all formulas. This does not seem to contradict any hardness assumptions. Of course, no such Q is known to be efficiently automatable on Resolution lower bound statements, but this raises again the question of whether some non-trivial algorithm weakly automating Resolution might be plausible.

1.5 Structure of the paper

The paper is structured as follows. After the preliminaries in Section 2, we dedicate Section 3 to formally defining the Proof Analysis Problem and stating some basic facts about it. Section 4 proves Theorem 1.1, describing the algorithm for the search version of $\mathrm{PAP}_{\mathrm{Res}}$ and Theorem 1.2. Section 5 proves Theorem 1.3, giving **NP**-hardness of $\mathrm{PAP}_{\mathsf{EF}}$ and stronger systems. In Section 6 and Section 7 we formalize, respectively, the Resolution lower bound and upper bound on Ref formulas that yield Theorem 1.4 and Theorem 1.6. Section 8 translates this to the propositional setting to show that Extended Frege has polynomial-size proofs of the **NP**-hardness of automating Resolution. Finally, Section 9 gives the proof of the characterization in Theorem 1.7 and proves Theorem 1.8 on the equivalence of automatabilty of Ref formulas, while Section 9.3 proves Theorem 1.9 and Theorem 1.11 on the unprovability of Resolution lower bounds.

2 Preliminaries

We assume the reader to be familiar with the central concepts of computational complexity theory. Below, we review the essential definitions and facts involving proof complexity and bounded arithmetic that feature in the paper. For a more comprehensive treatment of proof complexity, we refer to Krajíček [Kra19]. For bounded arithmetic, the recent survey of Oliveira [Oli25] covers all the necessary material in the style of the meta-mathematics of computational complexity, which aligns with the style of our work. Other classical texts in logic and bounded arithmetic also cover these contents (see, e.g., [HP93; Kra95; Bus97; Bus98]).

2.1 Levin reductions

For decision problems $A, B \subseteq \{0,1\}^*$, we use the notation $A \leq_{\mathrm{m}}^{\mathrm{p}} B$ to express that A *many-one reduces* (or *Karp reduces*) to B, meaning that there is a polynomial-time computable function f such that for all $x \in \{0,1\}^*$ we have $x \in A$ if and only if $f(x) \in B$. We will be concerned with a strengthening of Karp reductions for *search problems*. A search problem for us is a relation $R \subseteq \{0,1\}^* \times \{0,1\}^*$. A search problem

R is in **FP** if there exists a polynomial-time function that on input x, outputs some y such that $(x, y) \in R$, if such a y exists.

We will say that R_1 *Levin reduces* to R_2 and write $R_1 \leq^{\mathrm{p}}_{\mathrm{Levin}} R_2$ if there exists a triple of polynomial-time computable functions (f, g, h) such that for all $x, w \in \{0,1\}^*$, it holds that (i) if $(x, w) \in R_1$, then $(f(x), g(x, w)) \in R_2$, and (ii) if $(f(x), w) \in R_2$, then $(x, h(x, w)) \in R_1$. Whenever membership in R_1 and R_2 is checkable in polynomial time, it holds that the sets $\operatorname{dom} R_1$ and $\operatorname{dom} R_2$ are in **NP** and $\operatorname{dom} R_1 \leq^{\mathrm{p}}_{\mathrm{m}} \operatorname{dom} R_2$.

2.2 Proof complexity

Following the classical definition of Cook and Reckhow [CR79], a *propositional proof system* S for the set Taut of propositional tautologies is a polynomial-time function $S : \{0,1\}^* \to \textsc{Taut}$ whose range is exactly Taut. We think of S at the polynomial-time verifier mapping proofs to the statements they prove; i.e., if $S(\pi) = \varphi$, then we say π is an S-proof of φ. It is often convenient to think of a proof system as establishing *unsatisfiability*; thus, if φ is an unsatisfiable formula and π is an S-proof of the tautology $\neg\varphi$, then we say that π is an *S-refutation* of φ, or an S-proof of the unsatisfiability of φ. Since we deal exclusively with classical logic, here and below we tacitly gloss over the distinction between the formulas $\neg\neg\varphi$ and φ; this is particularly useful for literals ℓ, where $\neg\ell$ is sometimes used to denote the complementary literal.

For a tautology φ and a proof system S, we denote by $\operatorname{size}_S(\varphi) \coloneqq \min_{\pi : S(\pi)=\varphi} |\pi|$ the *size* of its smallest S-proof. A proof system S is *polynomially bounded* if there exists a constant $c \in \mathbb{N}$ such that for all $\varphi \in \textsc{Taut}$ we have $\operatorname{size}_S(\varphi) \leq |\varphi|^c$. For a sequence $\varphi = \{\varphi_n\}_{n\in\mathbb{N}}$ of tautologies, we write $S \vdash_{\mathrm{poly}} \varphi$ or simply $S \vdash_{\mathrm{poly}} \varphi_n$ to express that $\operatorname{size}_S(\varphi_n) = |\varphi|^{O(1)}$ as n grows. When we want to emphasize that it is via a specific proof π that S proves φ_n, we write $\pi : S \vdash \varphi_n$. More generally, for $s \in \mathbb{N}$, we write $S \vdash_s \varphi_n$ to express that there exists an S-proof π of size $|\pi| \leq s$ such that $\pi : S \vdash \varphi_n$.

We say that a proof system S *simulates* another system Q, written $S \geq Q$, if there exists a constant $c \in \mathbb{N}$ such that for every $\varphi \in \textsc{Taut}$ we have $\operatorname{size}_S(\varphi) \leq \operatorname{size}_Q(\varphi)^c$. We additionally say that S *p-simulates* Q and write $S \geq_{\mathrm{p}} Q$ if there exists a polynomial-time computable function sending Q-proofs to S-proofs of the same formula; i.e., there exists a polynomial-time computable function f such that for every $\varphi \in \textsc{Taut}$ and every $\pi : Q \vdash \varphi$, we have $f(\pi) : S \vdash \varphi$. We say that two proof systems S and Q are *polynomially equivalent* if $S \geq_{\mathrm{p}} Q$ and $Q \geq_{\mathrm{p}} S$. A proof system S is *optimal* if $S \geq Q$ for every propositional proof system Q, and respectively *p-optimal* if $S \geq_{\mathrm{p}} Q$ for every propositional proof system Q.

A *literal* is a propositional variable or its negation. Given a formula $\varphi(x_1, \ldots, x_n)$, a *literal substitution* is a mapping of the form $\rho : \{x_1, \ldots, x_n\} \to \{x_1, \ldots, x_n, \neg x_1, \ldots, \neg x_n, 0, 1\}$ that replaces variables by other literals or substitutes constants in their place. We denote by $\varphi_{\restriction\rho}$ the substituted formula $\varphi(\rho(x_1), \ldots, \rho(x_n))$, with the convention that every resulting occurrence of $\neg\neg x_i$ is replaced by x_i. A *restriction* is a particular case of a variable substitution, where all variables are mapped to either 0, 1, or themselves. We say that a proof system S is *closed under substitutions* (respectively, *closed under restrictions*) if there exists a constant $d \in \mathbb{N}$ such that for every tautology φ and every literal substitution (respectively, restriction) ρ, it holds that $\operatorname{size}_S(\varphi_{\restriction\rho}) \leq \operatorname{size}_S(\varphi)^d$. All the explicit proof systems dealt with in this work (i.e., the ones described below, like Resolution or Frege or Extended Frege systems) are closed under literal substitutions. In these cases, a proof of the substituted formula can be obtained directly by applying the substitution line by line to every formula appearing in a proof π of φ, and we hence denote by $\pi_{\restriction\rho}$ the corresponding substituted proof of $\varphi_{\restriction\rho}$.

2.2.1 Resolution

A central proof system in this work is Resolution (Res). We usually see this as a refutation system for CNF formulas. Accordingly, we sometimes write $\pi : \mathrm{Res} \vdash \neg\varphi$ for a CNF formula φ, to mean that π is a Resolution refutation of φ, hence a proof of the tautology $\neg\varphi$. In this way, Res is a Cook-Reckhow

proof system for the fragment of TAUT made of the formula of the form $\neg\varphi$, where φ is an unsatisfiable CNF formula. Through the standard Tseitin transformation of an arbitrary propositional formula into equisatisfiable CNF form, Res can also be seen as a Cook-Reckhow proof system for TAUT itself; we do not need the details of this in this paper.

A *literal* is a propositional atom or its negation, a *clause* is a disjunction of literals, and a *CNF formula* is a conjunction of clauses. We see clauses as sets of literals, and write simply $C \subseteq D$ to express that C is a subclause of D. A *Resolution refutation* of an unsatisfiable CNF formula $\varphi = C_1 \wedge \cdots \wedge C_m$ over variables $x_1, \ldots, x_n$ is a sequence $D_1, \ldots, D_s$ of clauses over $x_1, \ldots, x_n$ such that $D_s = \bot$, denoting the empty clause, and for every $i \in [s-1]$, the clause D_i either (a) is one of the clauses $C_1, \ldots, C_m$ of φ, or (b) is a *weakening* of a previous clause, meaning that $D_i \supseteq D_j$ for some $1 \leq j < i$, or (c) has been obtained from two previous clauses $D_j = A \vee x$ and $D_k = B \vee \neg x$, for $j, k < i$, by an application of the *Resolution rule*:

$$\frac{A \vee x \qquad B \vee \neg x}{A \vee B} \text{ (Res)}$$

We say that $A \vee B$ was obtained by *resolving* over x. The *length* of π, denoted by $\text{length}(\pi)$, is s.

To every Resolution refutation π we can associate a directed acyclic graph in a natural way, and we often do so implicitly. We denote by $\text{depth}(\pi)$ the length of the longest path in the dag, starting from the root labeled by the empty clause $\bot$. The number of vertices in this graph is precisely $\text{length}(\pi)$.

We will also deal with a mild extension of the Resolution system, known as *k-DNF Resolution* [Kra01], denoted $\text{Res}(k)$ for $k \geq 1$. The system $\text{Res}(k)$ is also a refutational system, but lines are *k-DNF* formulas, which are unbounded fan-in disjuctions of *k-terms*, conjunctions of up to k literals. A clause is a 1-disjunction. The system consists of a weakeaning and an introduction rule,

$$\frac{A}{A \vee B} \text{ (Weak)} \qquad \frac{A \vee \ell_1 \qquad B \vee (\ell_2 \wedge \cdots \wedge \ell_s)}{A \vee B \vee (\ell_1 \wedge \cdots \wedge \ell_s)} \text{ ($\wedge$-Intro)}$$

together with a Cut rule that generalizes the Resolution rule,

$$\frac{A \vee (\ell_1 \wedge \cdots \wedge \ell_s) \qquad B \vee \neg\ell_1 \vee \cdots \vee \neg\ell_s}{A \vee B} \text{ (Cut)}$$

where A and B are k-DNF formulas and $s \leq k$.

It is easy to see that Resolution (Res) corresponds to $\text{Res}(1)$.

2.2.2 Frege systems

Through this work we reason about Resolution refutations within much stronger systems for propositional logic. A *Frege system* [CR79] consists of a finite set of axiom schemas and inference rules that are sound and implicationally complete for the language of propositional tautologies built from the Boolean connectives negation ($\neg$), conjunction ($\wedge$), and disjunction ($\vee$). A *Frege proof* is then a sequence of formulas where each formula is obtained by either substitution of an axiom schema or by application of an inference rule on previously derived formulas. The specific choice of rules does not affect proof size up to polynomial factors, as long as there are only finitely many rules and these are sound and implicationally complete [CR79]. We refer to Cook and Reckhow [CR79] or Krajíček [Kra19, §2.1] for specific examples of choices for these rules and axioms. One can alternatively define Frege systems in the formalism of Natural Deduction or the Sequent Calculus for classical propositional logic, but we will not be concerned with these syntactic details.

Of central importance for us is the *Extended Frege* (EF) system [CR79], in which proof lines can be succinctly written as Boolean circuits rather than formulas [Jeř05]. In general, for a circuit class $\mathbf{C}$, one can consider the proof system $\mathbf{C}$-Frege, in which lines are restricted to be Boolean circuits of that type. We are particularly interested in the $\mathbf{AC}^0_d$-Frege systems, in which lines are restricted to be Boolean circuits of

unbounded fan-in and constant depth d. We also consider more generally *bounded-depth Frege systems*, where the depth d is bounded, but not necessarily a constant.

For bounded-depth Frege systems, we have strong lower bounds available. The most famous such lower bound is the one for the *Pigeonhole Principle* (PHP). For every $m \in \mathbb{N}$ and $n \in \mathbb{N}$ such that $m > n$, the formula PHP_n^m stands for the CNF formula over variables $p_{i,j}$ for $i \in [m]$ and $j \in [n]$ consisting of the clauses

$$\bigvee_{j \in [n]} p_{i,j} \qquad \text{for all } i \in [m], \tag{PHP-1}$$

$$\neg p_{i,j} \vee \neg p_{i,j'} \qquad \text{for all } i \in [m] \text{ and } j, j' \in [n], j \neq j' \tag{PHP-2}$$

$$\neg p_{i,j} \vee \neg p_{i',j} \qquad \text{for all } i, i' \in [m], i \neq i' \text{ and } j \in [n] \tag{PHP-3}$$

We sometimes denote by PHP_n the formula PHP_n^{n+1}.

Strong lower bounds are known on the proof complexity of the pigeonhole principle for bounded-depth Frege systems [Ajt94; PBI93; KPW95]. Here we state only a simplified version of the best such lower bound, proven by Håstad [Hås23].

Theorem 2.1 ([Hås23])**.** *For every $d \leq O(\log n / \log\log n)$, depth-$d$ Frege systems requires size at least $\exp\left(\Omega(n^{1/(2d-1)})\right)$ to prove $\neg\mathrm{PHP}_n$.*

Finally, we often consider extensions of Extended Frege by sets of additional axioms. For a set $A \subseteq \textsc{Taut}$ of tautologies that is recognizable in polynomial time, the system $\mathsf{EF} + A$ refers to Extended Frege extended with the axiom schemas that allow (formula) substitution instances of any formula in A.

2.2.3 Automatability and proof search

The notion of of automatability, introduced by Bonet, Pitassi, and Raz [BPR00], is a natural formalization of efficient proof search in propositional proof system. We say that a proof system S is *automatable* if there exists a constant $c \in \mathbb{N}$ and an algorithm that given a propositional tautology φ, outputs an S-proof of φ in time $(|\varphi| + \mathrm{size}_S(\varphi))^c$, meaning that the proof search algorithm succeeds in finding a proof of size polynomial in the size of the shortest one.

Even when a system might not be automatable, it seems natural to ask whether there exists a system Q that p-simulates S and is itself automatable. In this case, we say that S is *weakly automatable* [AB04]. Weak automatability is equivalent to the existence of an automating algorithm where the output proof belongs to a system $Q \geq_{\mathrm{p}} S$ rather than S itself. In particular, weak automatability is closed downwards under p-simulation.

A more restrictive notion of proof search is given by the *Proof Size Problem*. Associated to any propositional proof system S we can define the *Proof Size Problem for S* (PSP_S), defined as the language

$$\mathrm{PSP}_S := \{(\varphi, 1^s) \mid \text{there is an } S\text{-proof of } \varphi \text{ in size } s\}.$$

Automating S entails approximating minimum proof-size to a polynomial, in polynomial time.

2.3 Bounded arithmetic

We heavily rely on the connection between propositional proof complexity and (weak) theories of arithmetic. We assume familiarity with basic knowledge of first-order logic and introduce the main theories we are concerned with.

2.3.1 The theories PV_1 and S^1_2

Theories of bounded arithmetic capture various forms of feasible reasoning and act as a uniform counterpart of propositional proof systems. The main tool to capture feasibility in mathematical reasoning is to bound the complexity of formulas over which one can apply induction.

Cook's PV_1. Cook's theory PV_1 [Coo75; KPT91] is an attempt to make precise the idea of polynomial-time reasoning. It is a universal theory whose vocabulary $\mathcal{L}_{\mathsf{PV}}$ consists of a function symbol for each polynomial-time function, and the axioms are precisely the recursive definitions of these functions via composition and limited recursion on notation, in the style of Cobham's functional definition of **FP** [Cob65]. The theory further admits induction on quantifier-free formulas, which define precisely polynomial-time predicates.

The formal definition of PV_1 is rather technical and the details are not particularly relevant to our proofs, so we refer the reader to Krajíček's textbook [Kra95, Definition 5.3] for the details. The reason we rarely care about the technicalities of PV_1 is that we often work instead in the theory S^1_2 of Buss, which happens to be conservative over PV_1 for the classes of formulas we are interested in. We discuss this next.

Buss's S^1_2. We see S^1_2 as a theory sitting in between Robinson's Arithmetic Q and Peano Arithmetic PA. Let $\mathcal{L}_{\mathsf{PA}}$ denote the language of Peano Arithmetic, $\mathcal{L}_{\mathsf{PA}} \coloneqq \{0, 1, +, \cdot, <\}$. The axioms of PA consist first of the axioms of Robinson's arithmetic Q, which define the basic behavior of the symbols of $\mathcal{L}_{\mathsf{PA}}$ (see, for example, [Kra19, §7.4.3] for a definition), together the *Induction Scheme*

$$(\varphi(0) \wedge \forall x(\varphi(x) \to \varphi(x+1)) \to \forall x \varphi(x), \tag{$\textsc{Ind}_\varphi$}$$

available for every formula φ.

The language of S^1_2 is the first-order language of bounded arithmetic, $\mathcal{L}_{\mathsf{BA}} \coloneqq \{0, 1, +, \cdot, <, |\cdot|, \lfloor \cdot/2 \rfloor, \#\}$. This extends the language of Peano Arithmetic $\mathcal{L}_{\mathsf{PA}}$ above by the symbols $|x|$, $\lfloor x/2 \rfloor$ and $x\#y$. The standard interpretation of $\lfloor x/2 \rfloor$ is clear. The notation $|x|$ denotes the length of the binary encoding of the number x, $\lceil \log(x+1) \rceil$, while the *smash symbol* $x\#y$ stands for $2^{|x| \cdot |y|}$.

For a term t in the language of bounded arithmetic and a variable x that does not appear in t, a formula of the form $\forall x(x < t \to \varphi(x))$ or $\exists x(x < t \wedge \varphi(x))$ is called a *bounded formula*. The quantifiers guarded by the bounds on x are called *bounded quantifiers* and we simply write $\forall x < t(\varphi(x))$ and $\exists x < t(\varphi(x))$. If the bounded quantifiers are of the form $\forall x < |s|$ of $\exists x < |s|$ for some term s, then they are called *sharply bounded* quantifiers. The *hierarchy of bounded formulas* consists of the classes Σ^b_n (and Π^b_n), for $n \geq 1$, which are defined by counting the alternations of bounded quantifiers ignoring the sharply bounded ones, starting with an existential (respectively, universal) one. The class Δ^b_n consists of all formulas that admit an equivalent definition in both Σ^b_n and Π^b_n. In particular, the class Δ^b_0 stands for all formulas with sharply bounded quantifiers only.

The theory S^1_2 of Buss [Bus86] extends Robinson's arithmetic Q by a set BASIC of simple axioms for the new function symbols (see, e.g., [Kra95, Definition 5.2.1] for the complete list). On top of this, the theory has the *Polynomial Induction* scheme ($\textsc{PInd}$) for Σ^b_1-formulas: for every $\varphi \in \Sigma^b_1$, the theory contains the axiom

$$\varphi(0) \wedge \forall x(\varphi(\lfloor x/2 \rfloor) \to \varphi(x)) \to \forall x \varphi(x). \tag{$\textsc{PInd}_\varphi$}$$

When working over S^1_2, we often invoke instead the schema for *Length Induction*,

$$\varphi(0) \wedge \forall x(\varphi(x) \to \varphi(x+1)) \to \forall x \varphi(|x|), \tag{$\textsc{LInd}_\varphi$}$$

made available for all Σ^b_1-formulas. This form of induction is provable from ($\textsc{PInd}_\varphi$) for $\varphi \in \Sigma^b_1$ [Kra95,

Lemma 5.2.5].

Unlike $\mathcal{L}_{\mathsf{PV}}$, the language $\mathcal{L}_{\mathrm{BA}}$ of bounded arithmetic does not contain a function symbol for every function in **FP**. However, every such $f \in \mathbf{FP}$ is Σ^b_1-definable in S^1_2, meaning that there exists a Σ^b_1 formula whose interpretation over the standard model $\mathbb{N}$ defines f and such that S^1_2 proves the totality of this definition. Thus, in the rest of the paper we choose to use the theory $\mathsf{S}^1_2(\mathcal{L}_{\mathsf{PV}})$, which is the theory S^1_2 in the language of bounded arithmetic extended by all PV function symbols, meaning that we have a fresh symbol for each function in **FP**, and induction is now available for all $\Sigma^b_1(\mathsf{PV})$ formulas. The theory $\mathsf{S}^1_2(\mathcal{L}_{\mathsf{PV}})$ is fully conservative over S^1_2. In what follows we abuse notation and denote this simply as S^1_2.

The final key fact for us is that all $\forall\Sigma^b_1$ formulas provable in S^1_2 are already provable in PV_1. That is, the theory S^1_2 is $\forall\Sigma^b_1$-conservative over PV_1 [Bus86]. We use this in some of the formalizations, where we carry out arguments in S^1_2 but later appeal to its proof in PV_1. For a proof of this fact, see, for example Krajíček's textbook [Kra95, Thm. 5.3.4 and Cor. 7.2.4].

2.3.2 Exact counting in PV_1 and S^1_2

When working in theories of bounded arithmetic, counting the size of different sets requires a great deal of care. We use the wide-spread *Log-notation* $n \in \mathrm{Log}$ as a short-hand for the formula $\exists x(|x| = n)$. A set X is called a *bounded definable set with parameter* y if there exists an arithmetic formula $\varphi(x, y)$ and some term $t(y)$ such that $X = \{x < t(y) \mid \varphi(x, y)\}$. We also adopt the standard set-theoretic notation denoting the interval $[0, a)$ directly by a. Then, a Boolean circuit $C : 2^k \to 2$ naturally defines a bounded definable set $X_C = \{x < 2^k \mid C(x) = 1\}$ with parameter C from which $k \in \mathrm{Log}$ can be extracted, and there exists a PV function Count which counts Log-sized initial segments of circuit-definable sets; i.e., in the standard model, $\mathrm{Count}(C, a)$ is the cardinality of the set $X_C \cap [0, |a|)$.

Importantly, if we know that $2^k \in \mathrm{Log}$, then we are able to *exactly count* the size of X_C when working in PV_1. This proves crucial when carrying out different combinatorial arguments. As an example, arguments involving the pigeonhole principle or different averaging arguments are all possible in PV_1 thanks to exact counting, as long as the sizes of the sets in question are in Log.

2.3.3 Buss's witnessing theorem

Witnessing theorems are a central tool in the field of bounded arithmetic connecting proofs and computation. Roughly speaking, they show that existential quantifiers can be turned into explicit functions computing witnesses for these quantifiers, such that the computational complexity of such a function depends tightly on the logical strength of the theory proving the statement. We mainly rely on Buss's witnessing theorem for S^1_2 (and hence also for PV_1) , capable of witnessing one level of existential quantifiers via polynomial-time functions.

Theorem 2.2 (Buss's witnessing theorem [Bus86])**.** *Let $\varphi(x, y)$ be a Σ^b_1 formula. If $\mathsf{S}^1_2 \vdash \forall x \exists y \varphi(x, y)$, then there exists a* PV *function w such that $\mathsf{S}^1_2 \vdash \forall x \varphi(x, w(x))$.*

Buss's original argument uses proof theory. We refer to Hájek and Pudlák [HP93, Theorem 4.32] for a model-theoretic proof.

2.3.4 Cook's propositional translation

Following Krajíček [Kra19, §8.6], we say that a theory of arithmetic T *corresponds* to a propositional proof system S if (i) T can prove the soundness of S and (ii) every universal consequence $\forall x \varphi(x)$ of T, where φ is quantifier-free, admits polynomial-size proofs in S when suitably grounded into a sequence of propositional formulas. (Pudlák alternatively says that S is a *weak system* of the theory T [Pud20].)

We are interested in the correspondence between PV_1 and Extended Frege (EF). In this case, the process used to turn first-order formulas into propositional ones is known as (Cook's) *propositional translation*, introduced in his seminal paper on PV [Coo75]. Given a quantifier-free formula $\varphi(x)$, Cook's translation is a polynomial-time construction sending φ to a sequence of polynomial-size propositional formulas $\{[\![\varphi]\!]_n\}_{n\in\mathbb{N}}$ such that for every $n \in \mathbb{N}$, the formula $[\![\varphi]\!]_n \in \textsc{Taut}$ if and only if $\mathbb{N} \models \varphi(n)$. See [Kra19, §12.3] or [Bus97, §6.1] for a complete definition of the construction. Cook then observed that, under this translation, PV_1 and EF do indeed correspond to each other.

Theorem 2.3 (Cook's correspondence theorem [Coo75])**.** *The theory* PV_1 *and the proof system Extended Frege correspond to each other. That is,*

(i) PV_1 *proves the soundness of* EF*;*

(ii) if $\varphi(x)$ *is a quantifier-free formula in the language* $\mathcal{L}_{\mathrm{BA}}(\mathsf{PV})$ *and* $\mathsf{PV}_1 \vdash \forall x\varphi(x)$*, then there exists a polynomial-time computable function* f *such that for every* $n \in \mathbb{N}$*, it holds that* $f(1^n) : \mathsf{EF} \vdash [\![\varphi]\!]_n$*.*

A proof of the theorem can be found in Krajíček's textbook [Kra19, Theorem 12.4.2].

2.3.5 Strong proof systems of theories of arithmetic

Let T be a consistent theory extending Robinson Arithmetic Q by a set of axioms that is decidable in polynomial time (i.e., T extends Q and is *polynomial-time axiomatizable*). We denote by P_T the propositional proof system in which a proof of the propositional tautology φ is given by a T-proof of the first-order sentence $\mathrm{Taut}(\ulcorner\varphi\urcorner)$ which states that the propositional formula φ is made true by every truth-assignment for its variables; here $\ulcorner\varphi\urcorner$ stands for the code of φ in a standard arithmetization of propositional formulas. Following Pudlák [Pud20], we call P_T the *strong proof system of* T.

We will use two key facts of these proof systems. First, if T is a theory for which P_T is defined and Q is a propositional proof system that corresponds to T in the sense of Section 2.3.4, then $P_T \geq_{\mathrm{p}} Q$ [Pud20, §4.2, Fact 2]. Second, by Gödel's second incompleteness theorem, T does not prove the soundness of P_T [Pud20, §4.2, Fact 3].

2.4 The Ref formulas

The main character in this paper is the so-called Ref *formula*. Given a propositional CNF formula φ and a size parameter $s \in \mathbb{N}$, the formula $\mathrm{Ref}_s(\varphi)$ states that φ has a Resolution refutation consisting of s clauses.

It is important to choose an encoding that is simultaneously natural from a modeling point of view while not making the formulas artificially hard to refute. At a basic level, the main property that such an encoding should satisfy is that for a concrete φ and s, the formula $\mathrm{Ref}_s(\varphi)$ should be satisfiable if and only if there is a Resolution refutation of φ in s clauses. It also seems natural to require that a Resolution refutation should be readable in polynomial time from a satisfying assignment to Ref.

While different encodings have appeared in the literature, they tend to agree on a few basic ideas. The formula $\mathrm{Ref}_s(\varphi)$ consists of s so-called *blocks* of variables, each representing a clause in the purported Resolution refutation. Each block has variables to represent the literals that appear in this block, how it was obtained (resolved or weakened from an axiom), and it contains *pointer variables* to indicate from which blocks it was derived.

The unary encoding of Pudlák. Pudlák [Pud03] uses the seemingly most standard encoding, which we refer to as the *unary encoding* for Ref. He used it to prove that the canonical pair of Resolution is symmetric. This encoding employs pointers in unary, meaning that for every block $B \in [s]$, there are up to s additional variables to point at the blocks from which B was derived.

The relativized unary encoding of Atserias and Müller. Atserias and Müller [AM20] start by studying Pudlák's encoding. They proved suitable so-called *index-width* lower bounds for it in Resolution, but they were unable to prove a *size* lower bound for it. They then introduced a *relativized* version, in which each block can be possibly *enabled* or *disabled*. If it is disabled, then the block is not used towards the refutation, and its associated clauses are immediately satisfied. These additional *enabling variables* made it possible to prove the size lower bound from the index-width lower bound for the unrelativized encoding. We refer to this second encoding as the *relativized unary encoding*.

We note, however, that the change of encoding is not the source for hardness. Garlík [Gar19] proved that even when using the original encoding of Pudlák, the formulas are hard for Resolution whenever the underlying CNF is unsatisfiable.

The binary encoding of de Rezende *et al.* In their alternative proof of the lower bound on Ref formulas, de Rezende *et al.* [dRGN+21] introduce an encoding of Ref where pointers are encoded in binary. Informally, for every block $B \in [s]$, there are $O(\log s)$ variables used encode the value $B' \in [s]$ of the block(s) from which B was derived. We refer to this as the *binary encoding*. While this encoding also includes the enabling variables of the relativized encoding, these are inessential, since one can always assign the pointers in a dummy fashion to effectively disable a block.

We contend that the unary relativized encoding is both the most natural as well as the most versatile. We see three reasons for this:

1. thanks to the enabling variables, one can naturally turn a Resolution refutation of $t < s$ clauses into a satisfying assignment to $\textsc{Ref}_s(\varphi)$ simply by disabling $s - t$ blocks that are not needed, while in the relativized encoding one needs to fill in the remaining $s - t$ clauses with some dummy content;

2. the enabling variables make the random restriction argument leading to the size lower bound much simpler to prove, and Garlík has shown that the hardness does not comes from this change in syntax;

3. when using a unary encoding rather than the binary one of de Rezende *et al.*, one can easily restrict some pointers to get an instance of $\textsc{Ref}_t(\varphi)$ for every $t < s$, while in the binary encoding, after disabling a block, the binary pointers might still be able to point to it, making the formulas more delicate to handle after applying a restriction.

