Efficient Adversaries

Erfan Khaniki
University of Oxford, UK

Ján Pich
University of Oxford, UK

Dmitry Sokolov
Université de Montréal, Canada

Abstract

The size of Frege proofs can be characterized in terms of prover-adversary games of Pudlák and Buss. We consider a generalization of prover-adversary games to many standard proof systems and show that some of the major proof complexity lower bounds such as the constant-depth Frege lower bound for the pigeonhole principle based on the method of k-evaluations, the Resolution lower bound for the weak pigeonhole principle based on the method of pseudo-width and Razborov's $\mathsf{Res}(k)$ lower bound for Nisan-Wigderson generators based on expansion and a width lower bound (which is used to derive the $\mathsf{Res}(k)$-hardness of formulas expressing circuit lower bounds) are constructive in the sense that they yield efficient algorithms computing winning strategies of adversaries in the generalized games. This is in contrast with our second result saying that if (a) such a constructive lower bound exists for Extended Frege system EF for formulas expressing succinct circuit lower bounds for SAT and (b) EF is strong enough to prove efficiently the correctness of anticheckers for SAT, then it is easy to separate the canonical pair of EF.

2012 ACM Subject Classification Theory of computation → Proof complexity; Theory of computation → Circuit complexity; Theory of computation → Complexity theory and logic

Keywords and phrases proof complexity, circuit complexity, lower bounds, barriers, truth-table formula, pigeonhole principle

Digital Object Identifier 10.4230/LIPIcs.CCC.2026.8

Funding *Erfan Khaniki*: This work was supported by the Royal Society University Research Fellowship URF\R1\211106 "Proof complexity and circuit complexity: a unified approach".
Ján Pich: This work was supported by the Royal Society University Research Fellowship URF\R1\211106 "Proof complexity and circuit complexity: a unified approach".
Dmitry Sokolov: This work was supported by the Swiss State Secretariat for Education, Research, and Innovation (SERI) under contract number MB22.00026.

Acknowledgements We are indebted to Moritz Müller for a simplification of the proof of Theorem 1, which allowed us to derive Theorem 2 for EF instead of G_1^*, and for many comments that clarified the presentation of the paper. We thank Neil Thapen for a correction of Lemma 19, Noel Arteche for pointing out to us that Item 2 in Theorem 1 implies $\mathsf{NP} \not\subseteq \mathsf{Circuit}[n^\ell]$, and Pavel Pudlák, Jan Krajíček and anonymous reviewers for suggestions and comments. This project was initiated at the Institute of Mathematics of the Czech Academy of Sciences. For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

1 Introduction

Proving circuit lower bounds for explicit Boolean functions is a central open problem in complexity theory, underlying questions such as the P vs NP problem.

The notorious difficulty of the problem has led to the development of a theory of formal barrier results. For example, already in the 80's Razborov [57] showed that the approximation method, used to derive monotone circuit lower bounds [55] and lower bounds against $\mathsf{AC}^0[p]$

[56, 69], cannot yield superquadratic lower bounds for non-monotone circuits. In the 90s, Razborov and Rudich [65] discovered the influential natural proofs barrier clarifying the limitations of major lower bound methods such as the method of random restrictions or the algebraic method of Razborov and Smolensky [56, 69]. More recently, the locality barrier explained why a successful implementation of hardness magnification, an approach to strong explicit circuit lower bounds sidestepping the natural proofs barrier, requires more substantial new ideas [13]. Such barrier results could potentially guide us towards the solution of the original problem, but it is also possible that they will eventually lead to the conclusion that strong circuit lower bounds are practically impossible to establish or to the discovery of new surprising algorithms, cf. [12].

The possibility that circuit lower bounds could lie beyond what can be achieved within feasible logical reasoning was investigated extensively by Razborov who used natural proofs to derive the unprovability of circuit lower bounds in certain theories of bounded arithmetic [59]. Such unprovability results can be typically interpreted as lower bounds on lengths of proofs of tautologies expressing circuit lower bounds in propositional proof systems. Indeed, after composing [59] Razborov switched to the more elegant and transparent framework of propositional logic and formulated his subsequent results in this setting; including a degree lower bound for deriving circuit lower bounds in Polynomial Calculus [60], a Resolution lower bound for formulas expressing circuit lower bounds based on the pseudo-width method [61, 62, 63] (which simplified and strengthened the original result of Raz [54]), and the hardness of circuit lower bounds for $\mathsf{Res}(k)$, an extension of Resolution operating with k-DNFs, for a small parameter k [64]. The last result remains the strongest propositional lower bound for tautologies encoding circuit lower bounds.[1]