We remark that the choice between unary and binary encodings is ultimately inessential, and all the results in this paper can be reproven for the binary encoding. We choose the unary encoding mainly for reason (3) above, which simplifies the write-up.

We now define the formula in detail.

The variables of $\textsc{Ref}_s(\varphi)$. Here, we assume φ is a CNF formula over n variables $x_1, \dots, x_n$ and m clauses and define the following variables, where $\mathrm{Lit}_n := \{x_1, \dots, x_n, \neg x_1, \dots, \neg x_n\}$.

$\mathsf{a\text{-}lit}_\ell^A$	:	literal $\ell \in \mathrm{Lit}_n$ is present in the clause $A \in [m]$ of φ;
enable^B	:	block $B \in [s]$ is enabled;
$\mathsf{derived}^B$	:	block $B \in [s]$ was obtained by a Resolution step;
weak_A^B	:	block $B \in [s]$ is obtained by weakening from clause $A \in [m]$ of φ;
lit_ℓ^B	:	literal $\ell \in \mathrm{Lit}_n$ is present in the block $B \in [s]$;
$\mathsf{res}_{x_i}^B$	:	block $B \in [s]$ was obtained by resolving over the variable x_i;
$\mathsf{lpoint}_{B'}^B$	:	block $B \in [s]$ was resolved on the left from block $B' \in [s]$, $B' < B$;
$\mathsf{rpoint}_{B'}^B$	:	block $B \in [s]$ was resolved on the right from block $B' \in [s]$, $B' < B$.

Building on these variables, the $\textsc{Ref}_s(\varphi)$ formula is defined as follows. We write the clauses as implications for the sake of readability.

Definition 2.4 (The Ref formulas)**.** Let $n, m, s \in \mathbb{N}$, and let $\mathrm{Lit}_n \coloneqq \{x_1, \ldots, x_n, \neg x_1, \ldots, \neg x_n\}$ denote the of possible literals over n variables. The $\textsc{Ref}_s(\varphi)$ formula is built from the variables defined above, together with the conjunction of the following clauses:

$$\left(\mathsf{enable}^B \wedge \mathsf{res}^B_{x_i} \wedge \mathsf{lpoint}^B_{B'} \wedge \mathsf{lit}^{B'}_{\ell}\right) \to \mathsf{lit}^B_{\ell} \quad \text{for } B, B' \in [s], B' < B, i \in [n], \ell \in \mathrm{Lit}_n \setminus \{x_i\}, \tag{Ref-1}$$

$$\left(\mathsf{enable}^B \wedge \mathsf{res}^B_{x_i} \wedge \mathsf{rpoint}^B_{B'} \wedge \mathsf{lit}^{B'}_{\ell}\right) \to \mathsf{lit}^B_{\ell} \quad \text{for } B, B' \in [s], B' < B, i \in [n], \ell \in \mathrm{Lit}_n \setminus \{\neg x_i\}, \tag{Ref-2}$$

$$\left(\mathsf{enable}^B \wedge \mathsf{weak}^B_A \wedge \text{a-}\mathsf{lit}^A_{\ell}\right) \to \mathsf{lit}^B_{\ell} \quad \text{for } B \in [s], A \in [m], \ell \in \mathrm{Lit}_n, \tag{Ref-3}$$

$$\left(\mathsf{enable}^B \wedge \mathsf{derived}^B\right) \to \bigvee_{i \in [n]} \mathsf{res}^B_{x_i} \quad \text{for } B \in [s], \tag{Ref-4}$$

$$\left(\mathsf{enable}^B \wedge \mathsf{derived}^B\right) \to \bigvee_{\substack{B' \in [s] \\ B' < B}} \mathsf{lpoint}^B_{B'} \quad \text{for } B \in [s], \tag{Ref-5}$$

$$\left(\mathsf{enable}^B \wedge \mathsf{derived}^B\right) \to \bigvee_{\substack{B' \in [s] \\ B' < B}} \mathsf{rpoint}^B_{B'} \quad \text{for } B \in [s], \tag{Ref-6}$$

$$\left(\mathsf{enable}^B \wedge \neg\mathsf{derived}^B\right) \to \bigvee_{A \in [m]} \mathsf{weak}^B_A \quad \text{for } B \in [s], \tag{Ref-7}$$

$$\left(\mathsf{enable}^B \wedge \mathsf{lpoint}^B_{B'}\right) \to \mathsf{enable}^{B'} \quad \text{for } B, B' \in [s], B' < B \tag{Ref-8}$$

$$\left(\mathsf{enable}^B \wedge \mathsf{rpoint}^B_{B'}\right) \to \mathsf{enable}^{B'} \quad \text{for } B, B' \in [s], B' < B \tag{Ref-9}$$

$$\neg\mathsf{lit}^s_{\ell} \quad \text{for } \ell \in \mathrm{Lit}_n, \tag{Ref-10}$$

$$\mathsf{enable}^s. \tag{Ref-11}$$

Remark 2.5. Our encoding of Ref has fewer axioms than that of [AM20]. For example, we do not require that if a block B is resolved on the left by variable x from block B', then B' should contain x, or we do not require that for every resolution step there is a unique resolved variable. We remark that soundness still holds and Ref is satisfiable if and only if there exists a size-s refutation, which can easily be read from the satisfying assignment to the Ref formula. We also remark that the lack of these axioms does not affect the extraction algorithm or the lower bound in any way: while removing axioms could in principle make the lower bound easier to prove, the algorithm works just as well if we added the missing axioms, and our lower bound proof still goes through with the additional axioms. This more succinct encoding, however, makes it easier to formalize the upper bound construction in Resolution. ■

Remark 2.6 (Number of variables). The formula $\textsc{Ref}_s(\varphi)$ is defined over $N = \Theta(s^2 + sm + sn + mn)$ variables and $M = \Theta(s^2n^2 + smn)$ clauses. For the case when the a-lit variables are restricted to encode a k-CNF formula over n variables and $s = n^c$ for some constant $c \geq 1$, we have $m = O(n^k)$ and $N = O(n^{\max\{2c, c+k\}})$. ■

Blocks and block-width. If a variable is part of a block B_i, we say that it *mentions* B_i. An important measure for us will be the *block-width* of a given clause C over the variables of $\textsc{Ref}_s(\varphi)$. This is defined as the number of different blocks mentioned by the variables of the literals in C, not counting the root block $B_\perp$. We denote this measure by $\mathrm{bw}(C)$, and generalize it to refutations by taking $\mathrm{bw}(\pi)$ to be the maximum block-width over all the clauses in π.

2.4.1 The Ref formulas for other proof systems

We will also be interested in the Ref formulas for proof systems other than Resolution. In general, for a Cook-Reckhow system Q, we denote by $\text{Ref}^Q(\varphi, \pi)$ the formula stating that π is a correct Q-refutation of an unsatisfiable φ. For convenience, in this context we always consider all proof systems as refutational systems. The formula $\text{Ref}^Q(\varphi, \pi)$ is simply the propositional formula that verifies that φ is accepted by the Boolean circuit that checks Q-refutations. This can be obtained by writing the computation of the circuit as a Boolean formula using the usual Tseitin encoding.

Remark 2.7 (Notation). By default, the Ref formula stands for the formula as defined in Definition 2.4 for Resolution refutations. If we want to refer to the Ref formula for a different proof system, we explicitly write Ref^Q for the system Q in question. ■

2.4.2 The Sat formula and reflection principles

We also have have the $\text{Sat}(\varphi, \alpha)$ formula, encoding that a CNF formula φ is satisfied by an assignment α. The variables we consider are

α_i	:	value assigned by α to variable x_i;
a-lit_ℓ^A	:	literal $\ell \in \text{Lit}_n$ is present in clause $A \in [m]$;
sat_ℓ^A	:	clause $A \in [m]$ is satisfied because literal $\ell \in \text{Lit}_n$ evaluates to 1 under α.

We use a-lit instead of lit to distinguish between these variables and the lit variables of $\text{Ref}(\varphi, s)$.

Definition 2.8 (The Sat formulas)**.** Let $n, m \in \mathbb{N}$, let φ denote the set of variables of the form a-lit_ℓ^A as above and let α denote the set of variables α_i as above. The formula $\text{Sat}(\varphi, \alpha)$ is the CNF formula over the variables in φ, α and additionally all the variables sat_ℓ^A above consisting of the conjunction of the following clauses,

$$\neg\text{sat}_\ell^A \vee \text{a-lit}_\ell^A \qquad \text{for all } A \in [m] \text{ and } \ell \in \{x_1, \dots, x_n, \neg x_1, \dots, \neg x_n\}, \tag{Sat-1}$$

$$\neg\text{sat}_{x_i}^A \vee \alpha_i \qquad \text{for all } A \in [m] \text{ and } i \in [n], \tag{Sat-2}$$

$$\neg\text{sat}_{\neg x_i}^A \vee \neg\alpha_i \qquad \text{for all } A \in [m] \text{ and } i \in [n], \tag{Sat-3}$$

$$\bigvee_{i \in [n]} \text{sat}_{x_i}^A \vee \text{sat}_{\neg x_i}^A \qquad \text{for all } A \in [m]. \tag{Sat-4}$$

One can similarly write a Sat formula for evaluating DNF formulas in the obvious ways. It is always clear from context which version of Sat we are using, so we use the same notation for both.

Proposition 2.9. *For every CNF formula φ, the following statements hold:*

(i) if φ has a Resolution refutation of size s, then $\text{Sat}_{\restriction\varphi}$ has a Resolution refutation of size $O(s)$;

(ii) if $\text{Sat}_{\restriction\varphi}$ has a Resolution refutation of size s, then φ has a Resolution refutation of size $O(s)$.

Proof. Assume φ has n variables $x_1, \dots, x_n$ and m clauses $C_1, \dots, C_m$. We first show how to go from a refutation of $\text{Sat}_{\restriction\varphi}$ to a refutation of φ. After fixing φ, the only variables left in Sat are of type sat_ℓ^A or α_i. Since Resolution is closed under literal substitutions, it suffices to define a substitution σ that replaces all the original variables by literals of φ, and argue that all the axioms of $\text{Sat}_{\restriction\varphi,\sigma}$ follow from clauses of φ. For every $i \in [n]$, every $\ell \in \{x_1, \dots, x_n, \neg x_1, \dots, \neg x_n\}$, and every $A \in [m]$, the substitution σ maps

$$\sigma(\alpha_i) \coloneqq x_i \qquad \text{and} \qquad \sigma(\text{sat}_\ell^A) \coloneqq \begin{cases} \ell & \text{if } \ell \in C_A \\ 0 & \text{if } \ell \notin C_A. \end{cases} \tag{2.1}$$

Let us inspect the axioms of $\text{SAT}_{\restriction\varphi,\sigma}$. For an axiom $\neg\text{sat}^A_\ell \vee \text{a-lit}^A_\ell$ of type (SAT-1), if the restriction of this axiom is present is $\text{SAT}_{\restriction\varphi}$ it is because $\ell \notin C_A$, or else a-lit^A_ℓ would be satisfied. Then, $\sigma(\text{sat}^A_\ell) = 0$ and the substitution is satisfied. If the axiom is of type (SAT-2), then either $x_i \notin C_A$, in which case $\sigma(\text{sat}^A_{x_i}) = 0$ and $\neg\text{sat}^A_{x_i} \vee \alpha_i$ is satisfied, or $\sigma(\text{sat}^A_{x_i}) = x_i$ and $\sigma(\alpha_i) = x_i$, which gives the trivial clause $\neg x_i \vee x_i$. The case for (SAT-3) is analogous. Finally, for a clause of type (SAT-4), it is easy to see that the substitution σ maps the clause precisely to the clause C_A itself, which is a clause of φ.

To go from a refutation of φ to a refutation of $\text{SAT}_{\restriction\varphi}$, we do the following. First, rename all the variables α_i by x_i. Then, for every $A \in [m]$, note that we can resolve the corresponding clause of type (SAT-4) with the unit clauses $\neg\text{sat}^A_\ell$ of type (SAT-1) to obtain $\bigvee_{\ell \in C_A} \text{sat}^A_\ell$. Now, for each sat^A_ℓ in this clause, cut with the corresponding clause of type (SAT-2) or (SAT-3) to obtain $\bigvee_{\ell \in C_A} \ell$, which is just C_A. In this way we have derived every axiom C_A of φ, and we can now proceed with the refutation of φ in the natural way. □

Finally, we define the reflection principle for any proof system.

Definition 2.10 (The REFL formulas)**.** Let Q be a Cook-Reckhow propositional proof system and let $n, m, s \in \mathbb{N}$. We define $\text{REFL}^Q_{n,m,s} := \text{SAT}(\varphi, \alpha) \wedge \text{REF}^Q(\varphi, \pi)$ where the SAT instance is for CNF formulas with n variables and m clauses and the REF^Q instance is for Q-refutations of such formulas of size s, and refer to the sequence of tautologies $\text{REFL}^Q := \{\neg\text{REFL}^Q_{n,m,s}\}_{n,m,s\in\mathbb{N}}$ as the *reflection principle for* Q.

A useful property of EF is the fact that any propositional system S, however strong, can always be seen as a Frege-like system due to the fact that $\text{EF} + \text{REFL}^S \geq_{\text{p}} S$ [Kra19, Theorem 8.4.3].

2.4.3 Reflection principles for first-order theories

The reflection principles above are the propositional analogues of the well-studied first-order reflection principles for theories of arithmetic. For a strong enough recursively axiomatizable theory of arithmetic T and a class Φ of sentences, the schema Φ-Refl$_T$ stands for the collection of all formulas of the form

$$\exists\pi\, \text{Prf}_T(\ulcorner\varphi\urcorner, \pi) \to \varphi \qquad (\text{Refl}_{T,\varphi})$$

for every $\varphi \in \Phi$. Here, the notation $\ulcorner\varphi\urcorner$ stands for the encoding of φ in a fixed suitable arithmetization of syntax, and Prf_T stands for the provability statement in this arithmetization.

3 The Proof Analysis Problem: definitions and basic facts

For a CNF formula $\varphi(x_1, \ldots, x_n)$, we denote by $\text{REF}_s(\varphi)$ the propositional formula claiming that there exists a Resolution refutation of φ in size s. Different encodings of this formula have been considered in the literature. For our purposes, REF consists of s of *blocks* of variables, each of them describing a clauses in a purported Resolution refutation of φ of size s. (See Section 2.4 for a full rendering of the variables and clauses involved in $\text{REF}_s(\varphi)$.)

We are interested in the following decision problem.

Definition 3.1 (The Proof Analysis Problem, PAP_Q)**.** Let Q be a propositional proof system. We define the *Proof Analysis Problem for* Q to be the language

$$\text{PAP}_Q := \{(\varphi, \pi, 1^s) \mid \varphi \in \text{SAT and } \pi : Q \vdash \neg\text{REF}_s(\varphi)\}.$$

We denote by $\text{PAP}_Q[s(n)]$ the problem where the size parameter s is restricted to be at least $s(n)$ and n denotes the number of variables of φ.

The problem asks, given the proof of a Resolution lower bound in a fixed proof system Q, to decide whether the underlying formula is satisfiable or not. Note that whenever φ is satisfiable there is no Resolution refutation and thus any lower bound holds, so the problem is well-defined.

Analogous to the notion of whether a proof system is automatable, PAP naturally induces a notion of whether, for a given proof system, its Resolution lower bounds are "analyzable".

Definition 3.2 (Analyzability)**.** We say that a propositional proof system Q is *analyzable* if there exists some constant $c > 0$ such that $\mathrm{PAP}_Q[n^c] \in \mathbf{P}$.

Remark 3.3. It might seem more intuitive to define a proof system Q to be analyzable if $\mathrm{PAP}_Q \in \mathbf{P}$, without restrictions on the size parameter. Note, however, that for most reasonable proof systems, the language PAP_Q taken as a whole contains some degenerate instances that make the problem trivially **NP**-hard. For example, if the size parameter is set to $s = 1$, then certainly proving a Resolution lower bound against φ is easy already for Resolution itself, and we can map a CNF formula φ to the PAP_Q-instance $(\varphi, \pi, 1)$ for some easy to construct Q-proof π that checks there is no Resolution refutation of φ in one clause. ■

It is easy to see that for every Cook-Reckhow system Q, the problem PAP_Q is in **NP**. Similarly, we note that unlike automatability, analyzability is naturally downwards-closed under p-simulations. Namely, if S is p-simulated by Q and Q is analyzable, so is S; this is not the case with automatability, where a search algorithm for Q may not be used to search for proofs in a weaker S.

The following is a corollary of the results of Atserias and Müller [AM20]. Here, by **P**-uniform we mean the standard notion of uniformity by which there is a polynomial-time descriptor Turing machine that on input 1^ℓ outputs the circuit solving the problem for inputs of size ℓ (see, e.g., [AB09, Definition 6.12; All25]).

Proposition 3.4. *It holds that* $\mathrm{PAP}_{\mathsf{Res}}[n^2]$ *is in* **P***-uniform* $\mathbf{AC}^0$*. That is, Resolution is analyzable.*

Proof. Let us first describe the general polynomial-time algorithm that puts $\mathrm{PAP}_{\mathsf{Res}}[n^2]$ in **P**, and we later elaborate on how this can be computed in **P**-uniform $\mathbf{AC}^0$. Indeed, by the Resolution lower bound on Ref formulas (Theorem 1.5), there exists $\varepsilon > 0$ such that for every $s \in \mathbb{N}$, if a formula φ over n variables is unsatisfiable, then a correct Resolution refutation π of $\mathrm{Ref}_s(\varphi)$ must have size $|\pi| > 2^{\varepsilon \cdot s/n}$. Given an input $(\varphi, \pi, 1^s)$ to $\mathrm{PAP}_{\mathsf{Res}}[n^2]$, to decide if the instance belongs in the language, it suffices to check (i) that π is a correct Resolution refutation of $\mathrm{Ref}_s(\varphi)$ and (ii) that $|\pi|$ is smaller than the lower bound $2^{\varepsilon \cdot s/n}$. If (i) fails, we immediately reject, and otherwise, if (ii) fails, the input size is large enough to brute-force SAT in polynomial time. Here we use the fact that $s \geq n^2$, hence $|\pi| \geq 2^{\varepsilon \cdot s/n} \geq 2^{\varepsilon n}$, and thus the input size is large enough.

Let us now argue that this entire computation is possible within **P**-uniform $\mathbf{AC}^0$. For the sake of precision, let us fix the following natural binary encoding for $\mathrm{PAP}_{\mathsf{Res}}$. An input $(\varphi, \pi, 1^s)$ will be of the form $(1^n, 1^m, C_1, \ldots, C_m, 1^t, \pi, 1^s)$. Here, the first part of the tuple corresponds to the encoding of φ, a CNF formula over n variables and m clauses $C_1, \ldots, C_m$, and we assume that these clauses are initially represented as strings of length $2n$ with indicators for every possible literal. The Resolution refutation of $\mathrm{Ref}_s(\varphi)$ is encoded by 1^t and π, where π is an assignment to the $N := N(n, m, t, s) = \mathrm{poly}(n, m, t, s)$ variables of $\mathrm{Ref}_t(\mathrm{Ref}_s(\varphi))$, as per Section 2.4, and the number of variables N can be easily computed in polynomial time. For the purpose of unique decoding, we assume that the tuple $(1^n, 1^m, C_1, \ldots, C_m, 1^t, \pi, 1^s)$ is encoded by bit-doubling: each bit is duplicated and 01 is used as separators. We assume that there are no separators between the clauses $C_1, \ldots, C_m$, so a correct input contains only five separators.

Now, when dealing with binary strings of even length ℓ, there are at most $O(\ell^4)$ possible ways of interpreting such strings as a tuple of the form $(1^n, 1^m, C_1, \ldots, C_m, 1^t, \pi, 1^s)$. This is because we can choose values for n, m, t and s in the interval $[\ell/2 - 5]$, where $\ell/2 - 5$ comes from the fact that we duplicated every bit and introduced 10 bits for the five separators between 1^n, 1^m, $C_1, \ldots, C_m$, 1^t, π, and 1^s, not counting separators between C_1 and C_m. One can then check that this choice of n, m, t and s conforms to the desired

pattern: the segment for the clauses C_1 to C_m has length exactly $2nm$, and the segment for π has length exactly $N = N(n, m, t, s)$, as per Remark 2.6. That is, it must hold that $\ell = 2(n+m+2nm+t+N(n, m, t, s)+s)+10$ and $s \geq n^2$. There are at most $O(\ell^4)$ such choices for $(n, m, t, s) \in [\ell/2 - 5]^4$, hence the upper bound. Furthermore, it is easy to see that, due to the bit-doubling, every string can only encode correctly one input of the form $(1^n, 1^m, C_1, \ldots, C_m, 1^t, \pi, 1^s)$, so the decoding is unique.

The **P**-uniform descriptor machine for inputs of even length ℓ now works as follows. On input 1^ℓ, for ℓ even, it tries all possible $O(\ell^4)$ ways of separating the lengths, and for each interpretation (n, m, t, s) of the lengths it constructs a different constant-depth Boolean circuit $D_{n,m,t,s}$, as follows.

(a) If the interpretations of the lengths is inconsistent, in the sense that the string cannot correspond to something of the form $(1^n, 1^m, C_1, \ldots, C_m, 1^t, \pi, 1^s)$, then it outputs the constant circuit 0.

(b) If the interpretation of the lengths is valid and $|\pi| < 2^{\varepsilon \cdot s/n}$, then it simply constructs the circuit that checks that π is a correct Resolution refutation of $\textsc{Ref}_s(\varphi)$ in at most t clauses; that is, it outputs the formula $\textsc{Ref}_t(\textsc{Ref}_s(\varphi))$, which itself depends on the variables encoding φ. Since Ref formulas are in CNF, nesting these together with φ will result in a total depth of 5 (see Section 7.3 for a more detailed treatment of how the depth increases when nesting the Ref formulas).

(c) If the interpretation of the lengths is valid and $|\pi| \geq 2^{\varepsilon \cdot s/n}$, then the descriptor outputs the conjunction of two circuits: one is the same as before, checking the correctness of π as a refutation of $\textsc{Ref}_s(\varphi)$ in at most t clauses, and the other is the trivial circuit of size $\text{poly}(n, m) \cdot 2^n$ that brute-forces the satisfiability of φ. More formally, this is a big disjunction of fan-in 2^n, where each wire goes to the formula $\textsc{Sat}(\varphi, \alpha)$ from Definition 2.8 for different hard-wired values of $\alpha \in \{0, 1\}^n$. Since $\textsc{Sat}(\varphi, \alpha)$ is a CNF formula, this brute-forcing circuit has depth 3, and combined with the circuit checking the correctness of π, the entire circuit has depth 5 in this case.

For each interpretation of the lengths there is also a circuit $\text{Correct}_{n,m,t,s}$ that verifies that the input correctly encodes a PAP_{Res} instance of the right size. This amounts to checking that the separators are in the right place and the double-bit encoding is correctly implemented, which can all be verified in depth 3.

Finally, the descriptor machine outputs the circuit

$$R_\ell := \bigvee_{(n,m,s,t) \in [\ell/2-5]^4} D_{n,m,t,s} \wedge \text{Correct}_{n,m,t,s} \tag{3.1}$$

consisting of the disjunction of all the circuits above for every interpretation of the lengths.

The final circuit R_ℓ correctly computes $\text{PAP}_{\text{Res}}[n^2]$ on inputs of even length ℓ, has depth 6 and polynomial size. Indeed, the constructions (a) and (b) above both have polynomial size, and whenever we construct the exponential-size circuit in (c), it is with respect to a segment of the string that has itself size exponential in n. Finally, the descriptor machine runs in polynomial time given only the length ℓ of the input string x, so we can conclude that $\text{PAP}_{\text{Res}}[n^2]$ is in **P**-uniform $\mathbf{AC}^0$. □

The fact that PAP_{Res} is so easy makes it natural to ask whether the same is true for the *search version* of the problem.

Definition 3.5 (Search version of PAP)**.** For a propositional proof system Q, we denote by FPAP_Q the *search version of the Proof Analysis Problem for* Q, defined as follows.

FPAP_Q (search version of PAP_Q)	
Input	A CNF formula φ, a size parameter s in unary and a proof π such that $\pi : Q \vdash \neg\textsc{Ref}_s(\varphi)$.
Output	Either a satisfying assignment for φ, if $\varphi \in$ SAT, or 0 otherwise.

Similarly to $\mathrm{PAP}_Q[s(n)]$, we define $\mathrm{FPAP}_Q[s(n)]$ to be the search problem where the size parameter is at least $s(n)$.

If we impose a polynomial upper bound on the size of π, then by the lower bound on Ref formulas, there is always a satisfying assignment for φ, and the problem $\mathrm{FPAP}_Q[n^c]$ for any $c > 0$ is in **TFNP**. The fact that $\mathrm{PAP}_{\mathrm{Res}}[n^2] \in \mathbf{P}$, does not, however, directly imply that $\mathrm{FPAP}_{\mathrm{Res}}[n^c] \in \mathbf{FP}$ for any $c > 0$. Namely, it is not clear that given a polynomial-size proof π of $\mathrm{Ref}_s(\varphi)$ one can extract a satisfying assignment of φ, even if one can conclude that φ is satisfiable. We show in Section 4 that $\mathrm{FPAP}_{\mathrm{Res}}$ is in **FP**—although the algorithm is not quite as straightforward as the one for the decision problem.

We see PAP and the analyzability of a proof system as closely related to automatability. The following proposition captures this idea and underlines the relevance on PAP in showing hardness of automatability.

Proposition 3.6. *Let $Q \geq$ Res. If Q is both analyzable and automatable, then* $\mathbf{P} = \mathbf{NP}$.

Proof. If Q is analyzable and automatable, this means that $\mathrm{PAP}_Q[n^c] \in \mathbf{P}$ for some constant $c > 0$, and that there is an automating algorithm A for Q. We call the polynomial-time algorithm for $\mathrm{PAP}_Q[n^c]$ an *analyzer* and observe that these two combined can solve 3SAT in polynomial time as follows. Given a 3-CNF formula φ, construct the formula $\mathrm{Ref}_{n^c}(\varphi)$, stating that φ does not have Resolution refutations of size n^c. Since $Q \geq$ Res, by the upper bound construction [Pud03, Theorem 4.1; AM20, Lemma 11], whenever φ is satisfiable, there will be size-$n^{O(1)}$ refutations in Resolution, and hence also in Q, and the automating algorithm A will succeed in finding some refutation in polynomial time. Feed this refutation to the Q-analyzer to decide whether $\varphi \in$ SAT. If the automating algorithm failed to output a polynomial-size proof, then we would already know that $\varphi \notin$ SAT. □

The previous proposition can be seen as an abstract way of stating the **NP**-hardness of automating Resolution too. Since $\mathrm{PAP}_{\mathrm{Res}}[n^2] \in \mathbf{P}$, that means that Resolution cannot be automatable unless $\mathbf{P} = \mathbf{NP}$. In Section 5 we study the possibility of analyzing algorithms for strong proof systems actually leading to the hardness of their automatability—and establish that this is highly unlikely.

4 The extraction algorithm

This section proves that the search version of the Proof Analysis Problem for Resolution is in **FP**. Our algorithm (in fact, two algorithms) arise from closely observing the lower bound on the Ref formulas and attempting to make it fully constructive, in a style that would make them amenable to formalization in weak theories of arithmetic like PV_1 (in the style of Cook and Pitassi [CP90]).

Recall that the lower bound can be presented in two steps: first, a random restriction argument takes a small refutation and produces a low block-width refutation of a restricted formula, followed by a block-width lower bound for this restricted formula, which overall bounds the size of the original refutation.

Our algorithm works analogously. On input a proof π of $\mathrm{Ref}_s(\varphi)$, it first finds a restriction ρ such that $\pi_{\restriction\rho}$ is a refutation of a restricted version of $\mathrm{Ref}_s(\varphi)$ and has low block-width. Then, we have a second algorithm that, inspired by the proof of the width lower bound, analyses this low block-width refutation and extracts a satisfying assignment.

We present the algorithm in two steps. First, two alternatives to perform block-width reduction are described in Section 4.1. These correspond, respectively, to a random restriction and a deterministic restriction argument similar to the one in the proof of the lower bound for the Ref formulas. In Section 4.2 we explain how to design the algorithm that analyzes low block-width refutations, which is essentially the Prover-Delayer strategy behind the block-width lower bound for the Ref formulas. Putting them together yields the desired procedure.

4.1 The block-width reduction algorithm

Recall that we crucially assume that in our definition of the Ref formula there is a variable enable^i for every block $i \in [s]$ that allows us to *disable* that block (see Definition 2.4). For succinctness, we often denote enable^i simply by e_i and refer to is as an *enabling variable*.

Let us first define a kind of restriction that will come up a lot in our arguments.

Definition 4.1 (Disabling restrictions)**.** We say that a restriction $\rho \in \{0,1,*\}^N$ to the N variables of $\text{Ref}_s(\varphi)$ is *d-disabling* if it satisfies that (i) exactly d blocks are disabled, and the rest are all enabled, (ii) every variable belonging to a disabled block is assigned a value, and (iii) no other variable is assigned.

A key property of disabling assignments is that they can never falsify any axioms of $\text{Ref}_s(\varphi)$, all the enabling variables disappear after the restriction and $\text{Ref}_s(\varphi)_{\restriction\rho}$ is essentially an instance of $\text{Ref}_{s-d}(\varphi)$, except there are some pointer variables pointing to the d disabled blocks that are still hanging.

A first approach to perform block-width reduction in inspired by random restriction arguments, and requires randomness.

Lemma 4.2 (Randomized block-width reduction)**.** *Let $p \in [0,1)$ and let N be the number of variables of the $\text{Ref}_s(\varphi)$ formula for some CNF formula φ over n variables. There exists a randomized algorithm R taking as input 1^N, and outputting an $\lfloor s/2 \rfloor$-disabling restriction $\rho \in \{0,1,*\}^N$, such that for every Resolution refutation π of the formula $\text{Ref}_s(\varphi)$, the following properties hold:*

(i) the restriction ρ does not falsify $\text{Ref}_s(\varphi)$;

(ii) with probability at least p, the block-width of $\pi_{\restriction\rho}$ is at most $O\left(\log|\pi| - \log(1-p)\right)$;

(iii) the running time of $R(1^N)$ is $O(N)$.

Proof. The proof follows closely the random restriction argument of de Rezende *et al.* [dRGN+21, Section 6.3]. For simplicity, let us assume s is even. (Note that, without loss of generality s can be even, since if π is a correct refutation of $\text{Ref}_s(\varphi)$ and s is odd, then π can be turned into a refutation of essentially $\text{Ref}_{s-1}(\varphi)$ by hitting π with the restriction that disables and fully restricts one block).

Assume the root block corresponds to block $B_\perp$, which is not counted towards block-width, and consider the following random restriction: pair all s blocks into $s/2$ pairs, and for each pair, with probability $1/2$, decide which block in the pair is going to be disabled. Now, if a block is disabled, all of its remaining variables are assigned uniformly at random. We denote by ρ the restriction obtained in this way, which the algorithm outputs.

We claim that with probability at least p, the restriction succeeds in lowering the block-width of any Resolution derivation π to $O(\log|\pi| - \log(1-p))$. Indeed, if ℓ is a literal corresponding to the variable e_i determining whether a certain block is disabled, then $\Pr_\rho[\ell_{\restriction\rho} = 1] = 1/2$. For every other literal ℓ in a block B_i,

$$\Pr_\rho[B_i \text{ is disabled and } \ell_{\restriction\rho} = 1\ ] = \Pr_\rho[B_i \text{ is disabled}] \cdot \Pr_\rho[\ell_{\restriction\rho} = 1 \mid B_i \text{ is disabled}] \tag{4.1}$$

$$= \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}\,. \tag{4.2}$$

Hence, for every literal ℓ not from $B_\perp$, $\Pr_\rho[\ell_{\restriction\rho} = 1] \geq 1/4$.

Now, if C is a clause of block-width at least w, we have that

$$\Pr_\rho[C_{\restriction\rho} \neq 1] \leq (3/4)^{w/2}\,, \tag{4.3}$$

where the $1/2$ in the exponent comes from the fact that if two consecutive blocks are present, meaning that they were paired together and only one of them was enabled, their values depend on each other.

Then, if π was indeed a Resolution derivation of $\textsc{Ref}_s(\varphi)$, by a union bound,

$$\Pr_{\rho}[\pi_{\restriction\rho} \text{ has a clause of block-width } \geq w] \leq \text{length}(\pi) \cdot (3/4)^{w/2} \leq |\pi| \cdot (3/4)^{w/2}, \tag{4.4}$$

which is the failure probability for property (ii) in the statement. For success probability at least p, we want to choose w such that $|\pi| \cdot (3/4)^{w/2} \leq 1-p$. This bound is met by choosing $w \geq 2(\log(|\pi|/(1-p))/(\log 4/3))$, meaning that with probability p, the restriction ρ will satisfy all clauses of at least this width.

It suffices to argue that properties (i) and (iii) are also satisfied. Indeed, by the way we designed the restriction, after applying ρ there are no disabling variables left and all variables in the disabled blocks have been restricted, to this is exactly $s/2$-disabling.

As for the running time, the algorithm is simply sampling the restriction, which takes time $O(N)$. □

We now move on to a fully deterministic algorithm that takes as input an actual refutation π and outputs a restriction that *always* manages to reduces the block-width.

Lemma 4.3 (Deterministic block-width reduction)**.** *There exists a constant $c > 0$ and a deterministic algorithm taking as input a Resolution refutation π of the formula $\textsc{Ref}_s(\varphi)$ over N variables and outputting a d-disabling restriction $\rho \in \{0,1,*\}^N$ with $d \leq c/2 \cdot \left(\sqrt{s \log |\pi|}\right)$ such that*

(i) the restriction ρ does not falsify $\textsc{Ref}_s(\varphi)$;

(ii) the block-width of $\pi_{\restriction\rho}$ is at most $c \cdot \left(\sqrt{s \log |\pi|}\right)$;

(iii) the algorithm runs in time $\text{poly}(|\pi|, s)$.

Proof. We employ a greedy strategy to construct the restriction, meaning that we look at all the clauses of high block-width and we iteratively choose to restrict a literal that kills a significant fraction of these clauses.

More formally, let w be a parameter to be optimized later, and given π, let W denote the set of all clauses in π with block-width at least w. Through the following iterative process we will enable and disable some blocks. Whenever we enable a block, we will also mark it as not active by keeping track of a set $\mathsf{ActiveBlocks} \subseteq [s]$, meaning that when choosing greedily the next literal to restrict, inactive blocks are not a valid choice; and whenever we disable a block, we add it to a set $D \subseteq [s]$ to keep track of it.

Algorithm 4.1: Deterministic block-width reduction

Repeat the following procedure iteratively, starting with $\rho := \emptyset$, $\mathsf{ActiveBlocks} := [s]$, and $D := \emptyset$, and stop whenever W is empty:

1. Find the most frequent block $i \in \mathsf{ActiveBlocks}$ among the ones mentioned in the clauses in W.
2. Look at the literal e_i of the variable used to disable block i.
   (a) If e_i appears positively in at least $1/3$ of all the clauses in W that mention block i, then set $\rho := \rho \cup \{e_i \mapsto 1\}$, $\mathsf{ActiveBlocks} := \mathsf{ActiveBlocks} \setminus \{i\}$, $W := W_{\restriction\rho}$, and go back to step (1).
   (b) If e_i does not appear positively in at least $1/3$ of the clauses in W mentioning block i, then set $\rho := \rho \cup \{e_i \mapsto 0\}$, $W := W_{\restriction\rho}$, $D := D \cup \{i\}$, and for every other variable x of block i,

i. if x appears in W positively more often than negatively, then set $\rho \coloneqq \rho \cup \{x \mapsto 1\}$, and $W \coloneqq W_{\restriction \rho}$;

ii. if x appears in W negatively more often than positively, then set $\rho \coloneqq \rho \cup \{x \mapsto 0\}$, and $W \coloneqq W_{\restriction \rho}$;

iii. repeat for every variable of block i.

(c) Once all the variables of block i have been taken care of, go back to step (1).

The procedure terminates once W is either empty or all clauses in W mention only blocks that are no longer in ActiveBlocks. At this point, $|D|$ blocks have been disabled. For the remaining blocks that were not mentioned by any clause in W, enable all of them by setting the corresponding variables $e_i \mapsto 1$. This completes the construction of the restriction ρ, and the algorithm outputs ρ.

The procedure runs for at most s iterations, since each iteration takes care of one of the blocks mentioned by the clauses in the initial W and we never deal with a block twice. Therefore, the algorithm runs in time $\mathrm{poly}(|\pi|, s)$.

As for the correctness of the algorithm, the restriction ρ is d-disabling by construction for $d = |D|$. It is left to argue that for a suitable choice of w, there exists a constant $c > 0$ such that $d \leq c/2(\sqrt{s \log |\pi|})$ and the block-width of $\pi_{\restriction \rho}$ is at most $c \cdot (\sqrt{s \log |\pi|})$.

We want to choose w so that after $\ell \leq s$ iterations, the set W becomes empty. At the first iteration, by an averaging argument, we know that the most frequent block is mentioned in at least a w/s fraction of $|W|$. More generally, at iteration ℓ, block-width might have decreased up to $w - (\ell - 1)$ and up to $\ell - 1$ blocks may have become inactive, so the same averaging argument tells us that the most frequent active block is mentioned in at least a $(w - (\ell - 1))/(s - (\ell - 1))$ fraction of the clauses. Furthermore, observe that if at a given iteration block i is the most frequent active block, we are not promised to kill all the clauses mentioning i, but we are guaranteed to kill at least $1/3$ of them. Indeed, if we enable block i that is because it appeared in at least $1/3$ of all the active clauses mentioning i; and otherwise we are guaranteed to restrict at least $1/2$ of the remaining $2/3$ fraction of the active clauses mentioning i, which amounts to a $1/3$ fraction.

Therefore, if ρ_ℓ is the restriction built after ℓ iterations,

$$|W_{\restriction \rho_\ell}| \leq |W| \cdot \left(1 - \frac{w}{3s}\right) \cdot \left(1 - \frac{w-1}{3(s-1)}\right) \cdot \dots \cdot \left(1 - \frac{w-(\ell-1)}{3(s-(\ell-1))}\right) \tag{4.5}$$

$$\leq |W| \cdot \left(1 - \frac{w-\ell}{3s}\right)^{\ell} \tag{4.6}$$

$$\leq |W| \cdot e^{-\ell \cdot \frac{w-\ell}{3s}}. \tag{4.7}$$

We want to ensure that for some $\ell \leq s$ we achieve $|W_{\restriction \rho_\ell}| < 1$. It suffices to have $|W| \cdot e^{-\ell \cdot \frac{w-\ell}{3s}} < 1$. Taking logarithms on both sides we have

$$\ln |W| < \ell \cdot \frac{w-\ell}{3s}, \tag{4.8}$$

which holds already for $\ell = w/2$, assuming $w > \sqrt{12s \ln |\pi|} \geq \sqrt{12s \ln |W|}$.

Now it suffices to choose a constant c such that $w \coloneqq c \cdot \sqrt{s \log |\pi|} > \sqrt{12s \ln |\pi|}$. In this way we get that after at most $\ell \coloneqq w/2$ iterations, $W_{\restriction \rho_\ell} = \emptyset$ and thus the block-width of $\pi_{\restriction \rho_\ell}$ is also at most $c \cdot \sqrt{s \log |\pi|}$. Furthermore, note that $|D| \leq \ell$, since the algorithm only runs for at most ℓ iterations, meaning that ρ_ℓ is d-disabling for $d \leq \ell = w/2 = c/2\sqrt{s \log |\pi|}$, as desired. □

4.2 The block-width analysis algorithm

Using one of the two algorithms above, we can take a refutation π of $\text{REF}_s(\varphi)$ and obtain a new refutation π' of the restricted formula $\text{REF}_s(\varphi)_{\restriction\rho}$ in low block-width. We can now show how to analyze this refutation, inspired by the block-width lower bound, and succeed in finding a satisfying assignment whenever one exists.

We first state the following simple but crucial technical fact used in the proof.

Fact 4.4. *Let φ be a Boolean formula in CNF over n variables. If C is a width-n clause over the variables of φ that is not the weakening of any clause of φ, then $\neg C$ encodes a satisfying assignment for φ.*

Now we can present the algorithm and prove its correctness.

Lemma 4.5 (Assignment extraction)**.** *There exists a deterministic algorithm E such that for every $s \in \mathbb{N}$, π a purported Resolution refutation of $\text{REF}_s(\varphi)$ for a CNF formula $\varphi(x_1, \ldots, x_n)$ with m clauses, and $\rho \in \{0, 1, *\}^N$ a d-disabling restriction to the N variables in $\text{REF}_s(\varphi)$, it holds that $E(\varphi, \rho, s, \pi)$ terminates in time $\text{poly}(|\pi|, s, n, m)$ and provides exactly one of the following outputs:*

(a) an incorrect derivation step in π;

(b) a clause $C \in \pi_{\restriction\rho}$ of block-width at least $1/3\lfloor(s - d - n)/n\rfloor$;

(c) a satisfying assignment for φ.

Proof. We traverse the refutation π inspired by the Delayer's strategy in the Prover-Delayer game that yields a block-width lower bound for the restricted REF formulas—and the correctness of the algorithm is essentially the proof of this block-width lower bound.

Before the traversal of π starts, we arrange the $s - d$ blocks enabled by ρ in a layered manner, so that there are n layers, each containing $\lfloor(s - d)/n\rfloor$ blocks (with the remainder blocks left from the flooring operations collected all in the last layer), plus one additional layer on top with a single block corresponding to the root. We see the root as laying at layer 0, and intuitively blocks in layer i will be obtained by resolving over variable x_{i+1}. Throughout the traversal, we keep a record $\alpha \in \{0, 1, *\}^N$ which we call the *reservation*, in which we "reserve" information needed to continue the traversal. (Intuitively, this record keeps the information that the Delayer has in mind when playing against the Prover.)

The algorithm proceeds as follows. First, let $\pi \coloneqq \pi_{\restriction\rho}$, initialize $\alpha \coloneqq \rho$, and collect in a set $D \subseteq [s]$ the d blocks that are disabled by the restriction ρ. Note that ρ is a d-disabling restriction, meaning that it only enables and disables blocks, and sets the value of all the variables in disabled blocks, meaning that the restriction cannot possibly falsify any axiom of $\text{REF}_s(\varphi)$. Note as well that in π, after hitting it with ρ, there are no longer resolution steps over enabling variables nor over variables belonging to a disabled block.

(In what follows, capitals letters in roman font, like A, refer to clauses in the refutation π, while capital letters in calligraphic font, like $\mathcal{A}$, refer to clauses encoded in the blocks of $\text{REF}_s(\varphi)$ which are determined by assignments to the variables of $\text{REF}_s(\varphi)$.)

Algorithm 4.2: Block-width analysis and assignment extraction

Let C be the root of π and traverse the proof dag following these instructions.

1. If C is obtained by an illegal derivation step, halt and output C; if C is a leaf of π, halt and output failure. Otherwise, continue to Step (2).
2. If C was obtained from weakening a clause $C' \subseteq C$, then set $C \coloneqq C'$ and move to Step 4.
3. If C was derived from clauses $A \vee v$ and $B \vee \neg v$ by resolving over v, attempt the following

reservations according to these rules (with the condition that blocks mentioned in the set D can never be reserved).

(a) If v belongs in the root block and α contains no information about the root block, update α so that it encodes that the root block corresponds to $\bot$ and take two free blocks in layer 1 encoding the clauses x_1 and $\neg x_1$ used to derive $\bot$.

(b) If v belongs to a block on layer $1 \leq i < n$, we have two cases:

i. if α has no information about this block, update α by reserving two unreserved blocks on layer $i+1$ so that the block of v encodes the clause $\bigvee_{j=1}^{i} x_j$ and it was obtained by resolving the clauses $x_{i+1} \vee \bigvee_{j=1}^{i} x_j$ and $\neg x_{i+1} \vee \bigvee_{j=1}^{i} x_j$, to be encoded in the two reserved blocks on layer $i+1$;

ii. if the block was reserved in α but it had no children attached, then α already determined the clause C that is to be encoded in this block. Then, update α by reserving two unreserved blocks on layer $i+1$ and so that that the block of v encodes the clause C and it was obtained by resolving the clauses $x_{i+1} \vee C$ and $\neg x_{i+1} \vee C$, to be encoded in the two reserved blocks on layer $i+1$.

iii. Otherwise, do nothing.

If this reservation fails because there are not enough free blocks available, halt and output the clause C.

(c) If v belongs to a block on layer n, we distinguish two cases:

i. if the block is not mentioned in α, then try to find the first axiom in φ such that $\bigvee_{j=1}^{n} x_j$ is a weakening of it, and reserve it so it encodes $\bigvee_{j=1}^{n} x_j$ and its pointers point to this axiom;

ii. if the block was reserved in α but it was not pointing to any axiom, then there is a clause $\mathcal{A}$ associated to it by the reservation; try to find the first clause in φ such that $\mathcal{A}$ is a weakening of it and point to it in α.

iii. Otherwise, do nothing.

If this reservation fails because no axiom could be found, then halt and output the assignment to the variables $x_1, \ldots, x_n$ given by $\neg\mathcal{A}$, the negation of the clause encoded by $\mathcal{A}$.

If the reservations succeeded, look at $\alpha(v)$, which is now guaranteed to be defined. If $\alpha(v) = 1$, then move to $C := B \vee \neg v$, and otherwise move to $C := A \vee v$.

4. Clean-up the reservations in α as follows: α should only contain information about (i) blocks disabled by ρ and (ii) blocks mentioned in C or possibly the children of these according to α. Erase all other information from α.

5. Go to Step 1.

Since the algorithm is only traversing a path inside the proof dag of π, the running time of this procedure is never longer than a polynomial in the size of π. As for the correctness of the algorithm, we now show that this behaves exactly as claimed. The central claim is that, at the beginning of each iteration, when looking at clause C, the following invariant is satisfied. Here, $\mathrm{bw}(\alpha)$ stands for the number of blocks mentioned by the variables assigned by α, and $\mathrm{bw}(C)$ is the block-width of a clause C.

Claim (Invariant). The following hold at the beginning of each iteration of the algorithm, when dealing with the clause C in the traversal of π:

(i) the reservation α falsifies C;

(ii) a block $B \notin D$ is only reserved in α if it is either mentioned in C or its parent according to α is mentioned in C, and, in particular, $\mathrm{bw}(\alpha) - d \leq 3\,\mathrm{bw}(C)$;

(iii) if α encodes any information about a block B form layer i, and $B \notin D$, then α also determines that B contains exactly i literals over the variables $x_1, \ldots, x_i$ and no two literals for the same variable;

(iv) the reservation α does not falsify any axiom of $\textsc{Ref}_s(\varphi)$.

Proof sketch. The invariant is readily verified at the initial iteration, when $C = \bot$ and $\alpha = \rho$, the d-disabling restriction given as input.

Now, by straightforward structural induction, it is easy to see that assuming that the invariant holds at the beginning of an iteration and the algorithm correctly proceeds to the next iteration without halting, the invariant holds again at the beginning of the new iteration. □

Note that, if the algorithm reached a clause C that happened to be a leaf of π, then by point (i) of the invariant α would be falsifying C, which would mean falsifying an axiom of $\textsc{Ref}_s(\varphi)$, contradicting point (iv) of the very same invariant.

Thus, if π is a correct Resolution refutation, that means that the algorithm always halts before reaching a leaf, and always for one of the following two reasons.

(a) The algorithm attempted the reservation of a block at level $1 \leq i < n$, but there were no free blocks left. This means that α already reserved at least $\lfloor (s-d)/n \rfloor - 1$ blocks on that layer, and so $\mathrm{bw}(\alpha) \geq \lfloor (s-d)/n \rfloor - 1 + d$, since each layer $i < n$ contains exactly $\lfloor (s-d)/n \rfloor$ blocks. By point (ii) of the invariant we have that that $\mathrm{bw}(\alpha) - d \leq 3\,\mathrm{bw}(C)$, so putting this together we have that when outputting C we are outputting a clause of block-width at least $1/3\lfloor (s-d-n)/n \rfloor$, as desired.

(b) The reservation α had a clause $\mathcal{A}$ encoded in a block at layer n, but it failed to find an axiom of φ that $\mathcal{A}$ was a weakening of. By point (iii) of the invariant, since the block is at layer n, it encodes a width-n clause, and by Fact 4.4, $\neg\mathcal{A}$ encodes a satisfying assignment of φ. In this case the algorithm outputs this assignment, which satisfies the desired behavior.

This completes the proof of correctness of the algorithm. □

4.3 Putting it together

The following is a formal restatement of Theorem 1.1.

Theorem 4.6. *It holds that* $\mathrm{FPAP}_{\mathsf{Res}}[n^3] \in \mathbf{FP}$. *That is, there exists a deterministic polynomial-time algorithm solving the search version of the Proof Analysis problem for Resolution whenever the lower bound parameter satisfies* $s \geq n^3$.

Proof. Let $(\varphi(x_1, \ldots, x_n), \pi, 1^s)$ be an instance of the Proof Analysis Problem with $s \geq n^3$. First, check whether π is indeed a correct Resolution refutation of $\textsc{Ref}_s(\varphi)$. If not, reject. Otherwise, run the algorithm from Lemma 4.3 on π, which outputs a d-disabling restriction ρ, with $d \leq (c/2)\sqrt{s \log |\pi|}$, such that $\pi_{\restriction \rho}$ is a Resolution refutation of $\textsc{Ref}_s(\varphi)_{\restriction \rho}$ of block-width at most $c \cdot \sqrt{s \log |\pi|}$, for some fixed constant c. Then run the extraction algorithm from Lemma 4.5 on π and ρ. Since π is a correct refutation, it must be the case that the extraction algorithm from Lemma 4.5 outputs either a clause of $\pi_{\restriction \rho}$ of block-width at least

$(s-d-n)/3n$ or a satisfying assignment of φ. In the latter case, we are done. In the former case, it holds that $1/3\lfloor(s-d-n)/n\rfloor \leq c \cdot \sqrt{s \log |\pi|}$, which implies $|\pi| > 2^{\varepsilon s/n^2} \geq 2^{\varepsilon \cdot n}$ for some small enough ε and sufficiently large n. Since π is so large we can, in time polynomial in the size of the input, go over all 2^n assignments to the variables of φ and output a satisfying assignment of φ if one exists, and otherwise reject. □

If we are interested in inputs where the lower bound is quadratic instead of cubic, then we can still achieve polynomial time at the cost of randomness.

Theorem 4.7. *There exists a zero-error randomized polynomial-time algorithm solving the search version of the Proof Analysis problem for Resolution whenever the lower bound parameter satisfies $s \geq n^2$. That is,* $\mathrm{FPAP}_{\mathrm{Res}}[n^2] \in \mathbf{FZPP}$.

Proof. We carry out the proof for a fixed constant success probability p, but the argument works for any $p < 1$. The algorithm is essentially the same as before, except we now use the randomized width-reduction procedure in Lemma 4.2 at the beginning, instead of the greedy deterministic one.

Observe that the randomized algorithm in Lemma 4.2 can be used with zero-error, because once a restriction ρ is sampled, we can check if it successfully reduces the block-width to the desired bound, and run it again as many times as needed, which puts us in **FZPP**.

As for why s can now be allowed to be n^2, observe that the randomized procedure achieves better width reduction, of $O(\log |\pi|)$ whenever p is a constant. Combining this with the $(s-d-n)/3n$ bound of block-width given by Lemma 4.5, which in this case becomes $s/6n - 1/3$ because $d = s/2$, we now have $|\pi| > 2^{\Omega(s/n)}$, meaning that $s \geq n^2$ suffices to obtain an exponential lower bound. □

As discussed in introduction, the deterministic algorithm in Theorem 4.6 gives us a Levin reduction between 3SAT and the Proof Size Problem for Resolution ($\mathrm{PSP}_{\mathrm{Res}}$, see Section 2.2.3). Here the search version of 3SAT if the one that finds satisfying assignments of satisfiable formulas, while the search version of $\mathrm{PSP}_{\mathrm{Res}}$ consists in finding a Resolution refutation of the right size.

Corollary 4.8. *There is a polynomial-time Levin reduction from the search problem for* 3SAT *to the Proof Size Problem for Resolution.*

Proof. In [AM20], a 3-CNF formula φ is mapped to the formula $\textsc{Ref}_{n^2}(\varphi)$. If, instead, we map it to $\textsc{Ref}_{n^3}(\varphi)$, then this is still a many-one reduction from 3SAT and hardness of automatability still follows, but this is now a Levin reduction: given a satisfying assignment for φ, we can always come up with a short Resolution refutation of $\textsc{Ref}_{n^3}(\varphi)$ using the standard upper bound (see Section 7); and given a Resolution refutation π of $\textsc{Ref}_{n^3}(\varphi)$ of polynomial-size, we can extract a satisfying assignment for φ using Theorem 4.6. □

4.4 Assignment extraction as information efficiency

The statement that Resolution refutations of $\textsc{Ref}_s(\varphi)$ must "leak" satisfying assignments has an eminently information-theoretic flavor. This is no coincidence, as Theorem 4.6 can be interpreted in terms of the *information efficiency* of Resolution refutations, a variant of Kolmogorov complexity in the context of proof complexity.

Krajíček [Kra22] introduced the notion of information efficiency to measure the complexity of describing propositional proofs based on the well-established concept of time-bounded Kolmogorov complexity. Fix a universal Turing machine U such that $U(e, x, 1^t)$ simulates the machine with code e on input x for t steps. For every string $x \in \{0,1\}^*$, the *conditional time-bounded Kolmogorov complexity* (or *conditional Levin complexity*) of x given a string $y \in \{0,1\}^*$ is defined as

$$\mathrm{Kt}(x \mid y) \coloneqq \min\{|e| + \lceil \log t \rceil \mid U(e, y, 1^t) = x\}\,. \tag{4.9}$$

Then, for a propositional proof system S, the *information efficiency* of a formula $\varphi \in \textsc{Taut}$ is defined as

$$\mathsf{info}_S(\varphi) \coloneqq \min\{\mathrm{Kt}(\pi \mid \varphi) \mid \pi : S \vdash \varphi\}\,. \tag{4.10}$$

That is, $\mathsf{info}_S(\varphi)$ is the is the smallest Levin-complexity of a proof π of φ, given the formula φ. It is easy to see that for every $\varphi \in \textsc{Taut}$, the bounds $\log \mathsf{size}_S(\varphi) \le \mathsf{info}_S(\varphi) \le O(\mathsf{size}_S(\varphi))$ hold. The lower the information efficiency of a formula, the easier its proofs are to describe.

Because of the assignment extraction algorithm, if π is a correct Resolution refutation of $\textsc{Ref}_s(\varphi)$, then the description of π must at least include the description of a satisfying assignment for φ. This means that the information efficiency of $\textsc{Ref}_s(\varphi)$ should be (up to constant factors) precisely the same as the Levin complexity of the easiest satisfying assignment for φ. The following statement makes this precise and can be seen as an information-theoretic analogue of Theorem 4.6.

Theorem 4.9 (Assignment extraction as information efficiency)**.** *For every CNF formula φ over n variables and $\mathrm{poly}(n)$ clauses and every $s \ge n^3$, if φ is satisfiable, then*

$$\mathsf{info}_{\mathsf{Res}}(\neg\textsc{Ref}_s(\varphi)) = \Theta\left(\min\{\mathrm{Kt}(\alpha \mid \varphi) \mid \varphi(\alpha) = 1\}\right).$$

On the other hand, if φ is unsatisfiable, then $\mathsf{info}_{\mathsf{Res}}(\neg\textsc{Ref}_s(\varphi)) = \Omega(n)$.

Proof. We prove first that $\mathsf{info}_{\mathsf{Res}}(\neg\textsc{Ref}_s(\varphi)) = O\left(\min\{\mathrm{Kt}(\alpha \mid \varphi) \mid \varphi(\alpha) = 1\}\right)$. Let $\alpha^\star$ be a satisfying assignment that minimizes $\mathrm{Kt}(\alpha \mid \varphi)$ with a description $e_{\alpha^\star}$ and time $t_{\alpha^\star}$, and let $P(\varphi, \alpha, s)$ denote the Turing machine that constructs Pudlák's upper bound. Then,

$$\begin{aligned}
\mathsf{info}_{\mathsf{Res}}(\neg\textsc{Ref}_s(\varphi)) &\le \mathrm{Kt}(P(\varphi, \alpha^\star, s) \mid \neg\textsc{Ref}_s(\varphi)) && (4.11)\\
&\le |e_{\alpha^*}| + \lceil \log t_{\alpha^*} \rceil + O(1) + O(\log n) && (4.12)\\
&= O(\mathrm{Kt}(\alpha^\star \mid \varphi)) && (4.13)\\
&= O\left(\min\{\mathrm{Kt}(\alpha \mid \varphi) \mid \varphi(\alpha) = 1\}\right), && (4.14)
\end{aligned}$$

where in eq. (4.12) the constant term $O(1)$ corresponds to the code of Pudlák's algorithm P, and since φ is a CNF with $\mathrm{poly}(n)$ clauses and P runs in polynomial time, the log-term accounting for running P is $\lceil \log n^{O(1)} \rceil = O(\log n)$.

For the other direction, let $\pi^\star$ be a Resolution refutation that minimizes $\mathsf{info}_{\mathsf{Res}}(\neg\textsc{Ref}_s(\varphi))$, with description $e_{\pi^\star}$ and time $t_{\pi^\star}$. Let A denote the assignment extraction algorithm from Theorem 4.6. Then,

$$\begin{aligned}
\min\{\mathrm{Kt}(\alpha \mid \varphi) \mid \varphi(\alpha) = 1\} &\le \mathrm{Kt}(A(\pi^\star, \varphi) \mid \varphi) && (4.15)\\
&\le |e_{\pi^\star}| + \lceil \log t_{\pi^\star} \rceil + O(1) + O(\log n) && (4.16)\\
&= O(\mathrm{Kt}(\pi^\star \mid \varphi)) && (4.17)\\
&= O(\mathrm{Kt}(\pi^\star \mid \neg\textsc{Ref}_s(\varphi))) && (4.18)\\
&= O\left(\mathsf{info}_{\mathsf{Res}}\left(\neg\textsc{Ref}_s(\varphi)\right)\right). && (4.19)
\end{aligned}$$

We have that eq. (4.15) follows because $s \ge n^3$ so Theorem 4.6 guarantees the extraction algorithm A succeeds in finding a satisfying assignment, while the $O(\log n)$ term in eq. (4.16) follows again because the algorithm A runs in time $n^{O(1)}$.

The case when φ is unsatisfiable is an immediate corollary of the lower bound on $\textsc{Ref}$ formulas (Theorem 1.5): since $\textsc{Ref}_s(\varphi)$ requires Resolution size $2^{\Omega(n)}$, then writing such a proof also requires exponential time and hence $\mathrm{Kt}(\pi \mid \neg\textsc{Ref}_s(\varphi)) \ge \log 2^{\Omega(n)} = \Omega(n)$ for every refutation π, giving $\mathsf{info}_{\mathsf{Res}}(\neg\textsc{Ref}_s(\varphi)) = \Omega(n)$. (Note that if φ does have a refutation of size s, then the information efficiency of $\neg\textsc{Ref}_s(\varphi)$ is not defined, so we cannot claim a matching upper bound of $O(n)$.) □

5 $\mathrm{PAP}_{\mathsf{EF}}$ is NP-complete

In light of the extraction algorithm in Theorem 4.6, it is natural to ask whether stronger proof systems are also analyzable. As shown in Proposition 3.6, the existence of a polynomial-time proof analysis algorithm for a proof system S implies that automating S is **NP**-hard. Could this be the route towards proving that automating systems like Extended Frege is **NP**-hard?