Ultimately, in [64] Razborov formulated intriguing conjectures about a conditional hardness of Nisan-Wigderson (NW) generators for Frege and Extended Frege system EF: suitable NW-generators should form good proof complexity generators in the sense that tautologies postulating the existence of elements outside their range require super-polynomial-size proofs in Frege systems, assuming $\mathsf{P/poly}$ is hard on average for NC^1, and in EF, assuming $\mathsf{NP}\cap\mathsf{coNP}$ is hard on average for $\mathsf{P/poly}$. If true, these conjectures would present a dramatic barrier establishing that feasibly-constructive logical reasoning (which can be defined in terms of short EF proofs or as the provability in Cook's theory PV_1 [29, 16]) cannot prove explicit circuit lower bounds – $\mathsf{P/poly}$ lower bounds in the case of Frege and lower bounds for a specific kind of $(\mathsf{NP}\cap\mathsf{coNP})/\mathsf{poly}$ circuits in the case of EF.[2] A similar earlier conjecture of Krajíček [30, Conjecture 7.9], implying that, under the existence of strong pseudorandom generators, subexponential circuit lower bounds are hard for EF was motivated by the role of dual weak pigeonhole principle in bounded arithmetic. In [35], Krajíček formulated the conjecture in an unconditional form.

The conjectures of Razborov and Krajíček are interesting not only because of the potential meta-mathematical consequences but also because they suggest the existence of a strong connection between circuit complexity and proof complexity.

---

[1] There are several related results about the impossibility of deriving strong circuit lower bounds in bounded arithmetic [46, 49, 39, 26] which appear to be incomparable to the lower bounds for propositional systems, but it seems that strengthening them in a more substantial way requires new propositional lower bounds. In the propositional setting, Austrin and Risse [5] showed that Sum-of-Squares proof system, which proves efficiently the pigeonhole principle, cannot prove efficiently circuit lower bounds for certain functions f, but the lower bound does not hold for all functions f. An analogous meta-mathematical theory is currently under development also in the algebraic setting [42].

[2] A lot of contemporary complexity theory is formalizable within EF, e.g., the PCP theorem has short EF proofs [47], see also [45]. Some candidates for PV_1-unprovable theorems from computational complexity are given in [14] but they exploit witnessing theorems which (under standard hardness assumptions) cannot be used to get the PV_1-unprovability of deterministic circuit lower bounds for SAT [46, 44].

Proof complexity presents a logic-oriented approach to the separation of NP and coNP which proceeds by proving lower bounds on lengths of proofs in increasingly more powerful propositional proof systems. Unfortunately, Razborov's unprovability results as well as other existing proof complexity lower bounds work only for very weak proof systems. In fact, in some sense proving proof complexity lower bounds has turned out to be harder than proving circuit lower bounds: turning AC^0 lower bounds into constant-depth Frege lower bounds required developing a substantially more involved adaptation of the method of random restrictions in the form of k-evaluations [1, 38, 51] and it still remains an open problem to derive lower bounds for Frege systems operating with $\mathsf{AC}^0[2]$ circuits despite the existence of explicit $\mathsf{AC}^0[2]$ lower bounds [56, 69]. From the perspective of the attempts to derive propositional lower bounds for tautologies expressing circuit lower bounds this is somewhat ironic as it suggests that deriving such hardness results unconditionally might be harder than proving circuit lower bounds themselves.

Generally, the relation between circuit complexity and proof complexity is rather mysterious. Formal connections between the two areas exist if we restrict ourselves to weak computational models. For example, the method of feasible interpolation derives proof complexity lower bounds for many weak proof systems from circuit lower bounds for monotone circuits and related models [34]. In the opposite direction, lifting theorems yield lower bounds for various monotone computational models from lower bounds for weak propositional proof systems, see e.g. [66, 21]. In the algebraic setting, IPS lower bounds imply $\mathsf{VP} \neq \mathsf{VNP}$ [22]. Unfortunately, these connections break when moving to strong proof systems and the Boolean setting. Recently a conditional connection between lower bounds for strong proof systems and strong circuit lower bounds such as $\mathsf{NP} \not\subseteq \mathsf{P/poly}$ has been established [50, 3], but it requires the assumption that the proof system proves efficiently some circuit lower bound such as E being hard on average for subexponential-size circuits.