This turns out to be unlikely. While for Resolution $\mathrm{PAP}_{\mathsf{Res}}[n^2] \in \mathbf{AC}^0$ and $\mathrm{FPAP}_{\mathsf{Res}}[n^3] \in \mathbf{FP}$, already the decision problem for Extended Frege $\mathrm{PAP}_{\mathsf{EF}}$ turns out to be **NP**-complete. This extends to every proof system S that p-simulates Extended Frege: it is **NP**-complete to decide whether a formula φ is satisfiable given an S-proof of a Resolution lower bound on φ. We dedicate Section 5.1 to prove this. Section 5.2 builds on this **NP**-hardness result to investigate whether finding polynomial-time analysis algorithms (for weaker systems where they exist) requires proving proof complexity lower bounds first.

5.1 Hardness proof

The idea of the hardness proof is to show that Extended Frege can prove Resolution lower bounds for certain formulas encoding instances of the Vertex Cover problem. The lower bound will be "agnostic" in the sense that it will not depend on whether the underlying Vertex Cover instance is satisfiable or not. This means that an algorithm that distinguishes between "true Resolution lower bounds" (those where the underlying formula is unsatisfiable) from "trivial ones" (those proven for satisfiable formulas) will be able to decide Vertex Cover and hence all of **NP**.

The formulas in question come from a convenient encoding of Vertex Cover in a way that embeds a pigeonhole principle. We define these next.

Definition 5.1 (The Vertex Cover formulas)**.** Let $G = (V, E)$ be a graph on n nodes and let k be a positive integer such that $k \leq n$. The CNF formula $\mathrm{VC}(G, k)$ has variables

$$\{v_1, \ldots, v_n\} \cup \{p_{i,j} \mid i \in [n], j \in [k]\},$$

and clauses

$$\neg v_i \vee \bigvee_{j=1}^{k} p_{i,j} \qquad \text{for all } i \in [n], \tag{VC-1}$$

$$v_i \vee v_{i'} \qquad \text{for all } (i, i') \in E, \tag{VC-2}$$

$$\neg p_{i,j} \vee \neg p_{i,j'} \qquad \text{for all } i \in [n] \text{ and } j, j' \in [k], j \neq j', \tag{VC-3}$$

$$\neg v_i \vee \neg v_{i'} \vee \neg p_{i,j} \vee \neg p_{i',j} \qquad \text{for all } j \in [k] \text{ and } i, i' \in [n], i \neq i'. \tag{VC-4}$$

The formula $\mathrm{VC}(G, k)$ is satisfiable if and only if G has a vertex cover of size at most k. The vertices in the cover are given by an assignment to the variables $v_1, \ldots, v_n$ and, to force that there are at most k vertices in the cover, the clauses on the variables $p_{i,j}$ enforce an instance of the pigeonhole principle with one pigeon for each vertex in the cover, and k holes.

Due to the embedded pigeonhole principle these formulas will be hard for Resolution. We show that such a Resolution lower bound is provable already in S^1_2. We need two main ingredients for this. First, the graphs on which we will show hardness will be the graphs obtained from the standard textbook reduction from 3SAT to Vertex Cover. Second, the proof of the Resolution lower bound in Extended Frege will come from a reduction to Haken's lower bound for the pigeonhole principle. The latter was already formalized by Cook and Pitassi in PV_1.

Theorem 5.2 (Cook and Pitassi, 1990 [CP90]). *There exist a positive $\varepsilon_0 \in \mathbb{Q}$ and $n_0 \in \mathbb{N}$ such that*

$$\mathsf{PV}_1 \vdash \forall n \forall \pi (\mathrm{Ref}_{\mathrm{Res}}(\mathrm{PHP}^n_{n-1}, \pi) \wedge n \geq n_0 \rightarrow ||\pi|| > \varepsilon_0 n).$$

Regarding the construction of the graphs, we need to make sure that S^1_2 can prove the correctness of the standard reduction from 3SAT to Vertex Cover. The proof is the standard textbook construction from 3SAT to Clique and then to Vertex Cover (as in, for example, [GJ79, §3.1.3]). We state it below but defer the proof to Appendix B.

Lemma 5.3 (3SAT $\leq_{\mathrm{p}}$ Vertex Cover in S^1_2). *There exists a* PV *function f such that S^1_2 proves the statement that for every 3-CNF formula φ with n variables and m clauses, $f(\varphi)$ outputs a graph $G_\varphi = (V, E)$ with $m \cdot n$ nodes such that the formula φ is satisfiable if and only if G_φ has a vertex cover of size $m \cdot (n-1)$.*

Now, under a suitably crafted restriction, the formula $\mathrm{VC}(G, k)$ for the particular graph $G = G_\varphi$ from Lemma 5.3 will become PHP^{n+1}_n, and S^1_2 proves Haken's lower bound, as shown by Cook and Pitassi (Theorem 5.2).

Theorem 5.4. *For every positive constant $c \in \mathbb{N}$, there exist $n_0 \in \mathbb{N}$ such that, if for every 3-CNF formula φ we write G_φ for the graph obtained in polynomial-time from φ by the reduction f in Lemma 5.3, then it holds that*

$$\mathrm{S}^1_2 \vdash \forall \varphi \forall \pi \forall n \leq \varphi \forall m \leq \varphi \left(n \geq n_0 \wedge \text{3-CNF}(\varphi, n, m) \wedge |\pi| \leq n^c \rightarrow \neg \mathrm{Ref}_{\mathrm{Res}}(\mathrm{VC}(G_\varphi, m(n-1)), \pi)\right).$$

Proof. Suppose for contradiction that π is indeed a Resolution refutation of $\mathrm{VC}(G, m(n-1))$. Consider the restriction ρ mapping $v_i \mapsto 1$ for all $i \in [mn]$. By inspecting axioms (VC-1) to (VC-4) we get that $\pi_{\restriction_\rho}$ is now a Resolution refutation of $\mathrm{PHP}^{mn}_{m(n-1)}$.

Next, restrict further as follows. Consider the variable substitution ρ' extending ρ as follows: For every $i \in [mn]$ and $j \in [m(n-1)]$, let k, k', r, r' be integers such that $i = kn + r + 1$ and $j = k'(n-1) + r' + 1$ with $0 \leq r < n$ and $0 \leq r' < n - 1$, and set $\rho' : p_{i,j} \mapsto 0$ if $k \neq k'$, and $\rho' : p_{i,j} \mapsto p_{r+1,r'+1}$ if $k = k'$. Let $\pi' := \pi_{\restriction_{\rho'}}$. It is now immediate to see that π' is a Resolution refutation of PHP^n_{n-1}. However, by Theorem 5.2, $||\pi'|| > \varepsilon_0 n$, implying $||\pi|| > \varepsilon_0 n$. This contradicts the assumption $|\pi| \leq n^c$, when $n \geq n_0$ and n_0 is chosen large enough. □

Corollary 5.5. *For every positive constant $c \in \mathbb{N}$, there exists a polynomial-time computable function t such that for every 3-CNF formula φ over a large enough number n of variables and m clauses, $t(\varphi)$ outputs an Extended Frege proof π such that*

$$\pi : \mathsf{EF} \vdash \neg \mathrm{Ref}_{n^c}\left(\mathrm{VC}(G_\varphi, m(n-1))\right).$$

Proof. The formula in Theorem 5.4 is $\forall \Pi^b_1$, so we can apply Cook's translation (Theorem 2.3) to get polynomial-size EF proofs of $[\![\neg \mathrm{Ref}_{\mathrm{Res}}(\mathrm{VC}(G_\varphi, m(n-1)), \pi)]\!]$. Extended Frege can then uniformly prove that the Ref-like formula obtained from the translation is equisatisfiable to the Ref formulas as we have defined them in Definition 2.4. We then have that EF proves $\neg \mathrm{Ref}_{n^c}\left(\mathrm{VC}(G_\varphi, m(n-1)),\right)$ in polynomial size and the proofs can be produced uniformly in polynomial time. □

The previous upper bound in Extended Frege works for every φ, regardless of its satisfiability. This is the key idea behind the final reduction.

Theorem 5.6. *For every positive constant $c \in \mathbb{N}$, the language* 3SAT *reduces to* $\mathrm{PAP}_{\mathsf{EF}}[n^c]$ *under polynomial-time many-one reductions.*

Proof. The reduction maps a 3-CNF formula φ over n variables and m clauses to the instance $(\psi, \pi, 1^s)$, where ψ is the Vertex Cover formula $\psi := \mathrm{VC}(G_\varphi, m(n-1))$, together with the Extended Frege proof π given by the map $t(\varphi)$ in Corollary 5.5, and the size parameter 1^s is 1^{n^c}.

By Corollary 5.5, the proof π is always a correct EF-proof, regardless of the satisfiability of φ. Now, if $\varphi \in$ 3SAT, then G_φ has a vertex cover of size $m(n-1)$ and hence ψ is satisfiable, so $(\psi, \pi, 1^{n^c}) \in \mathrm{PAP}_{\mathsf{EF}}$. On the other hand, if $\varphi \notin$ 3SAT, then G_φ does not have a vertex cover of size $m(n-1)$, so ψ is unsatisfiable. Since π is still a valid EF-proof, we get $(\psi, \pi, 1^{n^c}) \notin \mathrm{PAP}_{\mathsf{EF}}$. This proves that the reduction is correct. □

This yields the following formal restatement of Theorem 1.3.

Corollary 5.7. *For every propositional proof system Q that p-simulates Extended Frege and every polynomial $s(n)$, the problem* $\mathrm{PAP}_Q[s(n)]$ *is* **NP***-complete under polynomial-time many-one Levin reductions.*

Proof. Membership in **NP** is trivial, since $\mathrm{PAP}_Q \in$ **NP** for every Cook-Reckhow system Q. Hardness for $\mathrm{PAP}_{\mathsf{EF}}[s(n)]$ is given by Theorem 5.6, and since Q p-simulates EF, this means that an instance $(\psi, \pi, 1^t)$ of $\mathrm{PAP}_{\mathsf{EF}}$ with $t \geq s(n)$ can be turned into an instance $(\psi, \pi', 1^t)$, where π' is the Q-proof obtained by simulating π. □

5.2 Does analyzability require lower bounds?

Despite the **NP**-completeness results above, it remains open whether proof systems below Extended Frege are analyzable. For those proof systems that are in fact analyzable, it is natural to ask whether their analysis algorithms do inevitably rely on proof complexity lower bounds for the systems in question. After all, the only example of an analysis algorithm we have is the one for Resolution, and it heavily relies on the lower bound proof of [AM20]. It then seems reasonable to conjecture the following.

Conjecture 5.8 (Analyzability requires lower bounds)**.** *For every propositional proof system Q, if Q is analyzable, then Q is not optimal.*

The requirement that Q is not optimal is a natural formalization of lower bounds in this context: if Q is not optimal, then there exists an explicit family of tautologies that are hard for Q, but easy for some other system. Since the lower bounds for Ref-like formulas arising in the context of $\mathrm{PAP}_{\mathsf{Res}}$ are precisely Resolution lower bounds not provable in Resolution but easy for other systems, non-optimality captures the kind of lower bounds we expect are necessary.

We could also relax the conclusion of Q not being optimal by simply Q not being *p-optimal*, which means that there exists an explicit family of tautologies whose Q-proofs are either long, or short but hard to find. Note, in particular, that Q could fail to be p-optimal and still be polynomially bounded.

Conjecture 5.9 (Analyzability requires lower bounds or proofs that are hard to find)**.** *For every propositional proof system Q, if Q is analyzable, then Q is not p-optimal.*

Clearly, Conjecture 5.8 implies Conjecture 5.9. Building on the proof of **NP**-hardness of $\mathrm{PAP}_{\mathsf{EF}}$ we can prove that both conjectures are likely true —but very hard to prove.

Proposition 5.10. *The following hold:*

(i) Conjecture 5.8 is true if, and only if, **NP** $\neq$ **coNP***;*

(ii) Conjecture 5.9 is true if, and only if, **P** $\neq$ **NP***.*

Proof. We start by proving (i). For the forward direction, we assume **NP** = **coNP** and show that the conjecture fails. In this case there exists a polynomially bounded and hence optimal proof system Q. We define the following proof system $Q^\star$ based on Q. A proof π in this system is of one of the following forms:

(a) if $\pi = \langle \varphi, \tau_1, \neg\textsc{Ref}_s(\varphi), \tau_2 \rangle$, where φ is a CNF formula, τ_1 is a correct Q-proof of $\neg\varphi$, and τ_2 is a correct Q-proof of $\neg\textsc{Ref}_s(\varphi)$ for some $s \in \mathbb{N}$, then $Q^\star$ outputs $\neg\textsc{Ref}_s(\varphi)$;

(b) if $\pi = \langle \varphi, \alpha, \neg\textsc{Ref}_s(\varphi), \tau \rangle$, where φ is a CNF formula, $\varphi(\alpha) = 1$, and τ is a correct Q-proof of $\neg\textsc{Ref}_s(\varphi)$ for some $s \in \mathbb{N}$, then $Q^\star$ outputs $\neg\textsc{Ref}_s(\varphi)$;

(c) if $\pi = \langle \varphi, \tau \rangle$ and φ is not of the form $\neg\textsc{Ref}_s(\psi)$ for any s and ψ, and τ is a correct Q-proof of φ, then $Q^\star$ outputs φ;

(d) in any other case, $Q^\star$ outputs a trivial tautology $p \vee \neg p$.

We claim that $Q^\star$ is a Cook-Reckhow system that is both analyzable and optimal. First, it is sound, because Q is and every $Q^\star$-proof relies on correct Q-proofs; second, it is complete, because if φ is not a $\neg\textsc{Ref}$ formula, then one can always prove φ via case (c) above using the completeness of Q, and if $\varphi = \neg\textsc{Ref}_s(\psi)$ for some ψ, then since Q is complete, there is always a proof of $\neg\textsc{Ref}_s(\psi)$ together with either a Q-proof of $\neg\psi$ or a satisfying assignment for ψ. Finally, $Q^\star$ is clearly polynomial-time computable because Q is.

It remains to argue that $Q^\star$ is both analyzable and optimal. Indeed, $\mathrm{PAP}_{Q^\star} \in \mathbf{P}$: by definition, if π is a $Q^\star$-proof of $\neg\textsc{Ref}_s(\varphi)$ for some s, then π is of the form (a) or (b) above. If π is of the form (a), then π includes a proof of the unsatisfiability of φ, so the analyzer immediately rejects after checking the correctness of this proof. If π is of the form (b), then the analyzer can simply check that α is a correct satisfying assignment for φ and conclude that φ is satisfiable. At the same time, it is easy to see that $Q^\star$ is polynomially bounded, and hence optimal. Indeed, all non-$\neg\textsc{Ref}$ formulas have polynomial-size proofs by case (c) as Q is polynomially bounded. Similarly, in cases (a) and (b), there are always polynomial-size τ_1 and τ_2 to choose as Q is polynomially bounded, so the entire $Q^\star$ is polynomially bounded. This means that the conjecture fails for $Q^\star$.

Now, for the backwards direction, if the conjecture fails, we show that **NP** = **coNP**. For the conjecture to fail there must exist a proof system $S^\star$ that is analyzable yet optimal. Suppose $S^\star$ is analyzable by virtue of $\mathrm{PAP}_{S^\star}[n^c] \in \mathbf{P}$ for some $c > 0$. We then construct a non-deterministic polynomial-time procedure for the **coNP**-complete set 3Unsat. On an input 3-CNF formula φ with n variables and m clauses, non-deterministically guess an $S^\star$-proof of the formula $\neg\textsc{Ref}_{n^c}(\mathrm{VC}(G_\varphi, m(n-1)))$. By Corollary 5.5 polynomial-size such proofs exist in Extended Frege. Since $S^\star$ is optimal we have $S^\star \geq \mathsf{EF}$ and some polynomial-size proof is available in $S^\star$ too. Now, since $S^\star$ is analyzable, we can decide deterministically in polynomial-time whether φ is satisfiable or not by analyzing this proof.

We now prove (ii). If **P** = **NP**, then there exists a polynomially bounded and p-optimal proof system $Q^\star$. Since **P** = **NP** the system $Q^\star$ is trivially analyzable, but it is p-optimal, so the conjecture fails for $Q^\star$.

For the other direction, suppose **P** ≠ **NP** but Conjecture 5.9 fails. That means there exists a system that is analyzable yet p-optimal. Since S is p-optimal, it follows that S p-simulates EF, and by Corollary 5.7 we have that $\mathrm{PAP}_S[n^c]$ is **NP**-complete for every $c \in \mathbb{N}$. Since **P** ≠ **NP** we get that $\mathrm{PAP}_S[n^c] \notin \mathbf{P}$ for any $c > 0$, a contradiction with the fact that S is analyzable. □

6 The Atserias–Müller lower bound in PV_1

In this section we undertake the task of formalizing the Atserias–Müller lower bound in bounded arithmetic to prove Theorem 1.4. There are at least two possible approaches here, at least at a conceptual level. The first one is to formalize in PV_1 the correctness of the extraction algorithm from Section 4, and use this to derive the lower bound along the lines of the argument in Theorem 4.6, carefully verifying that we never exceed the reasoning power of the theory.

The alternative approach, which we choose to take, is to forget about the extraction algorithm at first glance and attempt a direct formalization of the theorem seeing it purely as a proof complexity lower bound. Then, from the form of the statement, witnessing theorems will recover the algorithm together with its

proof of correctness. We opt for this second route, since it is in fact the original hint for why the algorithm should exist in the first place, and because restating the proof in this format can be insightful in itself.

Of course, the proof we choose to formalize deviates from the original one in a few crucial points, although the overall structure is preserved. The main difference is the deterministic restriction replacing the random restriction argument in [AM20]. This deterministic procedure corresponds exactly to the greedy algorithm behind the width-reduction technique in Lemma 4.3.

We start by stating the lower bound in the language of PV_1, which posits that a correct Resolution refutation π of $\textsc{Ref}_s(\varphi)$ must have size at least $2^{\varepsilon s/n^2}$ whenever $\varphi \notin \mathrm{SAT}$, for some fixed positive ε. Since we do not have exponentiation in PV_1, we scale down the bound and state is as $||\pi|| > \varepsilon s/n^2$, and the first-order formula becomes

$$\mathrm{AM}_{\varepsilon,n_0} \coloneqq \forall\varphi\forall\pi\forall n\forall s\Big(\Big(n \geq n_0 \wedge \mathrm{CNF}(\varphi,n) \wedge \mathrm{Ref}_{\mathrm{Res}}(\textsc{Ref}_s(\varphi),\pi) \wedge \forall\alpha \leq \varphi\big(\neg\,\mathrm{Sat}(\varphi,\alpha)\big)\Big) \rightarrow ||\pi|| > \varepsilon s/n^2\Big). \tag{$\mathrm{AM}_{\varepsilon,n_0}$}$$

Here the different predicates and function symbols all have the obvious intended meaning and it is readily verified that they are all computable in polynomial time and there exist function symbols for them in the language of PV.

The formula itself is $\forall\Sigma^b_1$. This becomes clear when rewriting it in prenex form, which gives us the statement

$$\forall\varphi\forall n\forall s\forall\pi\exists\alpha \leq \varphi\Big(n < n_0 \vee \neg\,\mathrm{CNF}(\varphi,n) \vee \neg\,\mathrm{Ref}_{\mathrm{Res}}(\textsc{Ref}_s(\varphi),\pi) \vee \mathrm{Sat}(\varphi,\alpha) \vee ||\pi|| > \varepsilon s/n^2\Big). \tag{6.1}$$

If we were to apply Buss's witnessing theorem here, we would get a polynomial-time function witnessing the existential quantifier on α, which corresponds precisely to the extraction algorithm.

We carry out the formalization in three subsections. Section 6.1 proves the restriction argument, Section 6.2 carries out the block-width lower bound and Section 6.3 puts these together. We also recall that $\mathsf{S}^1_2(\mathsf{PV})$, which we denote simply as S^1_2, is $\forall\Sigma^b_1$-conservative over PV_1, so we will often carry out arguments in S^1_2 instead of PV_1 and use the induction principles available there without further comment.

6.1 The restriction argument

We want to prove the following $\forall\Sigma^b_1$ formula, stating the restriction argument:

$$\begin{aligned}\text{AM-Restriction}_c \coloneqq \forall\varphi\forall n\forall s\forall\pi\Big(&\mathrm{CNF}(\varphi,n) \wedge \mathrm{Ref}_{\mathrm{Res}}(\textsc{Ref}_s(\varphi),\pi) \\ &\rightarrow \exists\rho \leq \pi\Big(\mathrm{bw}(\pi_{\restriction\rho}) \leq c\cdot\Big\lceil\sqrt{s\log|\pi|}\Big\rceil \\ &\qquad \wedge \textsc{Ref}_s(\varphi)_{\restriction\rho} \neq \bot \\ &\qquad \wedge \mathrm{Disabling}(\rho,\varphi,n,s) \leq c/2\cdot\Big\lceil\sqrt{s\log|\pi|}\Big)\Big\rceil\Big)\Big).\end{aligned} \tag{AM-Restriction$_c$}$$

The formula AM-Restriction$_c$ states that for every CNF formula φ over n variables and every correct Resolution refutation π of the $\textsc{Ref}_s(\varphi)$ formula, there exists a d-disabling restriction $\rho \in \{0,1,*\}^N$ to the N variables of $\textsc{Ref}_s(\varphi)$ such that $\mathrm{bw}(\pi_{\restriction\rho}) \leq c\cdot\lceil\sqrt{s\log|\pi|}\rceil$ and $d \leq c/2\cdot\lceil\sqrt{s\log|\pi|}\rceil$.

Throughout this section we freely use rationals and reals and notation like $\sqrt{\cdot}$, always assuming that some suitable rational approximation is used under the hood. This is standard for formalizations in these

theories (see, for example, the standard style of the formalizations by Jeřábek [Jeř05]).

It is easy to see that there exists a PV function W that given π outputs the set of clauses in π of block-width strictly larger than $c \cdot \left\lceil \sqrt{s \log |\pi|} \right\rceil$. Similarly, there is a PV function $W_{\restriction \rho}$ that given π and ρ defines the set of clauses in $\pi_{\restriction \rho}$ of block-width strictly larger than $c \cdot \left\lceil \sqrt{s \log |\pi|} \right\rceil$, and the basic properties of these functions can be proven in S^1_2.

For a given restriction ρ that sets values of variables in ℓ different blocks, we will denote by $\tau := ((b_1, X_1), \ldots, (b_\ell, X_\ell))$ an ordering $b_1, \ldots, b_\ell \in [s]$ of the blocks that are mentioned by ρ, and, for each of them, an ordering X_i of all the variables in block b_i. We call τ a *trace* for ρ and for every $k \in [\ell]$ and S a set of variables of $\mathrm{Ref}_s(\varphi)$, we denote by $\rho_{k,S}$ the subrestriction of ρ defined as

$$\rho_{k,S}(x) := \begin{cases} \rho(x) & \text{if } x \text{ belongs in one of the blocks } b_1, \ldots, b_k \in \tau \text{ or if } x \in S \\ * & \text{otherwise} \end{cases} \tag{6.2}$$

and $\rho_{0,S} := *^N$. We will denote by ρ_k the subrestriction $\rho_{k,\emptyset}$.

Definition 6.1 (Most-killing property)**.** We will say that a restriction ρ enjoys the *most-killing property* with respect to a trace $\tau = ((b_1, X_1), \ldots, (b_\ell, X_\ell))$ if it holds that for every $k \in [\ell]$, the block $i \in [s] \setminus \{b_1, \ldots, b_{k-1}\}$ that is the most frequent block mentioned in $W_{\restriction \rho_{k-1}}$ happens to be precisely b_k and ρ_k satisfies the following conditions.

1. If e_i appears positively in at least $1/3$ of all the clauses in $W_{\restriction \rho_{k-1}}$ that mention block i, then $\rho_k(e_i) = 1$ and $\rho_k(x) = *$ for every other variable $x \neq e_i$ in block i.
2. If e_i does *not* appear positively in at least $1/3$ of all the clauses in $W_{\restriction \rho_{k-1}}$ that mention block i, then $\rho_k(e_i) = 0$ and for every other variable in block i, ρ_k assigns values to them respecting the ordering in $X_i = (e_i, x_{i,1}, x_{i,2}, \ldots)$ and with the following priority. For $j = 1, \ldots, |X_i| - 1$,
   (a) if $x_{i,j}$ appears positively more often than negatively in $W_{\restriction \rho_{k-1,\{e_i, x_{i,1}, \ldots, x_{i,j-1}\}}}$, then $\rho_k(x_{i,j}) = 1$;
   (b) if $x_{i,j}$ appears negatively more often than positively in $W_{\restriction \rho_{k-1,\{e_i, x_{i,1}, \ldots, x_{i,j-1}\}}}$, then $\rho_k(x_{i,j}) = 0$.

Note that verifying that ρ enjoys the most-killing property with respect to τ is possible in polynomial-time and hence there is a PV function that performs the check and S^1_2 can prove this.

We now prove the following claim by induction.

Lemma 6.2. *Let $c \in \mathbb{N}$. The following is provable in S^1_2. For every CNF formula φ over n variables and π a correct Resolution refutation of $\mathrm{Ref}_s(\varphi)$, let $w := c \cdot \left\lceil \sqrt{s \log |\pi|} \right\rceil$ and let W denote the set of clauses in π of block-width larger than w. Then, for every ℓ, the formula $\Psi(\varphi, \pi, s, \ell)$ defined as follows holds: if $\ell \leq w/2$, then there exists a restriction $\rho \in \{0, 1, *\}^N$, where N is the number of variables of the $\mathrm{Ref}_s(\varphi)$ formula, a trace $\tau = ((b_1, X_1), \ldots, (b_\ell, X_\ell))$ and $d \leq \ell$ such that:*

(i) ρ is d-disabling;

(ii) ρ has the most-killing property according to τ;

(iii) $(3s)^\ell \cdot |W_{\restriction \rho}| \leq |W| \cdot (3s - w + \ell)^\ell$.

Proof. Observe that the formula Ψ is Σ^b_1. First, the existential quantifiers on ρ, τ and d are bounded by π and φ. The properties (i) and (ii) are checkable in polynomial time and hence there are PV relations for

them and S_2^1 can prove their basic properties. To show that item (iii) can be properly expressed by a PV formula, note that because π is a correct Resolution refutation of $\text{REF}_s(\varphi)$, it must hold that $s \leq |\pi|$ and

$$w = c\left\lceil \sqrt{s \log |\pi|} \right\rceil \leq c\left\lceil \sqrt{s||\pi||} \right\rceil \leq c\left\lceil \sqrt{|\pi|||\pi||} \right\rceil \leq c|\pi|\,, \tag{6.3}$$

where the last equality holds because $||\pi|| \leq |\pi|$. As a consequence, $w \in \text{Log}$, which means that for every $\ell \leq w/2$, $\ell \in \text{Log}$ and thus s^ℓ exists because $s^\ell = 2^{|s|\ell} = s\#2^\ell$.

We show that $S_2^1 \vdash \forall \ell \Psi(\varphi, \pi, s, \ell)$. We proceed by induction on the parameter ℓ, the length of the trace τ. Note that this corresponds to Length Induction over Σ_1^b formulas and is hence available in S_2^1.

If $\ell = 0$, then we can take the empty restriction $\rho := *^N$, which disables $d = 0$ variables and the trace τ to be the empty sequence, which trivially satisfy the conditions.

We assume now that $\Psi(\varphi, \pi, s, \ell)$ holds and we prove $\Psi(\varphi, \pi, s, \ell + 1)$. If $\ell + 1 > w/2$, then we are done, so assume $\ell + 1 \leq w/2$. By induction hypothesis, there exist ρ, d and τ satisfying the desired properties for ℓ. We extend ρ into ρ' by following the conditions of the most-killing property. This amounts to applying one iteration of Algorithm 4.1. There is a PV function for this, as well as for extracting the most frequent block in the refutation and S_2^1 can prove the basic properties of these functions. The trace is then extended by the most frequent block $b_{\ell+1}$ mentioned in $W_{\restriction\rho}$ and the ordering of the variables can be any fixed ordering of the variables used in the construction of ρ' by Algorithm 4.1.

If ρ was d-disabling for $d \leq \ell$, then we have that ρ' is at most $(d+1)$-disabling, since one iteration of Algorithm 4.1 disables at most one additional block, and we have $d + 1 \leq \ell + 1$. By construction, it is easy to see that ρ' has the most-killing property with respect to τ'. It is only left to verify that

$$(3s)^{\ell+1} \cdot |W_{\restriction\rho'}| \leq |W| \cdot (3s - w + \ell + 1)^{\ell+1}\,. \tag{6.4}$$

The set W contains all the clauses in π of block-width at least w. The restriction ρ' sets the variables of $\ell + 1$ blocks, meaning that every clause in W not trivialized by ρ' has block-width at least $w - (\ell + 1)$, and these clauses are precisely the ones in $W_{\restriction\rho'}$ by definition.

Furthermore, the restriction ρ' disables at most $\ell + 1$ blocks, so an averaging argument implies that there is a non-restricted block mentioned in at least $|W_{\restriction\rho}|(w - (\ell + 1))/(s - (\ell + 1))$ clauses. In particular, the most frequent non-restricted block must be mentioned at least that often. Since $|W| \leq |\pi|$ and $s \leq |\pi|$, we have that $|W|$ and s are in Log and therefore we can use exact counting in S_2^1 to carry out this averaging argument (see Section 2.3.2 for details).

By the way we constructed ρ' following Algorithm 4.1 we are guaranteed to kill at least $1/3$ of all the clauses mentioning the most frequent unrestricted block $b_{\ell+1}$. Indeed, by inspecting the conditions of the most-killing property (Definition 6.1) if we enable block $b_{\ell+1}$ that is because it appeared in at least $1/3$ of all the clauses mentioning $b_{\ell+1}$; and otherwise we are guaranteed to restrict at least $1/2$ of the remaining $2/3$ fraction of the clauses mentioning $b_{\ell+1}$, which amounts to a $1/3$ fraction.

Together, this means that

$$|W_{\restriction\rho'}| \leq |W_{\restriction\rho}| \cdot \left(1 - \frac{w - (\ell + 1)}{3s}\right), \tag{6.5}$$

which is the same as

$$3s \cdot |W_{\restriction\rho'}| \leq |W_{\restriction\rho}| \cdot (3s - w + \ell + 1)\,. \tag{6.6}$$

As ρ' extends the restriction ρ, it is clear that $|W_{\restriction\rho'}| \leq |W_{\restriction\rho}|$, and by induction hypothesis we know that

$$(3s)^\ell \cdot |W_{\restriction\rho}| \leq |W| \cdot (3s - w + \ell)^\ell\,, \tag{6.7}$$

meaning that

$$3s \cdot |W_{\restriction \rho'}| \leq |W_{\restriction \rho}| \cdot (3s - w + \ell + 1) \tag{6.8}$$

$$\leq \frac{1}{(3s)^{\ell}} |W| \cdot (3s - w + \ell)^{\ell} \cdot (3s - w + \ell + 1) \tag{6.9}$$

$$\leq \frac{1}{(3s)^{\ell}} |W| \cdot (3s - w + \ell + 1)^{\ell+1}, \tag{6.10}$$

from which the desired $(3s)^{\ell+1} \cdot |W_{\restriction \rho'}| \leq |W| \cdot (3s - w + \ell + 1)^{\ell+1}$ follows. □

With our claim in hand, we are ready to show that PV_1 proves the desired restriction argument.

Lemma 6.3 (Formalized deterministic width-reduction in PV_1)**.** *It holds that* $\mathrm{PV}_1 \vdash \text{AM-Restriction}_4$.

Proof. Let φ be a CNF formula over n variables and π a correct Resolution refutation of $\textsc{Ref}_s(\varphi)$ for some s. Apply Lemma 6.2 in S^1_2 for $\ell := w/2$, where $w := 4 \cdot \left\lceil \sqrt{s \log |\pi|} \right\rceil$. This gives us a restriction ρ that is d-disabling for $d \leq \ell$ and $(3s)^{\ell} \cdot |W_{\restriction \rho}| \leq |W| \cdot (3s - w + \ell)^{\ell}$. We want to show that $\mathrm{bw}(\pi_{\restriction \rho}) \leq w$, that $\textsc{Ref}_s(\varphi)_{\restriction \rho} \neq \bot$ and that ρ is d-disabling, for $d \leq w/2$.

That ρ is d-disabling is guaranteed by the lemma, and because it is d-disabling, the restriction only sets the values of enabling variables and, whenever a block is enabled, it does not restrict any other variables in the block. As a consequence, no axiom of $\textsc{Ref}_s(\varphi)$ is ever falsified by ρ.

It is only left to verify that the block-width of $\pi_{\restriction \rho}$ is at most w, as desired. Lemma 6.2 guarantees that $|W_{\restriction \rho}| \leq |W| \cdot (1 - (w - \ell)/3s)^{\ell}$, and substituting our choice of ℓ we have that

$$|W_{\restriction \rho}| \leq |W| \cdot \left(1 - \frac{w/2}{3s}\right)^{w/2} \leq |W| \cdot 2^{-\frac{w^2}{12s}}, \tag{6.11}$$

where the last inequality is provable in S^1_2, as shown in Lemma A.2 of Appendix A.

We want $|W| \cdot 2^{-w^2/12s} < 1$, so

$$|W| \cdot 2^{-\frac{w^2}{12s}} < 1 \Leftrightarrow \log |W| < \frac{w^2}{12s} \tag{6.12}$$

$$\Leftrightarrow \sqrt{12s \log |W|} < w, \tag{6.13}$$

and this last inequality is verified for our choice of w.

It follows that $W_{\restriction \rho} = \emptyset$ and hence $\mathrm{bw}(\pi_{\restriction \rho}) \leq w$, as desired. Since S^1_2 is $\forall\Sigma^b_1$-conservative over PV_1, we have the sentence in PV_1. □

6.2 The block-width lower bound

The crucial part of the formalization, and the place where we most clearly see how satisfying assignments can be extracted from Resolution refutations, is in the width lower bound. The restriction argument we just covered reduces the block-width of a refutation regardless of whether the underlying formula is satisfiable or not; it is the width lower bound that only holds when the formula is unsatisfiable.

Let us first state the formula we want to prove, which we encode in the following $\forall\Sigma^b_1$ statement:

$$\text{AM-WLB} := \forall\varphi\forall n\forall s\forall\pi\Big(\mathrm{CNF}(\varphi, n) \wedge \mathrm{Ref}_{\mathrm{Res}}(\textsc{Ref}_s(\varphi), \pi) \wedge \forall\alpha \leq \varphi.\neg\mathrm{SAT}(\varphi, \alpha) \\ \rightarrow \exists C \leq \pi\Big(C \in \pi \wedge \mathrm{bw}(C) \geq 1/3(\lfloor s/n \rfloor - 1)\Big)\Big). \tag{AM-WLB}$$

When writing the formula in prenex form, pulling out the universal quantifier on α and turning it into an existential one, it becomes clear how the statement is $\forall\Sigma^b_1$ and the extraction of a satisfying assignment can be performed by witnessing the first existential quantifier:

$$\text{AM-WLB} \equiv \forall\varphi\forall n\forall s\forall\pi\exists\alpha \leq \varphi\exists C \leq \pi\Big(\text{CNF}(\varphi, n) \wedge \text{Ref}_{\text{Res}}(\textsc{Ref}_s(\varphi), \pi) \\ \rightarrow \text{Sat}(\varphi, \alpha) \vee \Big(C \in \pi \wedge \text{bw}(C) \geq 1/3(\lfloor s/n \rfloor - 1)\Big)\Big). \tag{6.14}$$

The crucial fact that makes the block-width lower bound go through, and which fundamentally distinguishes the SAT and Unsat cases is that, if $\varphi \in$ Unsat, then every width-n clause over the variables of φ can be weakened from some clause of φ, while a width-n clause that is *not* the weakening of any axiom will necessarily encode a satisfying assignment. This was earlier stated as Fact 4.4. We now reprove the statement in S^1_2.

Lemma 6.4 (Fact 4.4 in S^1_2)**.** *Let φ be a Boolean formula in CNF over n variables. If C is a width-n clause over the variables of φ that is not the weakening of any clause of φ, then $\neg C$ encodes a satisfying assignment for φ.*

Proof. By Δ^b_1-induction on the number n of variables $x_1, \ldots, x_n$ of φ, which is available in S^1_2.

If $n = 1$, then the only two width-1 clauses are x_1 and $\neg x_1$. The three possible CNF formulas are x_1, $\neg x_1$ and $x_1 \wedge \neg x_1$, and it is easy to check that the statement holds.

Suppose the statement holds for formulas with up to n variables, and let φ be a formula over $n + 1$ variables and let C be a clause of width $n + 1$. Assume $x_{n+1} \in C$, and restrict $x_{n+1} \mapsto 0$. We obtain a new formula φ_0 and a clause C_0. For every clause D in φ we have the following cases:

(a) the variable x_{n+1} appeared in D negatively, and hence D is satisfied by setting $x_{n+1} \mapsto 0$;

(b) the variable x_{n+1} appeared in D positively, or the variable did not appear at all; in both cases we have that this is a clause over variables $x_1, \ldots, x_n$ and, by assumption, C and hence also C_0 is not a weakening of it. By induction hypothesis, $\neg C_0$ and then also $\neg C$ encode a satisfying assignment to D.

Hence, every clause in φ is satisfied by $\neg C$. If $\neg x_{n+1} \in D$ the the same argument goes through by restricting $x_{n+1} \mapsto 1$. This completes the proof. □

To prove the rest of the width lower bound, we show the following lemma first, which essentially corresponds to the invariant proved within Lemma 4.5 to derive the correctness of the assignment extraction algorithm. We restate the invariant here as a property of a path in the Resolution refutation.

Definition 6.5 (Reservation invariant)**.** Let π be a Resolution refutation of $\textsc{Ref}_s(\varphi)$ for some CNF formula $\varphi(x_1, \ldots, x_n)$, let N be the number of variables of $\textsc{Ref}_s(\varphi)$, $d \leq \text{depth}(\pi)$, and let $C = (C_1, \ldots, C_d)$ be a length-d path in π starting from $C_1 = \bot$. Let the s blocks of $\textsc{Ref}_s(\varphi)$ be arranged in a layered manner, so that there are n layers, each containing $\lfloor s/n \rfloor$ blocks, with the remainder blocks left from the flooring operations collected all in the last layer, plus one additional layer on top with a single block corresponding to the root. We say that C satisfies the *reservation invariant* with respect to a sequence $A = (\alpha_1, \ldots, \alpha_d)$ of restrictions $\alpha_i \in \{0, 1, *\}^N$ if, for every $i \in [d]$,

(i) the restriction α_i falsifies C_i;

(ii) a block is only mentioned in α_i if it is either mentioned in C_i or its parent according to α_i is mentioned in C_i, and, in particular, $\text{bw}(\alpha_i) \leq 3\,\text{bw}(C_i)$;

(iii) if α_i restricts some variable in block B from layer ℓ, then it does so in the following way: B must contain exactly ℓ literals over the variables $x_1, \ldots, x_\ell$ and no two literals for the same variable;

(iv) the restriction α_i does not falsify any clause of $\textsc{Ref}_s(\varphi)$;

(v) the path determined by C corresponds to A in the following way: if C_i was derived from C_{i+1} by resolving over variable v, then $\alpha_{i+1}(v)$ is defined and v appears in C_{i+1} with polarity $1 - \alpha_{i+1}(v)$.

The lemma now states that a path exists that satisfies the invariant or, alternatively, correctly leads to a wide clause or a satisfying assignment.

Lemma 6.6. *The following is provable in S^1_2. For every CNF formula φ over n variables and π a correct Resolution refutation of $\textsc{Ref}_s(\varphi)$, for every d, the formula $\Psi(\varphi, \pi, s, d)$ defined as follows holds: if $1 \leq d \leq \mathrm{depth}(\pi)$, then there exists a sequence $C = (C_1, \ldots, C_d)$ of clauses of π and a sequence $A = (\alpha_1, \ldots, \alpha_d)$ of restrictions such that at least one of the following conditions holds:*

(a) there is $C_i \in C$ that satisfies $\mathrm{bw}(C_i) \geq 1/3(\lfloor s/n \rfloor - 1)$;

(b) there is C_i in and its corresponding restriction α_i that contain a satisfying assignment to φ;

(c) the sequence C is a path in π starting at $C_1 = \bot$ and the reservation invariant holds for C with respect to A.

Proof. The formula Ψ is Σ^b_1 because the existential quantifiers on C and A are bounded by π and items (a)-(c) are all checkable by PV functions and S^1_2 proves their basic properties. We can then proceed by induction on the parameter d, which corresponds to Length Induction over a Σ^b_1 formula, available in S^1_2.

For the base case, $d = 1$ and we can pick $C_1 = \bot$ and $\alpha_1 = *^N$. It is immediate to verify that the reservation invariant is satisfied, so item (c) holds.

Now, suppose the statement holds for d, and we prove it for $d+1$. If $d+1 > \mathrm{depth}(\pi)$, then we are done. Otherwise, by induction hypothesis there exists a sequence $C = (C_1, \ldots, C_d)$ and a series of restrictions $A = (\alpha_1, \ldots, \alpha_d)$. If (a) or (b) hold for them, then we simply extend C with $C_{d+1} \coloneqq \bot$ and $\alpha_{d+1} \coloneqq *^N$ and either (a) or (b) will still hold.

If we are in case (c) but (a) and (b) failed, then α_d does not falsify any clause of $\textsc{Ref}_s(\varphi)$ (as per item (iv) of the reservation invariant), and hence the clause was derived from some previous clause in π, either by weakening or by a resolution step. Here we simply move to one of the children in π and extend α_d into α_{d+1} following the reservation strategy in Algorithm 4.2. More precisely, we run one iteration of Algorithm 4.2 starting the traversal from C_d and taking α_d as the restriction kept in memory by the algorithm (this corresponds to Steps 2-4 of Algorithm 4.2). Crucially, there is a PV function that carries out this computation.

Let us analyze the outcome of this procedure. If the iteration of Algorithm 4.2 succeeds in extending the reservation from α_d to α_{d+1}, then by the way α_{d+1} is constructed from α_d and by the fact that the induction hypothesis guaranteed that α_d satisfies the invariant, we have that α_{d+1} will immediately satisfy the reservation invariant too. In this case, item (c) holds and we are done.

We argue that the reservation process cannot fail. If it did, it must be that the process failed at Step 3b or 3c of Algorithm 4.2.

- If the process failed in Step 3b, the algorithm attempted the reservation of a block at layer $1 \leq i < n$, but there were no free blocks left. This means that α_d already reserved at least $\lfloor s/n \rfloor - 1$ blocks on that layer, and so $\mathrm{bw}(\alpha_d) \geq \lfloor s/n \rfloor - 1$, since each layer $i < n$ contains exactly $\lfloor s/n \rfloor$ blocks. By point (ii) of the reservation invariant we have that $\mathrm{bw}(\alpha_d) \leq 3\,\mathrm{bw}(C_d)$, so putting this together we have that C_d has block-width at least $1/3(\lfloor s/n \rfloor - 1)$, contradicting that we were not in case (a).

- If the process failed in Step 3c, this implies the reservation α_d had a clause $\mathcal{A}$ encoded in a block at layer n, but it failed to find a clause of φ that $\mathcal{A}$ was a weakening of. By point (iii) of the invariant, since the block is at layer n, $\mathcal{A}$ is a width-n clause, and by Lemma 6.4 in S^1_2, $\neg\mathcal{A}$ encodes a satisfying assignment of φ, contradicting that we were not in case (b).

This completes the induction. □

We are now ready to prove AM-WLB.

Lemma 6.7 (Formalized block-width lower bound in PV_1)**.** *It holds that* $\mathrm{PV}_1 \vdash$ AM-WLB.

Proof. Working in S^1_2, let π be a correct Resolution refutation of $\textsc{Ref}_s(\varphi)$ for a CNF formula φ over n variables. We can apply Lemma 6.6 for $d = \mathrm{depth}(\pi)$ to obtain a sequence of clauses $C = (C_1, \ldots, C_d)$ of π and a sequence of restrictions $A = (\alpha_1, \ldots, \alpha_d)$, satisfying one of conditions (a)-(c) in the statement. We claim that condition (c) cannot occur. Suppose it did. If C was really a path in π starting at the root $C_1 = \bot$, since π is a correct Resolution refutation this means the underlying graph of π is a DAG and therefore the clause C_d at depth d must be a leaf. It is not hard to see that S^1_2 can prove this. This implies that C_d is a clause of $\textsc{Ref}_s(\varphi)$. Now, items (i) and (iv) of the reservation invariant in Definition 6.5 contradict each other: the restriction α_d must falsify C_d, which is a clause of $\textsc{Ref}_s(\varphi)$; but item (iv) promised that α_d would not falsify any clause of the $\textsc{Ref}_s(\varphi)$.

The only viable option then is that we are in case (a) or (b). In case (a), some $C_i \in \pi$ that appears in the sequence $(C_1, \ldots, C_d)$ has block-width at least $1/3(\lfloor s/n \rfloor - 1)$. In case (b), we immediately get a satisfying assignment for φ.

Since AM-WLB is a $\forall\Sigma^b_1$ sentence and S^1_2 is $\forall\Sigma^b_1$-conservative over PV_1, we get that $\mathrm{PV}_1 \vdash$ AM-WLB. □

6.3 Formalization of the final lower bound statement

Theorem 6.8. *There exist a positive* $\varepsilon \in \mathbb{Q}$ *and* $n_0 \in \mathbb{N}$ *such that* $\mathrm{PV}_1 \vdash \mathrm{AM}_{\varepsilon,n_0}$.

Proof. Let $\varepsilon \in \mathbb{Q}$ and $n_0 \in \mathbb{N}$ be universal constants that can be computed from the rest of the argument. Working in PV_1, let π be a Resolution refutation of $\textsc{Ref}_s(\varphi)$ for a CNF formula φ over n variables, with $n \geq n_0$. We apply Lemma 6.3 to conclude that $\mathrm{PV}_1 \vdash \text{AM-Restriction}_4$, from which we get in PV_1 that there is some restriction ρ such that $\pi_{\restriction\rho}$ is a Resolution refutation of $\textsc{Ref}_s(\varphi)_{\restriction\rho}$ and the block-width of $\pi_{\restriction\rho}$ is at most $4 \cdot \sqrt{s \log |\pi|}$. Furthermore, ρ is d-disabling for some $d \leq 2\lceil\sqrt{s \log |\pi|}\rceil$.

Before proceeding, extend ρ into a restriction ρ' that further enables all blocks not touched by ρ and sets the values of pointers that are pointing at disabled blocks. In this way, $\textsc{Ref}_s(\varphi)_{\restriction\rho'}$ becomes exactly $\textsc{Ref}_{s-d}(\varphi)$, as ρ was d-disabling, and $\pi_{\restriction\rho'}$ is a refutation of $\textsc{Ref}_{s-d}(\varphi)$.

Assume now that φ is unsatisfiable, or else we are already done. By Lemma 6.7 we have that $\mathrm{PV}_1 \vdash$ AM-WLB. Since we assume φ is unsatisfiable, applying this block-width lower bound to $\pi_{\restriction\rho'}$ we conclude that there exists a clause in $\pi_{\restriction\rho'}$ of block-width at least $1/3(\lfloor (s-d)/n \rfloor - 1)$. It then holds that

$$1/3(\lfloor (s-d)/n \rfloor - 1) \leq 4 \cdot \left\lceil \sqrt{s \log |\pi|} \right\rceil, \tag{6.15}$$

and it is not hard to see that PV_1 can derive from this inequality that $||\pi|| > \varepsilon s/n^2$ for some small enough ε and large enough n. □

7 Pudlák's upper bound in Resolution

The fact that the $\textsc{Ref}_s(\varphi)$ formulas admit short Resolution refutations whenever φ is satisfiable is originally due to Pudlák [Pud03, Theorem 4.1]. However, technically speaking, our definition of Ref is in the *relativized*

form as used by Atserias and Müller, who crucially showed that the upper bound still works even in the presence of the disabling variables [AM20, Lemma 11], and the proof goes through for the encoding with pointer variables in binary too [dRGN+21, Lemma 2.1.i].

Our goal is to show that the upper bound construction can be proven correct in Resolution itself. The main technicality to overcome is that we want to show that from a satisfying assignment α to a CNF formula φ, one can construct not only a Resolution refutation of $\textsc{Ref}(\varphi)$, but that this refutation can itself be encoded as a satisfying assignment to the formula $\textsc{Ref}(\textsc{Ref}(\varphi))$, and that the correctness of this can be certified in Resolution.

Let $P(\varphi, \alpha, s)$ denote the circuit that constructs a refutation of $\textsc{Ref}_s(\varphi)$ given the satisfying assignment α. Intuitively, we want to derive a propositional formula that states

$$(\textsc{Sat}(\varphi, \alpha) \land \pi = P(\varphi, \alpha, s)) \to \textsc{Ref}(\textsc{Ref}_s(\varphi), \pi)\,. \tag{7.1}$$

Our goal is to refute the negation of this formula in Resolution, which presents two obstacles. First, when negating the formula to obtain a contradiction to be refuted in Resolution, we have

$$\textsc{Sat}(\varphi, \alpha) \land \pi = P(\varphi, \alpha, s) \land \neg\textsc{Ref}(\textsc{Ref}_s(\varphi), \pi)\,. \tag{7.2}$$

Since $\textsc{Ref}$ formulas are in CNF, its negation is a DNF, so this is still not in a format amenable to Resolution. The second issue is that the circuit P needs to be presented in a simple way that can also be handled by Resolution.

In Section 7.1 we introduce pseudo-negations, a construction using extension variables to simulate the negation of $\textsc{Ref}$ formulas. In Section 7.2 we describe the construction in natural language, in a style that streamlines the previous existing proofs. In Section 7.3 we describe the circuit P, which turns out to be a very low depth circuit that Resolution will be able to reason about. Finally, Section 7.4 puts these together to derive the correctness of the construction in Resolution.

7.1 Pseudo-negations in Resolution

In general, we cannot negate a CNF formula and rewrite in CNF without blowing up the size of the formula, but we can simulate the negation using extension variables.

Definition 7.1 (Pseudo-negations)**.** Let $\varphi = C_1 \land \cdots \land C_m$ be a CNF formula over n variables $x_1, \ldots, x_n$ and m clauses, each of width at most k, and let $C_i = \ell_{i,1} \lor \cdots \lor \ell_{i,k}$ denote the literals in each clause. We define the *pseudo-negation of* φ, denoted $\sim\varphi$, to be the following CNF formula over variables $x_1, \ldots, x_n$ and new variables $\mathsf{mistake}_1, \ldots, \mathsf{mistake}_m$

$$\sim\varphi := \left(\bigvee_{i=1}^{m} \mathsf{mistake}_i\right) \land \bigwedge_{i=1}^{m} \bigwedge_{j=1}^{k} \left(\neg\mathsf{mistake}_i \lor \neg\ell_{i,j}\right)\,.$$

The pseudo-negation operator introduces extension variables simulating the negation of the CNF. We remark that $\sim\varphi$ satisfies the basic properties of negation within Resolution. The following lemma is straightforward and we omit the proof.

Lemma 7.2. *For every CNF formula φ, the formulas $\neg\varphi$ and $\sim\varphi$ are equisatisfiable in the following way: every satisfying assignment to $\neg\varphi$ can be extended into a assignment to $\sim\varphi$, and every satisfying assignment to $\sim\varphi$ restricts into a satisfying assignment for $\neg\varphi$. Furthermore, $\varphi \land \sim\varphi$ has linear-size refutations in Resolution.*

Using pseudo-negations, we can encode the formula (7.2) as

$$\textsc{Sat}(\varphi, \alpha) \land (\pi = P(\varphi, \alpha, s)) \land \sim\textsc{Ref}(\textsc{Ref}_s(\varphi), \pi)\,. \tag{7.3}$$

An important application of pseudo-negations is the ability of carrying out inferences in the style of *modus ponens*. In Frege systems, given a proof of $\varphi \to \psi$ and a proof of φ, one can immediately obtain a proof of ψ. In the case of Resolution, the implication can be written using pseudo-negations to simulate the following behavior (which resembles more a contraposition argument than *modus ponens* proper). The following lemma is due to Jeřábek [Jeř25] and will turn out to be crucial in Section 9. It allows us to perform contraposition on pseudo-negations in Resolution.

Lemma 7.3 (Contraposition in Resolution [Jeř25])**.** *Let $\varphi(x)$ and $\psi(x)$ be two CNF formulas over the same set of variables. If $\psi(x)$ has a Resolution refutation in l steps and size s, and $\varphi(x) \wedge {\sim}\psi(x, \mathsf{mistake})$ has a Resolution refutation of k steps and size t, then $\varphi(x)$ has a Resolution refutation in $l + km$ steps and size $s + tm + k|\psi|$.*

Proof. Let π be a refutation of $\varphi(x) \wedge {\sim}\psi(x, \mathsf{mistake})$ with k steps and total size t. Fix $i \in [m]$, and restrict the refutation π with the substitution that maps $\mathsf{mistake}_i \mapsto 1$ and $\mathsf{mistake}_{i'} \mapsto 0$ for $i' \in [m] \setminus \{i\}$. We obtain a refutation of $\varphi(x) \wedge \bigwedge_{j\in[k]} \neg\ell_{i,j}$ with (at most) the same number of steps and size. Remove the axioms $\neg\ell_{i,j}$, and include C_i in all lines in the refutation. Since $\ell_{i,j} \in C_i$ and we only made changes at the axioms $\neg\ell_{i,j}$, the only steps affected can be initial derivations from axioms. Since in particular the axiom $\neg\ell_{i,j}$ is a literal, this cut is just a Resolution step of the form

$$\frac{D \vee \ell_{i,j} \qquad \neg\ell_{i,j}}{D}$$

which will turn into a trivial weakening step that derives $D \vee C_i$ from $D \vee C_i$. This can be collapsed to a single step, giving us a derivation of C_i from $\varphi(x)$ with k steps and size at most $t + k|C_i|$.

Doing this for all $i \in [m]$ in parallel, we get derivations of all clauses of $\psi(x)$ from $\varphi(x)$. Combining this with a refutation of $\psi(x)$ with l steps and size s, we will obtain a refutation of $\varphi(x)$ with $l + km$ steps and size $s + tm + k|\psi|$. □

7.2 Description of the construction

The next step in the construction of the upper bound is to describe the conjunct $\pi = P(\varphi, \alpha, s)$ in the formula (7.2). The circuit P relates φ and α, which are variables of $\textsc{Sat}(\varphi, \alpha)$, to the variables π of $\textsc{Ref}(\textsc{Ref}_s(\varphi))$. Before we describe the circuit P, we give a natural language explanation of what the upper bound construction is doing.

We remark that in the following description, the formula φ and a satisfying assignment α have been fixed. This means that there are no longer any a-lit^A_ℓ variables encoding the formula (see Definition 2.8). To simplify notation, we sometimes view α as a function over literals so that $\alpha(x_i) = \alpha_i$ and $\alpha(\neg x_i) = \neg\alpha_i$.

General structure of the construction. The goal of the refutation is to derive, for every $B \in [s]$, the clause

$$\mathsf{True}(B, \alpha) \coloneqq \neg\mathsf{enable}^B \vee \bigvee_{\substack{\ell\in\mathrm{Lit}_n \\ \alpha(\ell)=1}} \mathsf{lit}^B_\ell\,, \tag{$\mathsf{True}(B,\alpha)$}$$

encoding that if block B is enabled then it contains a clause that is satisfied by α. Each $\mathsf{True}(B, \alpha)$ is derived from $\mathsf{True}(C, \alpha)$ for all $C \in [B - 1]$, and from $\mathsf{True}(s, \alpha)$ one can easily derive the empty clause.

We will derive the clause $\mathsf{True}(B, \alpha)$ by first deriving the clauses

$$\mathsf{derived}^B \vee \mathsf{True}(B, \alpha) \qquad \text{and} \qquad \neg\mathsf{derived}^B \vee \mathsf{True}(B, \alpha) \tag{7.4}$$

and then applying one Resolution step.

The derivation of $\mathsf{derived}^B \vee \mathrm{True}(B,\alpha)$**.** For every $i \in [m]$ pick an arbitrary literal $\ell \in A_i$ made true by α. By weakening the axiom (Ref-3, B, i, ℓ), which is $\neg\mathsf{enable}^B \vee \neg\mathsf{weak}_i^B \vee \mathsf{lit}_\ell^B$ since the variables a-lit_ℓ^i are no longer present once a formula has been fixed, we obtain the clause

$$\neg\mathsf{enable}^B \vee \neg\mathsf{weak}_i^B \vee \bigvee_{\substack{\ell \in \mathrm{Lit}_n \\ \alpha(\ell)=1}} \mathsf{lit}_\ell^B . \tag{$L_1(B,i)$}$$

Now cut successively (Ref-7, B) with ($L_1(B,i)$) for $i \in [m]$ to get $\mathsf{derived}^B \vee \mathrm{True}(B,\alpha)$. That is, there will be m lines $L_2(B,1)$ to $L_2(B,m)$, each of the form

$$\neg\mathsf{enable}^B \vee \mathsf{derived}^B \vee \bigvee_{j=i+1}^{m} \mathsf{weak}_j^B \vee \bigvee_{\substack{\ell \in \mathrm{Lit}_n \\ \alpha(\ell)=1}} \mathsf{lit}_\ell^B , \tag{$L_2(B,i)$}$$

such that $L_2(B,i+1)$ is obtained by resolving $L_2(B,i)$ with $L_1(B,i+1)$ over variable weak_{i+1}^B. Note that $L_2(B,m)$ is precisely $\mathsf{derived}^B \vee \mathrm{True}(B,\alpha)$.

It is not hard to see that, for each $B \in [s]$, this derivation of $\mathsf{derived}^B \vee \mathrm{True}(B,\alpha)$ consists of $\Theta(m)$ resolution steps.

The derivation of $\neg\mathsf{derived}^B \vee \mathrm{True}(B,\alpha)$**.** We assume that for every $C \in [B-1]$ we have derived $\mathrm{True}(C,\alpha)$. We first carry out the following derivation for every $C \in [B-1]$ and $i \in [n]$, leading to auxiliary clauses ($R_2(B,C,i)$) defined below.

Let us suppose that $\alpha(x_i) = 0$ (the dual case is analogous but following the right-hand side pointers). For $j \in [n]$, let $\ell_j = x_j$ if $\alpha(x_j) = 1$ and $\ell_j = \neg x_j$ if $\alpha(x_j) = 0$. Note that for each $j \in [n]$ we have the axiom (Ref-1, B, C, i, ℓ_j), which is

$$\neg\mathsf{enable}^B \vee \mathsf{res}_{x_i}^B \vee \neg\mathsf{lpoint}_C^B \vee \neg\mathsf{lit}_{\ell_j}^C \vee \mathsf{lit}_{\ell_j}^B . \tag{$R_1^A(B,C,i,j)$}$$

Successively resolving $\mathrm{True}(C,\alpha)$ with ($R_1^A(B,C,i,j)$) over variable $\mathsf{lit}_{\ell_j}^C$ for $j \in [n]$ we get

$$\neg\mathsf{enable}^B \vee \neg\mathsf{enable}^C \vee \neg\mathsf{res}_{x_i}^B \vee \neg\mathsf{lpoint}_C^B \vee \bigvee_{\substack{\ell \in \mathrm{Lit}_n \\ \alpha(\ell)=1}} \mathsf{lit}_\ell^B . \tag{$R_1(B,C,i)$}$$

This consists of n lines, which we refer to as $R_1(B,C,i,j)$ for $j \in [n]$. Note that $R_1(B,C,i,n)$ refers to the same line ($R_1(B,C,i)$). We cut ($R_1(B,C,i)$) over variable enable^C with the clause $R_2^A(B,C,i) \coloneqq \neg\mathsf{enable}^B \vee \neg\mathsf{lpoint}_C^B \vee \mathsf{enable}^C$, which is (Ref-8, B, C), to obtain

$$\neg\mathsf{enable}^B \vee \neg\mathsf{res}_{x_i}^B \vee \neg\mathsf{lpoint}_C^B \vee \bigvee_{\substack{\ell \in \mathrm{Lit}_n \\ \alpha(\ell)=1}} \mathsf{lit}_\ell^B . \tag{$R_2(B,C,i)$}$$

We now cut (Ref-4, B) with ($R_2(B,C,i)$) over variable $\mathsf{res}_{x_i}^B$ for every $i \in [n]$ to get

$$\neg\mathsf{enable}^B \vee \neg\mathsf{derived}^B \vee \neg\mathsf{lpoint}_C^B \vee \bigvee_{\substack{\ell \in \mathrm{Lit}_n \\ \alpha(\ell)=1}} \mathsf{lit}_\ell^B . \tag{$R_3(B,C,i)$}$$

Finally cutting over the variable lpoint_C^B the clause $R_3^A(B,i) \coloneqq \neg\mathsf{enable}^B \vee \neg\mathsf{derived}^B \vee \bigvee_{C \in [B-1]} \mathsf{lpoint}_C^B$,

which is (Ref-5, B), with ($R_3(B, C, i)$) for each $C \in [B-1]$, we get

$$\neg\mathsf{enable}^B \vee \neg\mathsf{derived}^B \vee \bigvee_{\substack{\ell \in \mathrm{Lit}_n \\ \alpha(\ell)=1}} \mathsf{lit}_\ell^B, \tag{$R_4(B)$}$$

which is exactly $\neg\mathsf{derived}^B \vee \mathrm{True}(B, \alpha)$. This step consists of $B-1$ lines, which we denote by $R_4(B, C, i)$ for $C \in [B-1]$, where the last line $R_4(B, B-1, i)$ is ($R_4(B)$).