As a consequence of the lack of formal links between circuit and proof complexity, the barrier results in circuit complexity do not seem to say much about the possibility of deriving proof complexity lower bounds for strong proof systems. The question of identifying a "natural proofs"-like barrier in proof complexity was raised already by Beame and Pitassi [6]. Given that some prominent proof complexity lower bounds are harder to derive than their counterparts in circuit complexity, it might seem even more ironic that we have not been able to identify such a barrier, see Section 1.2 for a discussion of earlier attempts. Is there an explanation for our inability to prove Frege lower bounds like those in Razborov's conjecture? In this paper we introduce a notion that has a potential to answer this question.

1.1 Our contribution

Following the paradigm of natural proofs, which turn constructive circuit lower bounds into efficient algorithms, we want to find a way to interpret (sufficiently constructive) proof complexity lower bounds as a construction of new surprising algorithms.

In order to get closer to such a conclusion we focus on prover-adversary games used by Pudlák and Buss [53] to characterize the lengths of proofs in Frege systems. We outline a slightly modified formulation of these games which is sufficiently general to capture uniformly the existing lower bounds for many different proof systems.

An s-size lower bound for a formula ϕ in a proof system P is often obtained by defining a strategy for an adversary who responds to queries of the prover in the following manner. Given a supposed s-size P-refutation π of $\neg\phi$, the prover sends to the adversary the last formula ψ of π and the adversary responds with 1 or 0. After that, the prover sends formulas

$\psi_1, \ldots, \psi_k$ from which ψ was derived in π by a single application of a derivation rule and the adversary responds with 1 or 0 to each of them. The prover then proceeds to a subproof of π which derives a formula $\psi_i \in \{\psi_1, \ldots, \psi_k\}$ assigned 0 by the adversary, if such a formula exists. If such a formula does not exist and the adversary assigned 0 to ψ, the adversary has been proven to be inconsistent (caught). The game continues in this way possibly until the players reach the axioms or $\neg\phi$. The adversary has to assign 0 to the contradiction (which is the last formula ψ, if π is a refutation) and 1 to the axioms and $\neg\phi$ (indicating that the adversary posses an assignment falsifying ϕ). If we can show that the adversary maintains consistency, we can conclude that π was not a correct refutation of $\neg\phi$. For a more elaborated discussion of the prover-adversary games see Section 2.

Intuitively, we define a constructive proof complexity lower bound by the existence of an "efficient adversary"; more precisely, by an efficient algorithm generating the responses of the adversary on given formulas such that for each short P-refutation (not necessarily of $\neg\phi$) the adversary (who might not be given the entire refutation) stays consistent in the game described above, cf. Definitions 8-9. The adversary always has to assign 1 to $\neg\phi$ and the axioms of P but not necessarily to other initial formulas of a given P-refutation. A randomized efficient adversary is allowed to use randomness so that for each short P-refutation the adversary can choose a random string and use it to determine its responses. To indicate that an adversary corresponds to a lower bound (for a formula ϕ) in a proof system P we speak of P-adversaries (claiming that $\neg\phi$ is satisfiable).

Note that in order to make a proof complexity lower bound for a formula ϕ constructive, we do not need to construct adversaries for short refutations of $\neg\phi$ – such refutations do not exist if the lower bound holds. In a sense, an efficient adversary rules out more than just the existence of a short refutation of $\neg\phi$, it could be said that it rules out the existence of a short "pseudoproof" of ϕ, cf. Proposition 10.

Our first theorem shows that constructive EF lower bounds for formulas $\mathsf{lb}_A(\mathsf{SAT}, n^2)$ (expressing that SAT is not computable by n^2-size circuits on a set of inputs $A \subseteq \{0,1\}^n$) given by randomized efficient EF-adversaries[3] claiming that $\neg\mathsf{lb}_A(\mathsf{SAT}, n^2)$ is satisfiable can be used to break the canonical pair of EF assuming that EF is capable of efficiently deriving a statement about antickeckers. Informally, the canonical pair of a proof system P consists of two disjoint NP sets: the set of all satisfiable formulas and the set of all unsatisfiable formulas whose unsatisfiability has a short P-proof. The canonical pair of P is hard to separate if there is no small circuit accepting all inputs from the first set and rejecting all inputs from the second set, cf. [58].