We note that, for each $B \in [s]$, we derive $\neg\mathsf{derived}^B \vee \mathrm{True}(B, \alpha)$ in $\Theta(sn^2)$ resolution steps.

Contradiction from $\mathrm{True}(s)$**.** Once we have derived $\mathrm{True}(s)$, we resolve it with (Ref-11) to get

$$\bigvee_{\substack{\ell \in \mathrm{Lit}_n \\ \alpha(\ell)=1}} \mathsf{lit}_\ell^B, \tag{7.5}$$

which we then resolve with the axioms (Ref-10, ℓ) for the n literals ℓ such that $\alpha(\ell) = 1$ to get the empty clause.

We observe that the total number of clauses in this resolution refutation of $\mathrm{Ref}_s(\varphi)$, for a satisfiable formula φ, is $\Theta(s(m + sn^2))$.

7.3 The construction as a low-depth circuit

The expression $\pi = P(\varphi, \alpha, s)$ in the formula (7.3) is a short-hard for a conjunction of clauses describing the upper bound construction as presented in Section 7.2. As noted above, the number of clauses in this refutation is $\tau = \Theta(s(m + sn^2))$. We only define $\pi = P(\varphi, \alpha, s)$ for π being a proof of length $t \geq \tau$, as we will only consider this short-hand in this parameter regime. This subsection is dedicated to the description of the conjunction of clauses encoding $\pi = P(\varphi, \alpha, s)$. Each clause C that appears in the construction, including the axioms of Ref that are used in cuts, is mapped to some $\mathcal{B} \in [\tau]$ which corresponds to its position in the proof. Since this a one-to-one mapping, we sometimes abuse notation and identify C with $\mathcal{B}$.

The key observation is that each variable of π is a function of at most two variables of φ and α, and thus the construction can be expressed as a depth-2 circuit. This is because if step i of the construction is, say, to derive a clause C (over variables of $\mathrm{Ref}_s(\varphi)$), then we will add the unit clauses

$$\mathsf{enable}^i \wedge \bigwedge_{z \in C} \mathsf{lit}_z^i \wedge \bigwedge_{z \notin C} \neg\mathsf{lit}_z^i, \tag{7.6}$$

encoding that clause C is the clause at step i in the proof. If the construction obtains this clause by a Resolution step we add the conjunct $\mathsf{derived}^i$ and if it is by a weakening of an axiom we add the conjunct $\neg\mathsf{derived}^i$. Similarly we include appropriate unit clauses of the variables of type lpoint, rpoint and weak. For the most part, the construction is a ready-made template that only depends on α and φ in a few crucial points. In this case, we will include clauses that encode this dependence. For example, suppose that at step i we derive the clause

$$\bigvee_{\substack{\ell \in \mathrm{Lit}_n \\ \alpha(\ell)=1}} \mathsf{lit}_\ell^B \tag{7.7}$$

that does depend on α. Here B is a block of the inner Ref formula, while this clause itself will be instantiated as a block $\mathcal{B} = i$ of the outer Ref. We then add, for $j \in [n]$, the clauses encoding

$$\mathsf{lit}^{\mathcal{B}}_{\mathsf{lit}^B_{x_j}} \leftrightarrow \alpha_j \qquad \text{and} \qquad \mathsf{lit}^{\mathcal{B}}_{\mathsf{lit}^B_{\neg x_j}} \leftrightarrow \neg\alpha_j\,. \tag{7.8}$$

We first describe in more detail how the variables related to the lines $(L_1(B,i))$ and $(L_2(B,i))$ are defined. Line $(L_1(B,i))$ is the only one that depends on both α and φ; the rest of the construction depends solely on α. For subsequent lines in the construction, we simply describe the parts that depend on α as the others are completely determined by the ready-made template, as explained above.

Fix $i \in [m]$, and let $\mathcal{L}$ denote the block corresponding to $(L_1(B,i))$. We include the unit clauses

$$\mathsf{enable}^{\mathcal{L}} \wedge \neg\mathsf{derived}^{\mathcal{L}} \tag{7.9}$$

that determine that this line is enabled and is not derived. We also include unit clauses $\neg\mathsf{res}^{\mathcal{L}}_z$, $\neg\mathsf{lpoint}^{\mathcal{L}}_{\mathcal{B}}$ and $\neg\mathsf{rpoint}^{\mathcal{L}}_{\mathcal{B}}$ for all $\mathcal{B} \in [t]$ (since $\mathcal{L}$ is not obtained by a Resolution step). We set the lit variables as explained above, that is, we have clauses encoding

$$\mathsf{lit}^{\mathcal{L}}_{\mathsf{lit}^B_{x_j}} \leftrightarrow \alpha_j \qquad \text{and} \qquad \mathsf{lit}^{\mathcal{L}}_{\mathsf{lit}^B_{\neg x_j}} \leftrightarrow \neg\alpha_j\,, \tag{7.10}$$

the unit clauses

$$\mathsf{lit}^{\mathcal{L}}_{\neg\mathsf{enable}^B} \wedge \mathsf{lit}^{\mathcal{L}}_{\neg\mathsf{weak}^B_i}\,, \tag{7.11}$$

and the unit clauses $\neg\mathsf{lit}^{\mathcal{L}}_{\ell'}$ encoding that no other literal ℓ' (not appearing above) is present in the clause. It remains to define the $\mathsf{weak}^{\mathcal{L}}$ variables. Note that $(L_1(B,i))$ can be obtained by weakening any of the axioms (Ref-3, B, i, ℓ), for any $\ell \in A_i$ made true by α. We therefore set

$$\mathsf{weak}^{\mathcal{L}}_{(\text{Ref-3},B,i,x_j)} \leftrightarrow \mathsf{a\text{-}lit}^i_{x_j} \wedge \alpha_j \qquad \text{and} \qquad \mathsf{weak}^{\mathcal{L}}_{(\text{Ref-3},B,i,\neg x_j)} \leftrightarrow \mathsf{a\text{-}lit}^i_{\neg x_j} \wedge \neg\alpha_j\,. \tag{7.12}$$

Here the variables $\mathsf{a\text{-}lit}^i_{x_j}$ and $\mathsf{a\text{-}lit}^i_{\neg x_j}$ are part of the $\textsc{Sat}(\varphi,\alpha)$ formula (as in Definition 2.8), and determine which literals appear in what clauses of φ. Observe that we are technically establishing that the clause $(L_1(B,i))$ will have many weakening pointers. While this is somewhat unusual in real Resolution refutations, the refutation is still sound and none of the axioms in Definition 2.4 are violated. The advantage is that we do not have to establish which is the, say, first satisfied literal of every axiom. Even though (potentially) having multiple weakening pointers is not strictly necessary, it simplifies the clauses needed to express $\pi = P(\varphi,\alpha,s)$.

Now fix $i \in [m]$ and let $\mathcal{L}$ be the block corresponding to the $(L_2(B,i))$. We add the unit clauses

$$\mathsf{enable}^{\mathcal{L}} \wedge \mathsf{derived}^{\mathcal{L}} \wedge \mathsf{res}^{\mathcal{L}}_{\mathsf{weak}^B_i} \wedge \bigwedge_{z \neq \mathsf{weak}^B_i} \neg\mathsf{res}^{\mathcal{L}}_z \tag{7.13}$$

that determine that this line is enabled, it is derived and obtained by resolving over variable weak^B_i. We also include unit clauses enforcing all $\mathsf{weak}^{\mathcal{L}}$ variables to be 0 (since $\mathcal{L}$ is not a weakening of any axiom). The $\mathsf{lit}^{\mathcal{L}}$ variables are encoded as explained above. As for the pointers, we include the unit clauses

$$\mathsf{rpoint}^{\mathcal{L}}_{L_1(B,i)} \wedge \mathsf{lpoint}^{\mathcal{L}}_{L_2(B,i-1)} \wedge \bigwedge_{\mathcal{B} \neq L_1(B,i)} \neg\mathsf{rpoint}^{\mathcal{L}}_{\mathcal{B}} \wedge \bigwedge_{\mathcal{B} \neq L_2(B,i-1)} \neg\mathsf{lpoint}^{\mathcal{L}}_{\mathcal{B}}\,. \tag{7.14}$$

For the second set of clauses, used in the derivation of $\neg\mathsf{derived}^B \vee \mathrm{True}(B,\alpha)$, we describe the parts that depend on α. These can be split into two groups, the parts that depend only on whether we consider lpoint or rpoint, that is, only on α_i, and those that depend also on ℓ_j. We start with the former. If in the construction we described above a clause on a line corresponding to, say, block $\mathcal{R}$ contains the literal

$\neg\mathsf{lpoint}^B_C$ for the case when $\alpha_i = 0$, then we include the clauses encoding

$$\mathsf{lit}^{\mathcal{R}}_{\neg\mathsf{rpoint}^B_C} \leftrightarrow \alpha_i \qquad \text{and} \qquad \mathsf{lit}^{\mathcal{R}}_{\neg\mathsf{lpoint}^B_C} \leftrightarrow \neg\alpha_i\,, \tag{7.15}$$

so that the literal $\neg\mathsf{rpoint}^B_C$ is present iff $\alpha(x_i) = 1$ and $\neg\mathsf{lpoint}^B_C$ is present if and only if $\alpha(x_i) = 0$. Similarly if it contains the literal lpoint^B_C, we include $\mathsf{lit}^{\mathcal{R}}_{\mathsf{rpoint}^B_C} \leftrightarrow \alpha_i$ and $\mathsf{lit}^{\mathcal{R}}_{\mathsf{lpoint}^B_C} \leftrightarrow \neg\alpha_i$. For $\mathcal{R} = R_4(B, C, i)$, we include clauses

$$\mathsf{res}^{\mathcal{R}}_{\mathsf{rpoint}^B_C} \leftrightarrow \alpha_i \qquad \text{and} \qquad \mathsf{res}^{\mathcal{R}}_{\mathsf{lpoint}^B_C} \leftrightarrow \neg\alpha_i\,, \tag{7.16}$$

encoding that the resolved variable is either rpoint^B_C or lpoint^B_C, depending on α_i. For $\mathcal{R} = R^A_2(B, C, i)$ we include

$$\mathsf{weak}^{\mathcal{R}}_{(\textsc{Ref}\text{-}9,B,C)} \leftrightarrow \alpha_i \qquad \text{and} \qquad \mathsf{weak}^{\mathcal{R}}_{(\textsc{Ref}\text{-}8,B,C)} \leftrightarrow \neg\alpha_i\,, \tag{7.17}$$

and for $\mathcal{R} = R^A_3(B, i)$ we include

$$\mathsf{weak}^{\mathcal{R}}_{(\textsc{Ref}\text{-}6,B)} \leftrightarrow \alpha_i \qquad \text{and} \qquad \mathsf{weak}^{\mathcal{R}}_{(\textsc{Ref}\text{-}5,B)} \leftrightarrow \neg\alpha_i\,. \tag{7.18}$$

We now move to the parts that (also) depend on ℓ_j. If the line corresponding to block $\mathcal{R}$ contains $\mathsf{lit}^{B'}_{\ell_j}$ for some block B' we include

$$\mathsf{lit}^{\mathcal{R}}_{\mathsf{lit}^{B'}_{x_j}} \leftrightarrow \alpha_j \qquad \text{and} \qquad \mathsf{lit}^{\mathcal{R}}_{\mathsf{lit}^{B'}_{\neg x_j}} \leftrightarrow \neg\alpha_j\,. \tag{7.19}$$

Similarly, if it contains $\neg\mathsf{lit}^{B'}_{\ell_j}$ for some block B' we include $\mathsf{lit}^{\mathcal{R}}_{\neg\mathsf{lit}^{B'}_{x_j}} \leftrightarrow \alpha_j$ and $\mathsf{lit}^{\mathcal{R}}_{\neg\mathsf{lit}^{B'}_{\neg x_j}} \leftrightarrow \neg\alpha_j$. If $\mathcal{R}$ is a line $R_1(B, C, i, j)$ that is obtained by resolving over variable $\mathsf{lit}^C_{\ell_j}$, we include the clauses

$$\mathsf{res}^{\mathcal{R}}_{\mathsf{lit}^C_{x_j}} \leftrightarrow \alpha_j \qquad \text{and} \qquad \mathsf{res}^{\mathcal{R}}_{\mathsf{lit}^C_{\neg x_j}} \leftrightarrow \neg\alpha_j\,. \tag{7.20}$$

Finally, for $\mathcal{R}$ being $R^A_1(B, C, i, j)$, we include

$$\mathsf{weak}^{\mathcal{R}}_{(\textsc{Ref}\text{-}2,B,C,i,x_j)} \leftrightarrow \alpha_i \wedge \alpha_j\,, \qquad \mathsf{weak}^{\mathcal{R}}_{(\textsc{Ref}\text{-}1,B,C,i,x_j)} \leftrightarrow \neg\alpha_i \wedge \alpha_j\,, \tag{7.21}$$

$$\mathsf{weak}^{\mathcal{R}}_{(\textsc{Ref}\text{-}2,B,C,i,\neg x_j)} \leftrightarrow \alpha_i \wedge \neg\alpha_j\,, \qquad \text{and} \qquad \mathsf{weak}^{\mathcal{R}}_{(\textsc{Ref}\text{-}1,B,C,i,\neg x_j)} \leftrightarrow \neg\alpha_i \wedge \neg\alpha_j\,. \tag{7.22}$$

We have thus far described how we assign variable referring to blocks $\mathcal{B}$ of π for $\mathcal{B} \le \tau$ (where we recall $\tau = \Theta(s(m + sn^2))$ is the number of clauses in the refutation of $\textsc{Ref}_s(\varphi)$ described in Section 7.2). For $\mathcal{B} \in [t]$ and $\mathcal{B} > \tau$ we include unit clauses setting all variables to 0, in particular, we include clauses $\neg\mathsf{enable}^{\mathcal{B}}$ encoding that these blocks are not used in the refutations described by $P(\varphi, \alpha)$.

We denote by $\pi = P(\varphi, \alpha)$ all the conjuncts describing the refutation of $\textsc{Ref}_s(\varphi)$ as an assignment to the outer Ref formula.

7.4 Correctness of the construction in Resolution

We are ready to prove that the construction is provably correct in Resolution.

Theorem 7.4 (Pudlák's upper bound in Resolution). *For every $n, m, s \in \mathbb{N}$, there is a $\tau = \Theta(s(m + sn^2))$ such that for $t \ge \tau$ there exist uniform polynomial-size Resolution refutations of the formula*

$$\textsc{Sat}(\varphi, \alpha) \wedge {\sim}\textsc{Ref}_t(\textsc{Ref}_s(\varphi), \pi) \wedge \pi = P(\varphi, \alpha, s)\,,$$

where the SAT *formula is for CNF formulas with n variables and m clauses, and $\pi = P(\varphi, \alpha, s)$ stands for the conjunction of clauses describing the upper bound construction, as per Section 7.3. Moreover, the refutations can be generated uniformly in polynomial time.*

Proof. Let $\tau = \Theta(s(m + sn^2))$ be the number of clauses in the refutation of $\text{REF}_s(\varphi)$ described in Section 7.2. Let $\mathfrak{C}$ denote the set of clauses in the formula $\text{REF}_t(\text{REF}_s(\varphi), \pi)$. Observe that these clauses can be partitioned according to which type of axiom (REF-1 - REF-11) they belong to, as well as to the block $\mathcal{B} \in [t]$ they refer to. By using the pseudo-negation operation $\sim\text{REF}_t(\cdot)$, we have introduced a clause $\bigvee_{c\in\mathfrak{C}} \text{mistake}_c$.

Observe first that for every satisfiable φ and for every satisfying assignment α of φ, the assignment described by the conjuncts $\pi = P(\varphi, \alpha, s)$ is a sound Resolution refutation, meaning that no axiom of $\text{REF}_t(\cdot)$ can be violated. The goal will be to derive $\neg\text{mistake}_c$ for every $c \in \mathfrak{C}$, and then derive contradiction by cutting this with $\bigvee_{c\in\mathfrak{C}} \text{mistake}_c$. We first argue that if the variables of c only depend on α, then Resolution can derive $\neg\text{mistake}_c$. We are then left to show that Resolution can also derive $\neg\text{mistake}_c$ for the axioms that contain variables that depend on both α of φ.

Now, let $c \in \mathfrak{C}$ be a clause of $\text{REF}_t(\cdot)$. By inspection of the description of the refutation as a circuit in Section 7.3, it is clear that the only variables that depend on both a-lit and α variables are precisely the $\text{weak}^{\mathcal{L}}_{(\text{REF-3},B,i,\ell)}$ variables in line (7.12). For the axioms of constant width that do not contain the variable $\text{weak}^{\mathcal{L}}_{(\text{REF-3},B,i,\ell)}$, that is, axioms (REF-1 - REF-2) and (REF-8 - REF-11), as well as some of the axioms (REF-3), their variables depend on a constant number of α variables in the description $\pi = P(\varphi, \alpha, s)$. By brute force, since the step is sound for every α, Resolution can just derive for every one of them the unit clause $\neg\text{mistake}_c$. For axioms (REF-3) that contain $\text{weak}^{\mathcal{L}}_{(\text{REF-3},B,i,x_j)}$ the argument is similar, since for any value of φ and α, if $\text{weak}^{\mathcal{L}}_{(\text{REF-3},B,i,\ell)} = 1$ (that is, $\text{a-lit}^i_\ell = 1$ and $\alpha(\ell) = 1$) then indeed the clause at block $\mathcal{L}$ is a weakening of (REF-3, B, i, ℓ).

Now, for all of the axioms of types (REF-5) and (REF-6) note that all the variables appearing in them are completely set, that is, do not depend on α or φ, so Resolution can also derive the unit clause $\neg\text{mistake}_c$. The same holds for any other axiom whose variables are completely set. This leaves us with the axioms of type (REF-4) and (REF-7) whose variables are not completely set. For the axioms of type (REF-4) and (REF-7) that depend on variables defined in (7.16)-(7.18) and (7.20),

note that we set one of the variables in the wide disjunction (one of res or weak variables) to α_i and another to $\neg\alpha_i$. This is sound (for any value of α_i there will be at least one variable in the wide disjunction that is set to 1) and only depends on one α_i, so Resolution can derive $\neg\text{mistake}_c$. The case of axioms REF-7 that depend on (7.21) and (7.22) is similar: by a brute-force over the possible values of α_i and α_j, Resolution can derive that one of the $\text{weak}^{\mathcal{R}}$ variables must be 1.

Finally, we argue that if c is one of the axiom (REF-7) that depend on (7.12), Resolution can also derive $\neg\text{mistake}_c$. Note that $\pi = P(\varphi, \alpha, s)$ contains the clause

$$\text{weak}^{\mathcal{L}}_{(\text{REF-3},B,A,x_j)} \vee \neg\text{a-lit}^A_{x_j} \vee \neg\alpha_j \tag{7.23}$$

and the clause

$$\text{weak}^{\mathcal{L}}_{(\text{REF-3},B,A,\neg x_j)} \vee \neg\text{a-lit}^A_{\neg x_j} \vee \alpha_j\,, \tag{7.24}$$

for $A \in [m]$. By resolving the clause (7.23) with (SAT-1, A, x_j) and then with (SAT-2, A, j), we obtain the clause

$$\text{weak}^{\mathcal{L}}_{(\text{REF-3},B,A,x_j)} \vee \text{sat}^A_{x_j}\,. \tag{7.25}$$

Similarly, we can derive

$$\text{weak}^{\mathcal{L}}_{(\text{REF-3},B,A,x_j)} \vee \text{sat}^A_{\neg x_j} \tag{7.26}$$

by resolving (7.24) with (SAT-1, A, $\neg x_j$) and then with (SAT-3, A, j). Resolving (7.25) and (7.26) with (SAT-4)

we obtain

$$\bigvee_{i\in[n]} \left(\text{weak}^{\mathcal{L}}_{(\text{Ref-3},B,A,x_j)} \vee \text{weak}^{\mathcal{L}}_{(\text{Ref-3},B,A,\neg x_j)}\right). \tag{7.27}$$

Resolving this clause with the clauses $\neg\text{mistake}_c \vee \neg\text{weak}^{\mathcal{L}}_{\mathcal{A}}$ from $\sim\text{Ref}_t(\text{Ref}_s(\varphi),\pi)$, for all axioms $\mathcal{A}$ in $\{(\text{Ref-3},B,A,x_i),(\text{Ref-3},B,A,\neg x_i)\}_{i\in[n]}$, we derive $\neg\text{mistake}_c$. This completes the proof as we have derived $\neg\text{mistake}_c$ for all $c \in \mathfrak{C}$. □

8 NP-hardness of automating Resolution in EF

We now have the lower bound (Theorem 6.8) needed to show the NP-hardness of automatability formalized in PV_1. This is a $\forall\Sigma^b_1$ sentence, meaning that the existential quantifiers can be witnessed by polynomial-time functions and via Cook's translation turned into propositional formulas with short Extended Frege proofs.

Theorem 8.1 (Extraction algorithm in EF)**.** *There exists a uniform family of polynomial-size circuits $\{E_{n,m,s,t}\}_{n,m,s,t\in\mathbb{N}}$ such that for every CNF formula φ over n variables and m clauses, if φ is satisfiable and π is a Resolution refutation of $\text{Ref}_s(\varphi)$ for $s \geq n^3$, then the circuit $E_{n,m,s,t}(\varphi,\pi)$ outputs a satisfying assignment for φ. Furthermore, this is provable in Extended Frege, that is, there exists a polynomial $\ell(n,m,s,t)$ such that*

$$\mathsf{EF} \vdash_{\ell(n,m,s,t)} \text{Ref}_t(\text{Ref}_s(\varphi),\pi) \rightarrow \text{Sat}(\varphi, E_{n,m,s,t}(\varphi,\pi)),$$

and these proofs can be generated uniformly in polynomial time.

Proof. By Theorem 6.8, we know there exists a positive $\varepsilon \in \mathbb{Q}$ and $n_0 \in \mathbb{N}$ such that $\mathsf{PV}_1 \vdash \text{AM}_{\varepsilon,n_0}$, the first-order statement encoding the lower bound on Ref formulas ($\text{AM}_{\varepsilon,n_0}$). Since $\text{AM}_{\varepsilon,n_0}$ is a $\forall\Sigma^b_1$ formula, by Buss's witnessing theorem (Theorem 2.2), there exists a PV function E_0 such that

$$\begin{aligned}\mathsf{PV}_1 \vdash \forall\varphi\forall n\forall s\forall\pi\Big(& n < n_0 \vee \neg\,\text{CNF}(\varphi,n) \vee \neg\,\text{Ref}_{\text{Res}}(\text{Ref}_s(\varphi),\pi) \\ & \vee \text{Sat}(\varphi, E_0(\varphi,n,s,\pi)) \vee ||\pi|| > \varepsilon s/n^2\Big).\end{aligned} \tag{8.1}$$

This means that, when plugging a formula φ that is indeed a CNF formula over $n > n_0$ variables and the length of π is $|\pi| \leq 2^{\varepsilon n}$ and $s \geq n^3$, then the witnessing function E_0 correctly succeeds in finding a satisfying assignment of φ, and this is all provable in PV_1. Building on this E_0, we can further define a new polynomial-time function E that first checks whether $|\pi| \leq 2^{\varepsilon n}$, and then runs E_0, and otherwise performs a brute-force check for a satisfying assignment for φ. It is not hard to see, building on eq. (8.1), that PV_1 can now prove that this function E is a polynomial-time algorithm that always succeeds in finding a satisfying assignment if one exists, regardless of the size of π.

By applying Cook's translation (Theorem 2.3) to eq. (8.1) and restricting $s \geq n^3$, we obtain a uniform family of EF proofs showing the correctness of E as an extraction algorithm. Note that $\text{Ref}_{\text{Res}}(\varphi,\pi)$ is the first-order analogue of the propositional formula $\text{Ref}_s(\varphi,\pi)$ that we have been studying until now. Cook's translation on $\text{Ref}_{\text{Res}}(\varphi,\pi)$ gives a propositional formula with the same behavior as $\text{Ref}_s(\varphi)$. While $[\![\text{Ref}_{\text{Res}}]\!]_{n,s}$ may not be syntactically the same as $\text{Ref}_s(\varphi,\pi)$, it is not hard to see that $\mathsf{PV}_1 \vdash \forall n\forall s\,\big(\text{Taut}([\![\text{Ref}_{\text{Res}}]\!]_{n,s}) \leftrightarrow \text{Taut}(\ulcorner\text{Ref}_s(\varphi)\urcorner)\big)$, making them effectively equivalent.

Since the proofs generated by Cook's translation are all polynomial-size and uniform, we conclude that there exists a polynomial $\ell(n,m,s,t)$ such that for every $s \geq n^3$,

$$\mathsf{EF} \vdash_{\ell(n,m,s,t)} \text{Ref}_t(\text{Ref}_s(\varphi,\tau),\pi) \rightarrow \text{Sat}(\varphi, E_{n,m,s,t}(\varphi,\pi)), \tag{8.2}$$

where $E_{n,m,s,t}$ is the polynomial-size circuit encoding the computation of the extraction algorithm E on CNF formulas with n variables and m clauses, size parameter s and the refutation being analyzed consist of t clauses. □

Similarly, the upper bound construction of Pudlák, which we formalized in Resolution, clearly goes through in stronger systems too (Theorem 7.4).

Suppose now that $A = \{A_{n,m,s}\}_{n\in\mathbb{N}}$ is a family of circuits that automate Resolution, meaning that there is a constant $c \in \mathbb{N}$ such that on input a CNF formula φ over n variables and m clauses, the circuit $A_{n,m,s}(\varphi)$ outputs a Resolution refutation of φ of size s^c if a refutation of size s exists. Consider the propositional formulas stating the correctness of this algorithm,

$$\textsc{Aut}^A_{n,m,s} := \textsc{Ref}_s(\varphi, \pi) \to \textsc{Ref}_{s^c}(\varphi, A_{n,m,s}(\varphi)). \tag{$\textsc{Aut}^A$}$$

We want to prove that if Extended Frege can efficiently derive the ($\textsc{Aut}^A$) formulas above for some sequence of circuits $\{A_{n,m,s}\}_{n,m,s\in\mathbb{N}}$, then it can also show that there is a family of small circuits $\{C_n\}_{n\in\mathbb{N}}$ solving SAT on 3-CNF formulas. We encode this as

$$\textsc{Sat}^C_n := \textsc{Sat}(\varphi, \alpha) \to \textsc{Sat}(\varphi, C_n(\varphi)). \tag{$\textsc{Sat}^C$}$$

We show that EF can derive ($\textsc{Sat}^C$) from ($\textsc{Aut}^A$).

Theorem 8.2 (Resolution is **NP**-hard to automate, in EF)**.** *Suppose Resolution is automatable by a sequence of circuits $A = \{A_{n,m,s}\}_{n,m,s\in\mathbb{N}}$, and suppose $\mathsf{EF} \vdash_{\text{poly}} \textsc{Aut}^A_{n,m,s}$. Then, there exists a family of circuits $C = \{C_n\}_{n\in\mathbb{N}}$ of size $n^{O(1)}$ such that $\mathsf{EF} \vdash_{n^{O(1)}} \textsc{Sat}^C_n$.*

Proof. Let $E = \{E_{n,m,s,t}\}_{n,m,s,t\in\mathbb{N}}$ be the sequence of polynomial-size circuits carrying out the computation of the extraction algorithm, as in Theorem 8.1, and let $\{P_{n,m,s}\}_{n,m,s\in\mathbb{N}}$ be the sequence of circuits constructing the canonical refutations of Pudlák. By Theorem 8.1, Theorem 7.4 and the assumptions in the statement, we have that Extended Frege can efficiently prove

(i) the correctness of the upper bound on Ref formulas, $\mathsf{EF} \vdash_{\text{poly}} \textsc{Sat}_n(\varphi, \alpha) \to \textsc{Ref}_t(\textsc{Ref}_s(\varphi), P(\varphi, \alpha, s))$, for any $t \geq \tau$, where $\tau = \tau(n, m, s) = \text{poly}(n, m, s)$ is the parameter in Theorem 7.4;

(ii) the correctness of the automating circuits A, $\mathsf{EF} \vdash_{\text{poly}} \textsc{Ref}_s(\varphi, \pi) \to \textsc{Ref}_{s^c}(\varphi, A_{n,m,s}(\varphi))$;

(iii) the correctness of the extraction algorithm E, $\mathsf{EF} \vdash_{\text{poly}} \textsc{Ref}_t(\textsc{Ref}_s(\varphi), \pi) \to \textsc{Sat}_n(\varphi, E_{n,m,s,t}(\varphi, \pi))$, as long as $s \geq n^3$.