▶ **Theorem 1** (From adversaries to algorithms – Informal, cf. Theorem 11)**.**
Assume 1 - 2:

1. (The canonical pair of EF is hard.) *For each poly(n)-size circuit D with n inputs, there is z such that $D(z) = 0$ but z is a satisfiable formula or $D(z) = 1$ but there is a poly(n)-size EF-refutation of z.*
2. (Provably constructive antickeckers.) *There is $A \subseteq \{0,1\}^n$, $|A| = poly(n)$, and a poly(n)-size circuit F such that for each $z \in \{0,1\}^n$ there is a poly(n)-size EF-proof of the statement expressing that "if an n^2-size circuit B solves SAT on A, then a poly(n)-size circuit $F(B)$ solves SAT on z".*

Then for each randomized poly(n)-time EF-adversary claiming that $\neg\mathsf{lb}_A(\mathsf{SAT}, n^2)$ is satisfiable, there is a poly(n)-size EF-refutation (not necessarily of $\neg\mathsf{lb}_A(\mathsf{SAT}, n^2)$) catching the adversary with probability $\geq 1/2 - 1/n$ over the randomness of the adversary.

[3] The definition of EF-adversaries requires a careful treatment of the extension axioms, see Section 2.

Reading it contrapositively, Theorem 1 says that efficient adversaries defining constructive EF lower bounds for formulas $\mathsf{lb}_A(\mathsf{SAT}, n^2)$ yield efficient algorithms breaking the canonical pair of EF (a condition similar to the weak automatability of EF), assuming Condition 2 holds. Condition 2, the second assumption of the theorem, is less standard. It postulates the existence of a set of "anticheckers" A such that a small circuit which solves SAT on A can be turned into a circuit which solves SAT on any other input. Such A is known to exist unconditionally, cf. Section 2, but Condition 2 additionally says that this fact is efficiently provable in EF. It can be shown that Condition 2 follows if, e.g., EF proves efficiently that (a) E is hard on average for subexponential-size circuits and EF proves efficiently that (b) the existence of one-way permutation can be based on $\mathsf{NP} \not\subseteq \mathsf{P/poly}$. Both statements (a) and (b) are considered to be more approachable than non-uniform circuit lower bounds for SAT. The plausibility of Condition 2 can be further enhanced by replacing EF in Theorem 1 by stronger systems such as (a propositional equivalent of) ZFC, the theorem remains valid in this case as well. As pointed out by Noel Arteche (private communication), Condition 2 implies also that $\mathsf{NP} \not\subseteq \mathsf{Circuit}[n^\ell]$, for each ℓ. See Section 2 for a more detailed discussion.

The parameter $1/2 - 1/n$ in Theorem 1 is not essential for the purpose of a barrier result, any constant would work equally well.

The proof of Theorem 1 is straightforward: Since the adversary claims that it has a circuit solving SAT on A, using Condition 2 we can try to force it (with EF-queries) to claim that it has a circuit C which solves SAT everywhere and to simulate the computation of circuit C on any input z. This almost works except that we only get an efficient algorithm breaking the canonical pair of EF: If the adversary claims that C says that z is a satisfiable formula and z is unsatisfiable, we need the additional assumption that z has a short EF-refutation to reveal the inconsistency of the adversary (and force it to avoid such claims).

The proof works also for formulas $\neg\mathsf{tt}(\mathsf{SAT}, n^2) := \neg\mathsf{lb}_{\{0,1\}^n}(\mathsf{SAT}, n^2)$ instead of $\neg\mathsf{lb}_A$, but it crucially exploits the fact that the adversary is efficient w.r.t. the queries of the prover, not w.r.t. the formula for which the lower bound is being proved (which is of size $2^{O(n)}$ in this case). In fact, in the case of the tt-formulas we can avoid Condition 2 which can be replaced by a formalization of the search-to-decision reduction for SAT. Further, the switch from $poly(n)$-size to $2^{O(n)}$-size formulas allows us to replace the first assumption in Theorem 2 with $\mathsf{SAT} \notin \mathsf{P/poly}$. As a result we get