The circuit C_n solving SAT will be $C_n(\varphi) := E_{n,8n^3,n^3,t^c}(\varphi, A_{n,m,t}(\textsc{Ref}_{n^3}(\varphi)))$ for a large enough $t \geq \tau = \text{poly}(n)$, where τ is again the parameter in Theorem 7.4. Namely, we use the automating circuit A to produce a Resolution refutation from which we can extract a satisfying assignment using the extraction algorithm E. In between, the upper bound statement guarantees that a small refutation exists, and thus A will succeed in finding a not much larger one. The parameters are set up so that the three statements above correctly fit with each other: given a 3-CNF formula φ with n variables (and at most $8n^3$ clauses), we know that since φ is satisfiable, there is a size-τ refutation of $\textsc{Ref}_s(\varphi)$, where $\tau = \text{poly}(n, s)$ and $s = n^3$. The extraction algorithm is guaranteed to work when $s \geq n^3$, so we choose to run the automating algorithm on $\textsc{Ref}_{n^3}(\varphi)$. The automating algorithm is guaranteed to find a refutation of size t^c, and the extraction algorithm obtains a satisfying assignment from it. The fact that C is provably correct in EF then follows immediately by a chain of *modus ponens* on items (i)-(iii) above. □

9 Universality of Ref formulas

The formalizations in Theorem 6.8 and Theorem 7.4 have the interesting consequence of tightly relating the provability of arbitrary formulas to the provability of associated Resolution lower bounds. In Section 9.1 we make this precise for Extended Frege and stronger systems. In Section 9.2 we exploit this characterization to show that looking for proofs of tautologies in strong proof systems is equivalent to efficiently looking for proofs of Resolution lower bounds. Finally, Section 9.3 exploits similar ideas to provide examples of true exponential Resolution lower bounds that are unprovable in different propositional systems and first-order theories.

Our results apply to a wide range of proof systems, assuming some mild conditions on their behaviour.

Definition 9.1. We say that a proof system S is *reasonably strong* if it is polynomially equivalent to $\mathsf{EF} + A$ for some set A of tautologies recognizable in polynomial time.

Recall that for every Cook-Reckhow system S, we have $\mathsf{EF} + \textsc{Refl}^S \geq_{\mathrm{p}} S$ (see also Section 2.2.2 for the definition of systems of the form $\mathsf{EF} + A$). Reasonably strong proof systems have some features that make them nice to work with. We state them here for convenience.

Proposition 9.2 ([Kra25, Theorem 2.4.4])**.** *The following hold for every reasonably strong proof system S:*

(i) the system S is constructively closed under formula substitutions, i.e., given a proof of $\varphi(x_1, \ldots, x_n)$ in size s and formulas $\psi_1, \ldots, \psi_n$, we can obtain a proof of $\varphi(\psi_1, \ldots, \psi_n)$ in time $\mathrm{poly}(s, n, |\psi_1|, \ldots, |\psi_n|)$*;*

(ii) the system S is constructively closed under modus ponens*, i.e., given an S-proof of φ in size s and an S-proof of $\varphi \to \psi$ in size t, we can obtain an S-proof of ψ in time* $\mathrm{poly}(s, t)$*.*

When dealing with weaker systems, we often impose the following closure property on the ability of simulating *modus ponens* via pseudo-negations.

Definition 9.3. Let S be a propositional proof system. We say that the system S is *closed under contraposition for pseudo-negations* if the analogue of Lemma 7.3 holds for S, i.e., given an S-refutation of $\varphi \wedge {\sim}\psi$ in size s and an S-refutation of ψ in size t, there is an S-refutation of φ in size $\mathrm{poly}(s, t)$.

We note that, as expected, the usual notion of *modus ponens* subsumes the more *ad hoc* contraposition under pseudo-negations from Lemma 7.3.

Proposition 9.4. *Every propositional proof system closed under restrictions and* modus ponens *is also closed under contraposition for pseudo-negations.*

Proof. Suppose $\varphi(x)$ and $\psi(x)$ are unsatisfiable CNF formulas, and suppose that S has a proof π of the tautology $\neg(\varphi(x) \wedge {\sim}\psi(x, \mathrm{mistake}))$ in size s and a proof of $\neg\psi(x)$ in size t. Since S is closed under restrictions, for every clause C of ψ, we can restrict π by setting the mistake variables corresponding to C to 1 and the rest to 0, getting a proof of $\neg(\varphi(x) \wedge \neg C) \equiv \varphi(x) \to C$. Doing this in parallel for every clause C, we can combine them in a proof of $\varphi(x) \to \bigwedge_{C \in \psi} C$, which is precisely $\varphi(x) \to \psi(x)$. Now, by a *modus ponens* (or, rather, contraposition) inference on this with $\neg\psi(x)$ we get an S-proof of $\neg\varphi(x)$. □

We need one more definition before we proceed.

Definition 9.5. Let $\varphi(x_1, \ldots, x_n)$ be a CNF formula and let $1 \leq k \leq n$. We define $\varphi[k]$ to be the CNF formula $\varphi \wedge \mathsf{Ext}^k_n$, where Ext^k_n consists of all clauses encoding the extensions

$$y_{\{\ell_1, \ldots, \ell_k\}} \leftrightarrow \ell_1 \wedge \cdots \wedge \ell_k$$

for every set of literals $\{\ell_1, \ldots, \ell_k\}$ over the variables $x_1, \ldots, x_n$, where the y-variables are fresh variables.

It is well-known that if $\mathrm{Res}(k)$ refutes φ in size s, then Resolution refutes $\varphi[k]$ in size $O(ks)$, and if Resolution refutes $\varphi[k]$ in size s, then $\mathrm{Res}(k)$ refutes φ in size $O(ks)$ [AB04, Lemmas 1-2].

The following lemma is an important consequence of the formalization of Pudlák's upper bound.

Lemma 9.6. *Let $S \geq \mathrm{Res}$ be a propositional proof system closed under literal substitutions and contraposition for pseudo-negations. Then, there exists $p(n,m,s) = \mathrm{poly}(n,m,s)$ such that for every n-variate CNF formula φ with m clauses, and every $t \geq p(n,m,s)$ if $S \vdash \neg\mathrm{Ref}_t(\mathrm{Ref}_s(\varphi))$ in size ℓ, then $S \vdash \neg\varphi[2]$ in size $\mathrm{poly}(n,m,s,t,\ell)$.*

Proof. We work in the refutational setting, meaning that the fact $S \vdash \neg\mathrm{Ref}_t(\mathrm{Ref}_s(\varphi),\pi)$ in size ℓ is rephrased as saying that S has a size-ℓ refutation of $\mathrm{Ref}_t(\mathrm{Ref}_s(\varphi),\pi)$. By Theorem 7.4, we have that Resolution has polynomial-size refutations of

$$\mathrm{SAT}_{\restriction\varphi}(\alpha) \wedge {\sim}\mathrm{Ref}_t(\mathrm{Ref}_s(\varphi),\pi) \wedge \pi = P_{\restriction\varphi}(\alpha)\,, \tag{9.1}$$

where φ is fixed but α is encoded by free variables, and we define $p(n,m,s) = \tau(n,m,s) = \mathrm{poly}(n,m,s)$ to be the parameter in Theorem 7.4. Since S is closed under contraposition for pseudo-negations, Lemma 7.3 applies and we get that there are polynomial-size S-refutations of

$$\mathrm{SAT}_{\restriction\varphi}(\alpha) \wedge \pi = P_{\restriction\varphi}(\alpha)\,. \tag{9.2}$$

Now, by Proposition 2.9.(i), we can use a substitution on the variables of Sat to get a refutation of $\varphi(\alpha) \wedge \pi = P_{\restriction\varphi}(\alpha)$. By the way we defined the clauses in $\pi = P_{\restriction\varphi}(\alpha)$ in Section 7.3, once φ has been fixed, there are no variables of type a-lit left. By inspecting all the extension axiom from (7.9) to (7.22) we see that all extension axioms relate variables of π to literals over the α variables. Crucially, the width of these extensions is at most two, given by the extensions (7.21) and (7.22). Furthermore, all possible extensions over two α literals are available, meaning that this is in fact a refutation of $\varphi[2]$. □

9.1 Propositional fragments of the Atserias–Müller lower bound

As a consequence of Theorem 8.1 and Theorem 7.4, we can characterize the exact fragment of the Atserias–Müller lower bound efficiently provable by every strong enough propositional proof system. Namely, if EF can argue that φ is unsatisfiable, then it can also argue that $\mathrm{Ref}_{n^3}(\varphi)$ is hard for Resolution; and, conversely, if $\mathrm{Ref}_{n^3}(\varphi)$ is provably hard for Resolution, then EF can argue that φ itself is unsatisfiable. The key insight here is that the statement "$\mathrm{Ref}_{n^3}(\varphi)$ is hard for Resolution" is exactly a Ref formula itself, of the form $\mathrm{Ref}(\mathrm{Ref}(\varphi))$.

The following is a formal restatement of Theorem 1.7.

Theorem 9.7. *Let S be a reasonably strong proof system. There is $p(n,m,s) = \mathrm{poly}(n,m,s)$ such that for every CNF formula $\varphi(x_1,\ldots,x_n)$ with m clauses, every $t \in \mathbb{N}$, and every $s \geq n^3$,*

(i) if $S \vdash \neg\varphi$ in size ℓ, then $S \vdash \neg\mathrm{Ref}_t(\mathrm{Ref}_s(\varphi_n))$ in size $\mathrm{poly}(n,m,s,t,\ell)$;

(ii) if $S \vdash \neg\mathrm{Ref}_{p(n,m,s)}(\mathrm{Ref}_s(\varphi_n))$ in size ℓ, then $S \vdash \neg\varphi_n$ in size $\mathrm{poly}(n,m,s,\ell)$.

Furthermore, in both cases, given an S-proof of the left-hand side statement, one can obtain a proof of the right-hand side statement in polynomial time.

Proof. In the following Ref formulas φ is a fixed formula, while the only free variables are the π and α variables.

(i) If $S \vdash \neg\varphi$ via some proof π of size ℓ, then by Proposition 2.9, S also proves $\neg\mathrm{SAT}_{\restriction\varphi}(\alpha)$ and we can easily substitute the extraction circuit $E_{\restriction\varphi}(\pi)$ into α, getting a proof π' of $\neg\mathrm{SAT}_{\restriction\varphi}(E_{\restriction\varphi}(\pi))$. By Theorem 8.1,

EF proves $\textsc{Ref}_t(\textsc{Ref}_s(\varphi,\tau),\pi) \to \textsc{Sat}(\varphi, E(\varphi,\pi))$ and by the p-simulation $S \geq_{\mathrm{p}} \mathsf{EF}$ these proofs are also available in S. Hence, by contraposition and the restriction on φ, S derives $\neg\textsc{Ref}_t(\textsc{Ref}_s(\varphi,\tau),\pi)$, and it is easy to see that the total blow-up incurred by these additional derivations is at most $\mathrm{poly}(n, m, s, t, \ell)$.

(ii) This backwards direction follows from Lemma 9.6 above. Since S is reasonably strong, it is closed under restrictions and *modus ponens*, and by Proposition 9.4 S is also closed under contraposition for pseudo-negations. Hence, Lemma 9.6 applies, giving us polynomial-size refutations of $\varphi[2]$. The system S is in particular closed under clause substitutions, so we can substitute the extension axioms introduced by $\varphi[2]$ to get a refutation of φ. □

9.2 Automatability in terms of Ref formulas

We will say that a proof system S is *(weakly) automatable on* Ref *formulas* if there is an algorithm A that behaves like an automating algorithm whenever the input formula is a $\textsc{Ref}_s(\varphi)$ formula for some φ and s, and is allowed to behave in any other way otherwise.

We remark that these are the Ref formulas *for Resolution*, meaning that proof search in arbitrarily strong proof systems can be completely characterized by the proof search of Resolution lower bounds! The following theorems are the formal restatements of Theorem 1.8.

Theorem 9.8. *Let S be a reasonably strong proof system. Then, the following are equivalent:*

(i) S is automatable;

(ii) S is automatable on Ref *formulas.*

Proof. One direction is trivial. For the other implication, let A be an automating algorithm that is only guaranteed to work on tautologies of the form $\neg\textsc{Ref}_s(\cdot)$, and let φ be an unsatisfiable CNF formula with n variables and m clauses. We describe a new algorithm A' that correctly automates S on all inputs, meaning that if $\neg\varphi$ has an S-refutation in size t, then A' outputs a refutation in time $t^{O(1)}$.

By the formalization of the correctness of the extraction algorithm in Theorem 8.1, we obtained Theorem 9.7 from where it follows that if S has a refutation of φ in size t, then there exists an S-refutation of $\textsc{Ref}_{p(n,m)}(\textsc{Ref}_{n^3}(\varphi))$ for a fixed polynomial $p(n, m)$, in size $\mathrm{poly}(n, m, t)$. This is precisely a $\textsc{Ref}(\cdot)$ formula, so we can run $A(\textsc{Ref}_{p(n,m)}(\textsc{Ref}_{n^3}(\varphi)))$ to obtain a refutation π in S with at most a fixed polynomial blow-up. Now, by the backwards direction of Theorem 9.7, we known that S has a refutation of φ in size $\mathrm{poly}(n, m, t)$ and, we can obtain this efficiently from π. Thus, we have an automating algorithm that given φ, which had a refutation in S in size t, outputs a refutation in size $t^{O(1)}$. □

Since nothing in the argument above prevents us from talking about proofs in stronger systems, the result immediately carries over to weak automatability. For the following statement, we use the well-known fact that a proof system S is weakly automatable if, and only if, there exists an automatable proof system Q that simulates S. If we restrict ourselves to Ref formulas, then it is easy to see that S is weakly automatable on Ref formulas if, and only if, there is a proof system Q automatable on Ref formulas and simulating S over Ref formulas.

Theorem 9.9. *Let S be a reasonably strong proof system. Then, the following are equivalent:*

(i) S is weakly automatable;

(ii) S is weakly automatable on Ref *formulas.*

Proof. The argument is identical. The only difference is that instead of an automating algorithm for S we have a proof system Q that simulates S on Ref formulas and is automatable on Ref formulas. Then, if $\neg\varphi$ can be refuted in size t in S, we can still guarantee (by Theorem 9.7) that there is a proof of $\textsc{Ref}_{p(n,m)}(\textsc{Ref}_{n^3}(\varphi))$, for a fixed polynomial $p(n,m)$, in size $\text{poly}(n,m,t)$. We look for these proofs with the automating algorithm for Q. Since Q itself may not be reasonably strong itself, we turn this Q-proof into a proof in the system $\mathsf{EF} + \textsc{Refl}^Q$, which p-simulates Q, and for which Theorem 9.7 does apply. □

The argument above crucially requires that S is at least as strong as Extended Frege. The reason is that we require the correctness statement of the extraction algorithm to be provable in S. If φ has a refutation of size t, then Theorem 9.7.(i) guarantees that there is a size-$t^{O(1)}$ refutation of $\textsc{Ref}(\textsc{Ref}(\varphi))$, so the algorithm can focus on looking for these latter refutations instead. For an arbitrary proof system $S \geq \text{Res}$ that does not simulate EF, if S proves $\neg\varphi$ in size t, then we cannot guarantee that it also proves $\neg\textsc{Ref}(\textsc{Ref}(\varphi))$ in size $t^{O(1)}$.

This problem can be solved if S is Resolution itself; then, the semantics of the Ref formula compensate for the weakness of the system, since now the Ref formulas are about the proof system itself.

Theorem 9.10. *The following hold for Resolution:*

(i) Resolution is automatable if, and only if, it is automatable on Ref *formulas;*

(ii) Resolution is weakly automatable if, and only if, it is weakly automatable on Ref *formulas.*

Proof.

(i) This first item is a somewhat trivial byproduct of the hardness of automating Resolution [AM20]: if Resolution is automatable on Ref formulas, this can be used to decide SAT and hence **P** = **NP**; but then every proof system, including Resolution, is automatable.

(ii) Suppose Resolution is weakly automatable on Ref formulas. This means there exists a proof system Q that simulates Resolution on Ref formulas and is automatable on Ref formulas. Now, suppose a CNF formula φ with n variables and m clauses has a Resolution refutation in size t. Then, run the automating algorithm for Q on formulas of the form $\textsc{Ref}_{p(n,m,s)}(\textsc{Ref}_s(\varphi))$ for $p(n,m,s)$ the polynomial from Theorem 9.7 for $s = 1$, $s = 2$, and so on, via dovetailing, until a refutation is found in Q. Note that, since there exists a refutation of φ in size t, that means that $\textsc{Ref}_t(\varphi)$ is satisfiable and hence $\textsc{Ref}_{p(n,m,t)}(\textsc{Ref}_t(\varphi))$ is unsatisfiable but easy to refute in Resolution. Hence, the algorithm automating Q on Ref formulas must find a refutation π in size $\text{poly}(n,m,t)$.

At this point we would like to turn this refutation π into a refutation of φ. We can turn π into a refutation in $\mathsf{EF} + \textsc{Refl}^Q$, which p-simulates EF and is closed under *modus ponens*. Since Theorem 9.7 applies to these stronger systems, we just described an algorithm that finds a refutation of φ in time $\text{poly}(n,m,t)$ in the system $\mathsf{EF} + \textsc{Refl}^Q \geq_{\text{p}} \text{Res}$, where t is the size of the shortest Resolution refutation. We can conclude that Resolution is weakly automatable. □

By the classical result showing that Resolution is weakly automatable if and only if $\text{Res}(k)$ is weakly automatable for every constant $k \geq 1$ [AB04, Theorem 8], we have the following corollary.

Corollary 9.11. *For every $k \geq 1$, $\text{Res}(k)$ is weakly automatable if, and only if, $\text{Res}(k)$ is weakly automatable on* Ref *formulas.*

9.3 Unprovability of Resolution lower bounds

The power of Theorem 9.7 can be applied to study the provability of Resolution lower bounds. In what follows, we will say that a formula $\neg\text{Ref}_s(\varphi)$ constitutes a *true Resolution lower bound* if φ is unsatisfiable and does not have Resolution refutations of size s. Could there be a proof system that is polynomially bounded on all such formulas? And, in the first-order setting, is there perhaps a theory of arithmetic capable of proving all true Resolution lower bounds?

9.3.1 Propositional unprovability

We show that this is not the case: assuming S is strong enough and not polynomially bounded, there will be true Resolution lower bounds that S cannot derive in polynomial size.

We first exemplify this by giving explicit exponential lower bounds on Ref formulas for bounded-depth Frege. The following is a formal restatement of Corollary 1.10.

Theorem 9.12 (Hard Ref formulas for bounded-depth Frege)**.** *For every $d \leq O(\log n/\log\log n)$, there are polynomials $p(n)$ and $q(n)$ such that the formulas in the sequence $\{\neg\text{Ref}_{p(n)}(\text{Ref}_{q(n)}(\text{PHP}_n))\}_{n\in\mathbb{N}}$ are all tautological but require size $\exp\big(\Omega(n^{1/(2d+1)})\big)$ to be proven in depth-d Frege systems.*

Proof. The formula PHP_n has $N = \Theta(n^2)$ variables and $\Theta(n^3)$ clauses. Consider first the formulas $\text{Ref}_{q(n)}(\text{PHP}_n)$. Here, to apply the original bound on Ref formulas (Theorem 1.5), $q(n)$ has to be at least a square in the number of variables of PHP_n, so we define $q(n) := N^2 = \Theta(n^4)$. Let the size parameter $p(n)$ of the outer Ref formula be the polynomial p from Lemma 9.6 applied to N and $q(n)$. That the formulas $\text{Ref}_{p(n)}(\text{Ref}_{q(n)}(\text{PHP}_n))$ are tautologies follows form the fact that PHP_n is unsatisfiable, meaning that the lower bound on Ref formulas guarantees that $\text{Ref}_{q(n)}(\text{PHP}_n)$ is exponentially hard for Resolution and hence the second nesting of Ref remains unsatisfiable. The statement then follows from Lemma 9.6: if depth-d Frege refutes $\text{Ref}_{p(n)}(\text{Ref}_{q(n)}(\text{PHP}_n))$ in size ℓ, then depth-d Frege refutes $\text{PHP}_n[2]$ in size $\text{poly}(n,\ell)$. Substituting these 2-extension axioms increases the depth of every line in the proof by at most one, giving a depth-$(d+1)$ Frege refutation of PHP_n in size $\text{poly}(n,\ell)$, but, by Theorem 2.1, depth-$(d+1)$ Frege requires size $\exp\big(\Omega(n^{1/(2d+1)})\big)$ to refute PHP_n. □

There is nothing special about the pigeonhole principle in the argument above, and we can generalize this to any reasonably strong system where some lower bound is known. The following is a formal restatement of Theorem 1.9.

Theorem 9.13 (Propositional unprovability of Resolution lower bounds)**.** *For every propositional proof system $S \geq$ Res closed under* modus ponens *and clause substitutions, if S is not polynomially bounded, there exists a family of unsatisfiable CNF formulas $\{\psi_n\}_{n\in\mathbb{N}}$ on $N = \text{poly}(n)$ variables and of size $|\psi_n| = \text{poly}(n)$ such that*

(i) they require size at least $2^{N^{\Omega(1)}}$ to be refuted in Resolution;

(ii) there is a polynomial $p(n)$ such that the formulas $\{\neg\text{Ref}_{p(n)}(\psi_n)\}_{n\in\mathbb{N}}$ are tautological but do not have polynomial-size proofs in S.

Proof. Since S is not polynomially bounded, there exists a sequence of unsatisfiable 3-CNF formulas $\{\varphi_n\}_{n\in\mathbb{N}}$, each over n variables, that cannot be refuted by S in polynomial size. This means that the formulas of the form $\text{Ref}_{n^2}(\varphi_n)$ are unsatisfiable, or else the satisfying assignment would be a correct Resolution refutation which could be turned into a correct refutation in S. By the lower bound on Ref formulas (Theorem 1.5), $\text{Ref}_{n^2}(\varphi_n)$ requires size $2^{\Omega(n)}$ to be refuted in Resolution, meaning that $\text{Ref}_{p(n)}(\text{Ref}_{n^2}(\varphi_n))$ is unsatisfiable for every polynomial $p(n)$. If we choose p to be the polynomial in Lemma 9.6, it now follows from this same

lemma that S cannot possibly have polynomial-size proofs of the family $\{\neg\textsc{Ref}_{p(n)}(\textsc{Ref}_{n^2}(\varphi_n))\}_{n\in\mathbb{N}}$, or else there would be polynomial-size S-proofs of $\{\neg\varphi_n[2]\}_{n\in\mathbb{N}}$. Since S is closed under clause substitutions, we could substitute the 2-extensions of $\neg\varphi_n[2]$ and get polynomial-size proofs of $\{\neg\varphi_n\}_{n\in\mathbb{N}}$, a contradiction.

The formula family $\{\psi_n\}_{n\in\mathbb{N}}$ in the statement is then precisely given by $\psi_n \coloneqq \textsc{Ref}_{n^2}(\varphi_n)$. By Remark 2.6 the formula ψ_n has $N \coloneqq \mathrm{poly}(n)$ variables and polynomial size, which yields the rest of the parameters in the statement. □

The informal statement in Theorem 1.9 is obtained in particular by taking $p(n)$ to be N^2.

Remark 9.14. Since every proof system S can be p-simulated by $\mathsf{EF} + \textsc{Refl}^S$, and the latter is closed under *modus ponens* and clause substituions, that means that the previous statement applies to every proof system S. This implies that, unless **NP** = **coNP**, no proof systems can efficiently derive all true Resolution lower bounds. We note, however, that this follows already from the classical work of Iwama [Iwa97], who showed that the Proof Size Problem for Resolution ($\mathrm{PSP}_{\mathrm{Res}}$) is **NP**-complete. His many-one reduction mapped a 3-CNF formula φ over n variables to a new formula Φ over $O(n\log n)$ variables, such that φ is satisfiable if and only if Φ has a Resolution refutation of size $S(n) = O(n^3\log n)$. In the case when φ is unsatisfiable, however, Iwama only proves a lower bound of $S(n) + g(n)$, for some function $g(n) = O(S(n)/n)$. This suffices for his purposes, but it is certainly not enough to get the superpolynomial gap needed for hardness of automatability obtained in [AM20]. In our statement the parameters achieve almost optimal range: the formulas are exponentially hard, but a slightly subquadratic lower bound is not efficiently provable. ■

9.3.2 First-order unprovability of Resolution lower bounds

Our main result here is that a first-order version of the propositional unprovability results above hold unconditionally in the context of strong enough theories of arithmetic. The formal statement and proof follow.

Theorem 9.15. *Let T be a consistent first-order extending S^1_2 with a set of axioms recognizable in polynomial time and admitting $\mathbb{N}$ as a model. Then, there exits a sequence of unsatisfiable propositional formulas $\{\psi_{k,s}\}_{k,s\in\mathbb{N}}$ described uniformly by a polynomial-time algorithm given k and s in unary, where $\psi_{k,s}$ has $N = \mathrm{poly}(k,s)$ variables, and such that*

(i) Resolution refutations of the formula $\psi_{k,s}$ require size $2^{N^{\Omega(1)}}$;

(ii) there are constants $c > 0$ and $N_0 \in \mathbb{N}$ such that the lower bound expressed by the first-order sentence $\forall k\forall s\forall\pi\left(N > N_0 \wedge \mathrm{Ref}_{\mathrm{Res}}(\psi_{k,s},\pi) \rightarrow |\pi| > N^c\right)$ is independent of T.

Proof. Given the theory T, consider the strong proof system of T, denoted P_T, as defined in Section 2.3.5. Since T is polynomial-time axiomatizable, the system P_T is a well-defined Cook-Reckhow proof system. For convenience, we see P_T as a refutation system, and we consider the propositional formulas encoding its reflection principle $\textsc{Refl}^{P_T}_{k,s} \coloneqq \textsc{Ref}^{P_T}_s(\varphi,\tau) \wedge \textsc{Sat}_k(\varphi,\alpha)$, as defined in Section 2.4.1, where k denotes the number of variables of φ and s is the size of the proofs considered. We assume for convenience that φ is always a 3-CNF formula and has hence $O(k^3)$ clauses. Let us also assume that the formula $\textsc{Refl}^{P_T}_{k,s}$ itself is written as a 3-CNF formula and consists of $n \coloneqq \mathrm{poly}(k,s)$ variables. Consider now the propositional formula $\psi_{k,s} \coloneqq \textsc{Ref}_{n^2}(\textsc{Refl}^{P_T}_{k,s})$ stating that there is a size-n^2 Resolution refutation of $\textsc{Refl}^{P_T}_{k,s}$. It is easy to verify following Remark 2.6 that $\psi_{k,s}$ has $N \coloneqq \mathrm{poly}(n)$ variables and size $\mathrm{poly}(n)$.

Observe, first, that the sequence $\{\psi_{k,s}\}_{k,s\in\mathbb{N}}$ can be generated uniformly in polynomial time given k and s in unary. Next, note that all the formulas $\psi_{k,s}$ are unsatisfiable: otherwise, there would be size-n^2 Resolution refutations of the reflection principle of P_T, and this would amount to Resolution polynomially simulating P_T. In more detail, suppose F is a 3-CNF formula over k variables with a size-s refutation in P_T. If Resolution can refute $\textsc{Refl}^{P_T}_{k,s}(\varphi,\tau,\alpha)$ in size n^2, then by restricting φ by F and τ by the size-s P_T-refutation

in question, we get a size-n^2 refutation of $\textsc{Sat}(F, \alpha)$. Resolution can then refute F in its usual "native" encoding via Proposition 2.9, so we get a Resolution refutation of F in size $n^{O(1)} = \text{poly}(k, s)$, concluding $\text{Res} \geq P_T$. This is a contradiction, because since $T \supseteq \text{S}^1_2$ we know that P_T simulates Extended Frege [Pud20, Section 4.2, Fact 2], and it is well-known that Resolution is strictly weaker than EF.

Finally, because all of the formulas $\textsc{Refl}^{P_T}_{k,s}$ are unsatisfiable, the lower bound on $\textsc{Ref}$ formulas (Theorem 1.5) guarantees that for large enough $n = \text{poly}(k, s)$, the formula $\psi_{k,s} = \textsc{Ref}_{n^2}(\textsc{Refl}^{P_T}_{k,s})$ requires size $2^{\Omega(n^2/n)} = 2^{\Omega(n)}$ to be refuted in Resolution. Since $\psi_{k,s}$ has $N = \text{poly}(n)$ variables, the size lower bound in terms of N is $2^{N^{\Omega(1)}}$. In particular, that means that for every $c > 0$, there is some $N_0 \in \mathbb{N}$ such that the much weaker lower bound statement $\forall k \forall s \forall \pi \left(N > N_0 \wedge |\pi| \leq N^c \rightarrow \neg \text{Ref}_{\text{Res}}(\psi_{k,s}, \pi)\right)$ holds in the standard model $\mathbb{N}$, which is a model of T. Thus the lower bound is consistent with T.

We now provide a model where the sentence fails. By Gödel's second incompleteness theorem, T cannot prove the soundness of P_T [Pud20, Section 4.2, Fact 3]. Thus, there is a model M of T where the reflection principle of P_T fails. We claim that there is $c > 0$ and a suitable $N_0 \in \mathbb{N}$ such that the sentence $\forall k \forall s \forall \pi \left(N > N_0 \wedge |\pi| \leq N^c \rightarrow \neg \text{Ref}_{\text{Res}}(\psi_{k,s}, \pi)\right)$ fails in this model as well. Indeed, since the reflection of P_T fails in M, there exist some nonstandard 3-CNF formula $\varphi_0 \in M \setminus \mathbb{N}$, a P_T-refutation $\tau_0 \in M \setminus \mathbb{N}$ and $\alpha_0 \in M \setminus \mathbb{N}$ such that $M \models \text{Ref}_{P_T}(\varphi_0, \tau_0) \wedge \text{Sat}(\varphi_0, \alpha_0)$. This formula φ_0 has some nonstandard number of variables $k_0 \in M \setminus \mathbb{N}$, and similarly τ_0 has some length $s_0 := |\tau_0| \in M \setminus \mathbb{N}$. We then have that in the model M, the formula $\textsc{Refl}^{P_T}_{k_0,s_0}(\varphi, \tau, \alpha)$ is satisfiable.

Now, since M is a model of T and T extends S^1_2, by Theorem 7.4 we have that the upper bound on $\textsc{Ref}$ formulas holds in M. That is,

$$M \models \forall \varphi \forall n \forall m \forall \alpha \forall t \Big(\text{CNF}(\varphi, n, m) \wedge \text{Sat}(\varphi, \alpha) \rightarrow \exists \pi \left(|\pi| \leq c_P \cdot t(m + tn^2) \wedge \text{Ref}_{\text{Res}}(\textsc{Ref}_t(\varphi), \pi)\right)\Big), \tag{9.3}$$

where c_P is the constant in the Big-Theta term $\Theta(t(m+tn^2))$ bounding the size of upper bound construction in Theorem 7.4. (Technically speaking Theorem 7.4 proved the previous sentence in Resolution, but a simple inspection of the proof reveals that the construction is perfectly uniform and can be immediately formalized in S^1_2 —and, for that matter, even in much weaker theories.)

Hence, substituting φ for the 3-CNF formula $\textsc{Refl}^{P_T}_{k_0,s_0}$ on n variables and $m \leq 8n^3$ clauses, and t for n^2, we get

$$M \models \exists \pi \left(|\pi| \leq c_P \cdot (8n^5 + n^6) \wedge \text{Ref}_{\text{Res}}(\psi_{k_0,s_0}, \pi)\right) . \tag{9.4}$$

Since the number N of variables of ψ_{k_0,s_0} is $N = n^{\Theta(1)}$, the bound on $|\pi|$ is $c_P \cdot (8n^5 + n^6) \leq N^{O(1)}$ for large enough N, it follows that there is $c > 0$ and $N_0 \in \mathbb{N}$ such that the first-order sentence $\forall k \forall s \forall \pi \left(N > N_0 \wedge |\pi| \leq N^c \rightarrow \neg \text{Ref}_{\text{Res}}(\psi_{k,s}, \pi)\right)$ fails in M. □

With a couple of extra tricks, we can extend the result to any theory T containing just Robinson Arithmetic rather than all of S^1_2. We remark as well that there is nothing essential about T being polynomial-time axiomatizable, and one could generalize this to any recursively enumerable extension of Q via a padding trick [Cra53]. The following is a formal version of Theorem 1.11.

Corollary 9.16. *Let T be a consistent first-order extending* Q *with a set of axioms recognizable in polynomial time. Then, there exits a sequence of unsatisfiable propositional formulas $\{\psi_{k,s}\}_{k,s \in \mathbb{N}}$ described uniformly by a polynomial-time algorithm given k and s in unary, where $\psi_{k,s}$ has $N = \text{poly}(k, s)$ variables, and such that*

(i) Resolution refutations of the formula $\psi_{k,s}$ require size $2^{N^{\Omega(1)}}$;

(ii) there is a constant $c > 0$ and $N_0 \in \mathbb{N}$ such that the lower bound expressed by the first-order sentence $\forall k \forall s \forall \pi \left(N > N_0 \wedge \text{Ref}_{\text{Res}}(\psi_{k,s}, \pi) \rightarrow |\pi| > N^c\right)$ is unprovable in T, but true in $\mathbb{N}$.

Proof. Define $T' := \mathrm{S}^1_2 + \Pi_1\text{-Refl}_T$. Observe that T' satisfies all the conditions of Theorem 9.15. First, the theory is consistent because $\mathbb{N} \models T'$. This is because $\mathbb{N} \models \mathrm{S}^1_2$ and, as T is consistent and extends Q, which refutes every false Π_1-sentence, $\Pi_1\text{-Refl}_T$ is a true sentence. Second, T' is clearly recursively axiomatizable because S^1_2 is, and in particular the axioms can be recognized in polynomial time. Hence, Theorem 9.15 guarantees a sequence of unsatisfiable propositional formulas $\{\psi_{k,s}\}_{k,s\in\mathbb{N}}$ with the desired properties. In particular, the statement $\forall k \forall s \forall \pi \left(N > N_0 \wedge |\pi| \leq N^c \rightarrow \neg \mathrm{Ref}_{\mathrm{Res}}(\psi_{k,s}, \pi)\right)$ is unprovable in T', where c and N_0 are again given by Theorem 9.15. On the one hand, this means the sentence is not provable in T'. Thus it cannot be proven in T either, since, by construction, $T' = \mathrm{S}^1_2 + \Pi_1\text{-Refl}_T$ is Π_1-conservative over T and the sentence is a Π_1-sentence. On the other hand, as argued in the proof of Theorem 9.15, the lower bound statement is true in $\mathbb{N}$ by the lower bound on Ref formulas (Theorem 1.5). □

Acknowledgements

We are indebted to Ján Pich for insightful initial discussions on the problem and particularly for the idea behind the proof of Theorem 1.3, originally based on MCSP, as sketched in Section 1.2.2. We thank Emil Jeřábek [Jeř25] for the proof of Lemma 7.3 simulating contraposition in Resolution. We also thank Jonas Conneryd for useful comments and careful proofreading of a draft of this work, and Anupam Das, Stefan Grosser, Antonina Kolokolova, Jan Krajíček, Jakob Nordström, and Rahul Santhanam for helpful comments and suggestions.

This collaboration started at the Proof Complexity and Beyond workshop at Mathematisches Forschungsinstitut Oberwolfach, in March 2024. Part of this work was carried out during the Proof Complexity Workshop at the University of Oxford in September of 2024 and during the *Kaleidoscope de la complexité* spring school in April 2025 at the Centre International de Rencontres Mathématiques (CIRM) in Marseille, France. We also gratefully acknowledge that we have benefited greatly from being part of the Basic Algorithms Research Centre (BARC) environment financed by the Villum Investigator grant 54451.

Noel Arteche was supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Albert Atserias was supported by the Spanish Research Agency through grant PID2022-138506NB-C22 (PROOFS BEYOND) and the Severo Ochoa and María de Maeztu Program for Centers and Units of Excellence in R&D (CEX2020-001084-M). Susanna F. de Rezende received funding from the Knut and Alice Wallenberg grant KAW 2023.0116, ELLIIT, and the Swedish Research Council grant 2021-05104. Erfan Khaniki was supported by the Royal Society University Research Fellowship URF\R1\211106 “Proof complexity and circuit complexity: a unified approach”.

References

[AB04] A. Atserias and M. L. Bonet, “On the automatizability of resolution and related propositional proof systems,” *Inform. and Comput.*, vol. 189, no. 2, pp. 182–201, 2004. DOI: 10.1016/j.ic.2003.10.004.

[AB09] S. Arora and B. Barak, *Computational Complexity: A Modern Approach*. Cambridge University Press, 2009. DOI: a10.1017/CBO9780511804090.

[ACG24] N. Arteche, G. Carenini, and M. Gray, “Quantum automating $\mathbf{TC}^0$-Frege is LWE-hard,” in *39th Computational Complexity Conference, CCC 2024, July 22-25, 2024, Ann Arbor, MI, USA*, 2024, 15:1–15:25. DOI: 10.4230/LIPIcs.CCC.2024.15.

[AD08] A. Atserias and V. Dalmau, “A combinatorial characterization of resolution width,” *J. Comput. System Sci.*, vol. 74, no. 3, pp. 323–334, 2008. DOI: 10.1016/j.jcss.2007.06.025.

[Ajt94] M. Ajtai, "The complexity of the pigeonhole principle," *Combinatorica*, vol. 14, no. 4, pp. 417–433, 1994. DOI: 10.1007/BF01302964.

[AKPS24] N. Arteche, E. Khaniki, J. Pich, and R. Santhanam, "From proof complexity to circuit complexity via interactive protocols," in *51st International Colloquium on Automata, Languages, and Programming*, 2024, Art. No. 12, 20. DOI: 10.4230/lipics.icalp.2024.12.

[All25] E. Allender, **P**-*uniform vs.* **DLOGTIME**-*uniform* $\mathbf{AC}^0$, Theoretical Computer Science Stack Exchange, 2025. [Online]. Available: https://cstheory.stackexchange.com/q/55461.

[ALWZ21] R. Alweiss, S. Lovett, K. Wu, and J. Zhang, "Improved bounds for the sunflower lemma," *Annals of Mathematics*, vol. 194, no. 3, pp. 795–815, 2021. DOI: 10.4007/annals.2021.194.3.5.

[AM11] A. Atserias and E. Maneva, "Mean-payoff games and propositional proofs," *Inform. and Comput.*, vol. 209, no. 4, pp. 664–691, 2011. DOI: 10.1016/j.ic.2011.01.003.

[AM20] A. Atserias and M. Müller, "Automating resolution is **NP**-hard," *Journal of the ACM (JACM)*, vol. 67, no. 5, pp. 1–17, 2020. DOI: 10.1145/3409472.

[AR08] M. Alekhnovich and A. A. Razborov, "Resolution is not automatizable unless **W**[P] is tractable," *SIAM Journal on Computing*, vol. 38, no. 4, pp. 1347–1363, 2008. DOI: doi:10.1137/06066850X.

[AT24] A. Atserias and I. Tzameret, "Feasibly constructive proof of Schwartz-Zippel lemma and the complexity of finding hitting sets," *Electronic Colloquium on Computational Complexity (ECCC)*, no. TR24-174, 2024. [Online]. Available: https://eccc.weizmann.ac.il/report/2024/174/.

[Ats13] A. Atserias, "The proof-search problem between bounded-width resolution and bounded-degree semi-algebraic proofs," in *Theory and applications of satisfiability testing—SAT 2013*, ser. Lecture Notes in Comput. Sci. Vol. 7962, 2013, pp. 1–17. DOI: 10.1007/978-3-642-39071-5_1.

[BDG+04] M. L. Bonet, C. Domingo, R. Gavaldà, A. Maciel, and T. Pitassi, "Non-automatizability of bounded-depth Frege proofs," *computational complexity*, vol. 13, pp. 47–68, 2004. DOI: 10.1007/s00037-004-0183-5.

[Bel20] Z. Bell, "Automating regular or ordered resolution is **NP**-hard," *Electronic Colloquium on Computational Complexity (ECCC)*, no. TR20-105, 2020. [Online]. Available: https://eccc.weizmann.ac.il/report/2020/105.

[BP96] P. Beame and T. Pitassi, "Simplified and improved resolution lower bounds," in *Proceedings of 37th Conference on Foundations of Computer Science*, 1996, pp. 274–282. DOI: 10.1109/SFCS.1996.548486.

[BPR00] M. L. Bonet, T. Pitassi, and R. Raz, "On interpolation and automatization for Frege systems," *SIAM J. Comput.*, vol. 29, no. 6, pp. 1939–1967, 2000. DOI: 10.1137/S0097539798353230.

[BPT14] A. Beckmann, P. Pudlák, and N. Thapen, "Parity games and propositional proofs," *ACM Trans. Comput. Logic*, vol. 15, no. 2, May 2014. DOI: 10.1145/2579822.

[Bus86] S. R. Buss, *Bounded arithmetic*. Bibliopolis, Naples, 1986.

[Bus97] S. R. Buss, "Bounded arithmetic and propositional proof complexity," in *Logic of computation (Marktoberdorf, 1995)*, ser. NATO Adv. Sci. Inst. Ser. F: Comput. Systems Sci. Vol. 157, Springer, Berlin, 1997, pp. 67–121.

[Bus98] S. R. Buss, "First-order proof theory of arithmetic," in *Handbook of proof theory*, ser. Stud. Logic Found. Math. Vol. 137, North-Holland, Amsterdam, 1998, pp. 79–147. DOI: 10.1016/S0049-237X(98)80017-7.

[BW01] E. Ben-Sasson and A. Wigderson, "Short proofs are narrow—resolution made simple," *J. ACM*, vol. 48, no. 2, pp. 149–169, 2001. DOI: 10.1145/375827.375835.

[BY24] S. Buss and E. Yolcu, "Regular resolution effectively simulates resolution," *Inform. Process. Lett.*, vol. 186, Paper No. 106489, 4, 2024. DOI: 10.1016/j.ipl.2024.106489.

[CEI96] M. Clegg, J. Edmonds, and R. Impagliazzo, "Using the Groebner basis algorithm to find proofs of unsatisfiability," in *Proceedings of the Twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996)*, ACM, New York, 1996, pp. 174–183. DOI: 10.1145/237814.237860.

[Cob65] A. Cobham, "The intrinsic computational difficulty of functions," in *Logic, Methodology and Philos. Sci. (Proc. 1964 Internat. Congr.)*, North-Holland, Amsterdam, 1965, pp. 24–30.

[Coo75] S. A. Cook, "Feasibly constructive proofs and the propositional calculus (preliminary version)," in *Seventh Annual ACM Symposium on Theory of Computing (Albuquerque, N.M., 1975)*, Association for Computing Machinery, New York, 1975, pp. 83–97.

[CP90] S. Cook and T. Pitassi, "A feasibly constructive lower bound for resolution proofs," *Information Processing Letters*, vol. 34, no. 2, pp. 81–85, 1990. DOI: 10.1016/0020-0190(90)90141-J.

[CR79] S. A. Cook and R. A. Reckhow, "The relative efficiency of propositional proof systems," *J. Symbolic Logic*, vol. 44, no. 1, pp. 36–50, 1979. DOI: 10.2307/2273702.

[Cra53] W. Craig, "On axiomatizability within a system," *J. Symbolic Logic*, vol. 18, pp. 30–32, 1953, ISSN: 0022-4812. DOI: 10.2307/2266324.

[dRGN+21] S. F. de Rezende, M. Göös, J. Nordström, T. Pitassi, R. Robere, and D. Sokolov, "Automating algebraic proof systems is **NP**-hard," in *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, 2021, pp. 209–222. DOI: 10.1145/3406325.3451080.

[FFNR03] S. A. Fenner, L. Fortnow, A. V. Naik, and J. D. Rogers, "Inverting onto functions," *Inform. and Comput.*, vol. 186, no. 1, pp. 90–103, 2003. DOI: 10.1016/S0890-5401(03)00119-6.

[Gar19] M. Garlík, "Resolution lower bounds for refutation statements," in *44th International Symposium on Mathematical Foundations of Computer Science*, 2019, Art. No. 37, 13. DOI: 10.4230/LIPIcs.MFCS.2019.37.

[Gar24] M. Garlík, "Failure of feasible disjunction property for k-DNF resolution and **NP**-hardness of automating it," in *39th Computational Complexity Conference, CCC 2024, July 22-25, 2024, Ann Arbor, MI, USA*, vol. 300, 2024, 33:1–33:23. DOI: 10.4230/LIPIcs.CCC.2024.33.

[Gay21] A. Gaysin, "H-coloring dichotomy in proof complexity," *Journal of Logic and Computation*, vol. 31, no. 5, pp. 1206–1225, Apr. 2021. DOI: https://doi.org/10.1093/logcom/exab028.

[Gay23] A. Gaysin, *Proof complexity of CSP*, 2023. arXiv: 2201.00913.

[Gay24] A. Gaysin, *Proof complexity of universal algebra in a CSP dichotomy proof*, 2024. arXiv: 2403.06704.

[GJ79] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of* **NP***-Completeness*. W. H. Freeman, 1979.

[GKMP20] M. Göös, S. Koroth, I. Mertz, and T. Pitassi, "Automating cutting planes is **NP**-hard," in *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, 2020, pp. 68–77. DOI: 10.1145/3357713.3384248.

[GL10] N. Galesi and M. Lauria, "On the automatizability of polynomial calculus," *Theor. Comp. Sys.*, vol. 47, no. 2, pp. 491–506, Aug. 2010. DOI: 10.1007/s00224-009-9195-5.

[Hås23] J. Håstad, "On small-depth Frege proofs for PHP," in *2023 IEEE 64th Annual Symposium on Foundations of Computer Science—FOCS 2023*, IEEE Computer Soc., Los Alamitos, CA, 2023, pp. 37–49. DOI: 10.1109/FOCS57990.2023.00010.

[HP11] L. Huang and T. Pitassi, "Automatizability and simple stochastic games," in *Automata, languages and programming. Part I*, ser. Lecture Notes in Comput. Sci. Vol. 6755, Springer, Heidelberg, 2011, pp. 605–617. DOI: 10.1007/978-3-642-22006-7_51.

[HP93] P. Hájek and P. Pudlák, *Metamathematics of first-order arithmetic*. Springer-Verlag, Berlin, 1993. DOI: 10.1007/978-3-662-22156-3.

[IR22] D. Itsykson and A. Riazanov, "Automating OBDD proofs is **NP**-hard," in *47th International Symposium on Mathematical Foundations of Computer Science (MFCS 2022)*, vol. 241, 2022, 59:1–59:15. DOI: 10.4230/LIPIcs.MFCS.2022.59.

[Iwa97] K. Iwama, "Complexity of finding short resolution proofs," in *Mathematical foundations of computer science 1997 (Bratislava)*, ser. Lecture Notes in Comput. Sci. Vol. 1295, Springer, Berlin, 1997, pp. 309–318. DOI: 10.1007/BFb0029974.

[Jeř05] E. Jeřábek, "Weak pigeonhole principle, and randomized computation," Ph.D. dissertation, Faculty of Mathematics and Physics, Charles University, Prague, 2005.

[Jeř25] E. Jeřábek, *Modus ponens style inferences in Resolution*, Theoretical Computer Science Stack Exchange, 2025. [Online]. Available: https://cstheory.stackexchange.com/q/55062.

[Kha22a] E. Khaniki, "New relations and separations of conjectures about incompleteness in the finite domain," *Journal of Symbolic Logic*, vol. 87, no. 3, pp. 912–937, 2022. DOI: 10.1017/jsl.2021.99.

[Kha22b] E. Khaniki, "Nisan-Wigderson Generators in Proof Complexity: New Lower Bounds," in *37th Computational Complexity Conference (CCC 2022)*, ser. Leibniz International Proceedings in Informatics (LIPIcs), vol. 234, 2022, 17:1–17:15. DOI: 10.4230/LIPIcs.CCC.2022.17.

[Kha24] E. Khaniki, "Jump operators, interactive proofs and proof complexity generators," in *2024 IEEE 65th Annual Symposium on Foundations of Computer Science (FOCS)*, IEEE, 2024, pp. 573–593.

[KM00] J. Köbler and J. Messner, "Is the standard proof system for SAT p-optimal?" In *International Conference on Foundations of Software Technology and Theoretical Computer Science*, Springer, 2000, pp. 361–372. DOI: 10.1007/3-540-44450-5_29.

[KP98] J. Krajíček and P. Pudlák, "Some consequences of cryptographical conjectures for S^1_2 and EF," *Information and Computation*, vol. 140, no. 1, pp. 82–94, 1998. DOI: 10.1006/inco.1997.2674.

[KPT91] J. Krajíček, P. Pudlák, and G. Takeuti, "Bounded arithmetic and the polynomial hierarchy," *Annals of Pure and Applied Logic*, vol. 52, no. 1, pp. 143–153, 1991. DOI: https://doi.org/10.1016/0168-0072(91)90043-L.

[KPW95] J. Krajíček, P. Pudlák, and A. Woods, "An exponential lower bound to the size of bounded depth Frege proofs of the pigeonhole principle," *Random Structures Algorithms*, vol. 7, no. 1, pp. 15–39, 1995. DOI: 10.1002/rsa.3240070103.

[Kra01] J. Krajíček, "On the weak pigeonhole principle," *Fundamenta Mathematicae*, vol. 170, no. 1-2, pp. 123–140, 2001. DOI: 10.4064/fm170-1-8.

[Kra19] J. Krajíček, *Proof complexity*. Cambridge University Press, Cambridge, 2019. DOI: 10.1017/9781108242066.

[Kra22] J. Krajíček, "Information in propositional proofs and algorithmic proof search," *J. Symb. Log.*, vol. 87, no. 2, pp. 852–869, 2022. DOI: 10.1017/jsl.2021.75.

[Kra23] J. Krajíček, "A proof complexity conjecture and the incompleteness theorem," *The Journal of Symbolic Logic*, pp. 1–5, 2023. DOI: 10.1017/jsl.2023.69.

[Kra24] J. Krajíček, "Extended Nullstellensatz proof systems," *Proceedings of the American Mathematical Society*, vol. 152, pp. 4881–4892, 2024. DOI: doi.org/10.1090/proc/16709.

[Kra25] J. Krajíček, *Proof Complexity Generators*. Cambridge University Press, Cambridge, 2025. DOI: https://doi.org/10.1017/9781009611664.

[Kra95] J. Krajíček, *Bounded arithmetic, propositional logic, and complexity theory*. Cambridge University Press, Cambridge, 1995. DOI: 10.1017/CBO9780511529948.

[LLR24] J. Li, Y. Li, and H. Ren, "Metamathematics of Resolution lower bounds: A TFNP perspective," *Electronic Colloquium on Computational Complexity (ECCC)*, no. TR24-190, 2024. [Online]. Available: https://eccc.weizmann.ac.il/report/2024/190/.

[MP20] M. Müller and J. Pich, "Feasibly constructive proofs of succinct weak circuit lower bounds," *Annals of Pure and Applied Logic*, vol. 171, no. 2, p. 102 735, 2020. DOI: 10.1016/j.apal.2019.102735.

[MP24] N. Mazor and R. Pass, "Gap MCSP Is Not (Levin) **NP**-Complete in Obfustopia," in *39th Computational Complexity Conference (CCC 2024)*, ser. Leibniz International Proceedings in Informatics (LIPIcs), vol. 300, 2024, 36:1–36:21. DOI: 10.4230/LIPIcs.CCC.2024.36.

[MPW19] I. Mertz, T. Pitassi, and Y. Wei, "Short proofs are hard to find," in *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)*, vol. 132, 2019, 84:1–84:16. DOI: 10.4230/LIPIcs.ICALP.2019.84.

[Mül21] M. Müller, "Typical forcings, **NP** search problems and an extension of a theorem of Riis," *Annals of Pure and Applied Logic*, vol. 172, no. 4, p. 102 930, 2021. DOI: doi.org/10.1016/j.apal.2020.102930.

[Nar22] M. Narusevych, *Models of bounded arithmetic and variants of pigeonhole principle*, 2022. arXiv: 2208.14713.

[Nar24] M. Narusevych, *An independence of the MIN principle from the PHP principle*, 2024. arXiv: 2406.14930.

[Oli25] I. C. Oliveira, "Meta-mathematics of computational complexity theory," *Electronic Colloquium on Computational Complexity (ECCC)*, no. TR25-041, 2025. [Online]. Available: https://eccc.weizmann.ac.il/report/2025/041/.

[Pap24] T. Papamakarios, "Depth-d Frege systems are not automatable unless $\mathbf{P} = \mathbf{NP}$," in *39th Computational Complexity Conference (CCC 2024)*, vol. 300, 2024, 22:1–22:17. DOI: 10.4230/LIPIcs.CCC.2024.22.

[PBI93] T. Pitassi, P. Beame, and R. Impagliazzo, "Exponential lower bounds for the pigeonhole principle," *Computational Complexity*, vol. 3, no. 2, pp. 97–140, 1993. DOI: 10.1007/BF01200117.

[PP24] J. Park and H. T. Pham, "A proof of the Kahn–Kalai conjecture," *J. Amer. Math. Soc.*, vol. 37, pp. 235–243, 2024. DOI: 10.1090/jams/1028.

[PS22] J. Pich and R. Santhanam, "Learning algorithms versus automatability of Frege systems," in *49th International Colloquium on Automata, Languages, and Programming (ICALP 2022)*, vol. 229, 2022, 101:1–101:20. DOI: 10.4230/LIPIcs.ICALP.2022.101.

[PS23] J. Pich and R. Santhanam, *Towards* $\mathbf{P} \neq \mathbf{NP}$ *from Extended Frege lower bounds*, 2023. arXiv: 2312.08163.

[Pud03] P. Pudlák, "On reducibility and symmetry of disjoint **NP** pairs," *Theoretical Computer Science*, vol. 295, no. 1-3, pp. 323–339, 2003. DOI: https://doi.org/10.1016/S0304-3975(02)00411-5.

[Pud20] P. Pudlák, *Reflection principles, propositional proof systems, and theories*, 2020. arXiv: 2007.14835.

[Raz04] A. A. Razborov, "Resolution lower bounds for perfect matching principles," *J. Comput. System Sci.*, vol. 69, no. 1, pp. 3–27, 2004. DOI: 10.1016/j.jcss.2004.01.004.

[Raz95] A. A. Razborov, "Bounded arithmetic and lower bounds in Boolean complexity," in *Feasible mathematics, II (Ithaca, NY, 1992)*, ser. Progr. Comput. Sci. Appl. Logic, vol. 13, Birkhäuser Boston, Boston, MA, 1995, pp. 344–386.

[ST21] R. Santhanam and I. Tzameret, "Iterated lower bound formulas: A diagonalization-based approach to proof complexity," in *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, 2021, pp. 234–247. DOI: https://dl.acm.org/doi/10.1145/3406325.3451010.

[Tha16] N. Thapen, "A tradeoff between length and width in resolution," *Theory Comput.*, vol. 12, Paper No. 5, 14, 2016. DOI: 10.4086/toc.2016.v012a005.

A Exponential approximation in PV_1

Müller and Pich [MP20] formalized in PV_1 the following useful bound.

Proposition A.1 ([MP20, Proposition 2.5])**.** *Provably in* PV_1*, for all* $n, m \in \mathrm{Log}$ *such that* $n < m$*, it holds that* $1 - n/m \leq 2^{-n/m}$.

For our application in Lemma 6.3 we need to be able to take powers of this on both sides.

Lemma A.2. *In* PV_1*, for all* $n, m, p \in \mathrm{Log}$*, if* $n < m$ *and* $p \geq 1$*, it holds that*

$$\left(1 - \frac{n}{m}\right)^p \leq 2^{-\frac{n}{m}p}.$$

Proof. Since $p \in \mathrm{Log}$, we take n and m to be fixed and proceed by Length Induction on p. The base case $p = 1$ if covered by Proposition A.1. If the bound holds for p, then for $p + 1$ we have

$$\left(1 - \frac{n}{m}\right)^{p+1} = \left(1 - \frac{n}{m}\right)^p \cdot \left(1 - \frac{n}{m}\right) \leq 2^{-\frac{n}{m}p} \cdot 2^{-\frac{n}{m}} = 2^{-\frac{n}{m}(p+1)}. \tag{A.1}$$

The first equality holds because $(1 - n/m) \in \mathrm{Log}$ and PV_1 can prove the basic properties of powers for Log-sized objects; the next inequality holds by induction hypothesis and Proposition A.1. This completes the proof. □

B Proof of Lemma 5.3

Lemma 5.3. *There exists a* PV *function* f *such that* S^1_2 *proves the statement that for every 3-CNF formula* φ *with* n *variables and* m *clauses,* $f(\varphi)$ *outputs a graph* $G_\varphi = (V, E)$ *with* $m \cdot n$ *nodes and such that* φ *is satisfiable if and only if* G_φ *has a vertex cover of size* $m \cdot (n - 1)$.

Proof. Without loss of generality, let us assume every clause C_i in φ has three different literals and is of the form $\ell_{i,1} \vee \ell_{i,2} \vee \ell_{i,3}$. Build the formula

$$\varphi' \coloneqq \bigwedge_{i=1}^{m} \left(\ell_{i,2} \vee \ell_{i,3} \vee \bigvee_{j=1}^{n-2} \ell_{i,1}\right). \tag{B.1}$$

The formula φ' is just φ except every clause has been weakened with $n - 2$ copies of $\ell_{i,1}$ so that every clause has width n. It is clear that φ and φ' have the same satisfying assignments.

We construct first the usual reduction from SAT to Clique. We craft a graph K consisting of one node per literal in the formula. This means there are mn nodes, and two nodes are connected if and only if the literals come from different clauses and one is not the negation of the other.

See, first, that the graph can be partitioned into m sets $S_1, \ldots, S_m$, each with n nodes corresponding to the n literals of each clause. Within each set the nodes are disconnected, so each of these sets constitutes an independent set, and φ is satisfiable if and only if K has a clique of size m. If φ is satisfiable by assignment α, then so is φ' and we can find a clique in K as follows. From every clause C_i in φ', take the first literal satisfied by the assignment α. By construction, the nodes corresponding to these literals are all adjacent to each other, so they form a clique of size m. On the other hand, if K has a clique of size m, then each node in the clique belongs to one and only of the sets S_i. Furthermore, the literals underlying these nodes are compatible: none is the negation of any other. Consider the partial assignment defined by the underlying literals, and fill the rest of the assignment arbitrarily. Again, by construction, this forms a satisfying assignment to φ' and hence to φ.

Now, to go from Clique to Vertex Cover, we reprove in S_2^1 the standard fact that a graph $G = (V, E)$ on n nodes has a clique of size k if and only if $\overline{G} = (V, \overline{E})$ has a vertex cover of size $n - k$, where $\overline{E}$ is the set of pairs of vertices that are not in E. If C is a clique of size k in G, then, by definition, C is an independent set in $\overline{G}$. Then, $V \setminus C$ is a vertex cover for $\overline{G}$; if not, then there is $(u, v) \in E$ such that $u \notin V \setminus C$ and $v \notin V \setminus C$. Again, by definition of set difference, $u \in C$ and $v \in C$, which contradicts C being an independent set. For the backwards direction, if $V \setminus C$ is a vertex cover in $\overline{G}$, then C is an independent set in $\overline{G}$, or else there would be an edge in $\overline{E}$ not covered by any vertex in $V \setminus C$. Then, C is a clique in G. This entire argument is a manipulation of basic set-theoretic definitions and does not require induction in S_2^1.

For our specific graph K obtained from the reduction to Clique, we have mn nodes and thus $\varphi \in$ 3SAT if, and only if, $\overline{G}$ has a vertex cover of size $mn - m = m(n - 1)$. The function f in the statement outputs the graph $G_\varphi := \overline{K}$, which can be constructed in time $\mathrm{poly}(|\varphi|)$. This completes the proof. □