▶ **Theorem 2** (Informal, cf. Theorem 11)**.**
Assume that $\mathsf{SAT} \notin \mathsf{P/poly}$*. Then for each randomized p-time* EF*-adversary claiming that* $\neg\mathsf{tt}(\mathsf{SAT}, n^2)$ *is satisfiable, there is a* $2^{O(n)}$*-size* EF*-refutation catching the adversary with probability* $\geq 1/2 - 1/n$ *over the randomness of the adversary.*

Constructive proof complexity lower bounds. We complement Theorem 1 by showing that some of the major existing proof complexity lower bounds are indeed constructive.

As the first example we design efficient adversaries for the strongest known lower bound for lb-tautologies due to Razborov [64], who showed that there are no $2^{t^{\Omega(1)}}$-size $\mathsf{Res}(\Omega(\log t))$-refutations of $\neg\mathsf{lb}_W(f,t)$, for $n^2 \leq t \leq 2^n$ and any $W \subseteq \{0,1\}^n$. More precisely, we design efficient adversaries for "proof complexity generator" formulas $\tau_W(A,b)$ (expressing that a system of linear equations $Ax = b$ is solvable) instead of $\neg\mathsf{lb}_W(f,t)$. This is justified by the fact that Razborov's proof proceeds by reducing the lower bound for $\mathsf{lb}_W(f,t)$ to a lower bound for $\neg\tau_W(A,b)$ and Theorem 1 holds even if the formula $\neg\mathsf{lb}_W(\mathsf{SAT}, n^2)$ is replaced by any formula ϕ such that EF proves efficiently $\phi \to \neg\mathsf{lb}_W(\mathsf{SAT}, n^2)$, cf. Remark 13. The focus on the τ-formulas allows us to avoid the problem of encoding the reduction of lower bounds for $\mathsf{lb}_W(f,t)$ to lower bounds for $\neg\tau_W(A,b)$ in the strategy of the adversary, see Section 3.1 for more details.

▶ **Theorem 3** (Constructive Res(k) lower bound of Razborov – Informal, cf. Theorem 14)**.** *For each sufficiently small $\delta > 0$, there is $\epsilon > 0$ and a poly(t^{t^δ})-size randomized* Res(k)*-adversary C claiming that $\tau_W(A, b)$ is satisfiable, where $k = \epsilon \log t$ and $|W| = poly(t)$, such that for each $2^{t^{\Omega(\delta)}}$-size* Res(k)*-refutation π of any set of clauses, with probability $\geq 9/10$ over the randomness of C, π does not catch C.*

The complexity of the adversary in Theorem 3 is subexponential, but the formal version of Theorem 1 (Theorem 11) captures this setting of parameters. Our proof of Theorem 3 does not yield better than 2^n-size adversaries for $W = \{0,1\}^n$, which is the setting that would correspond to a constructive Res(k)-adversary for $\neg\mathsf{tt}(f, t)$, so we can match Theorem 3 with Theorem 1 but not with Theorem 2.

The proof of Theorem 3 requires a detailed analysis of Razborov's original proof combined with a different proof of the final width lower bound. We did not manage to exploit [64] in a more black-box fashion. Verifying Theorem 3 thus requires reading the entire proof of Razborov (with a more sparse restriction while keeping the original expansion parameters) except the switching lemma which can be treated mostly as a black-box. As a result of this our write-up becomes dense, with many references to the intricacies of the proof in [64]. We replace the width lower bound in Razborov's proof by an adaptation of a width lower bound presented by Krajíček [34, Lemma 13.4.5]. In the adaptation, instead of using Krajíček's expansion, we use Razborov's "lossless" expansion.

The second example we present is a constructive version of the method of pseudowidth, a sophisticated analogue of the popular method of width [7], which was used to derive Resolution (Res) lower bounds for the weak pigeonhole principle and for formulas encoding circuit (or even DNF) lower bounds [54, 61, 62, 63, 20]. Pseudowidth and width are currently incomparable: It is not known how to prove Res lower bounds for the weak (enough) pigeonhole principle or for formulas encoding DNF lower bounds using width and it is not known how to prove Res lower bounds for proof complexity generator formulas $\tau_W(A, b)$ considered in Theorem 3 using pseudowidth.

We present a constructive version of the lower bound for the weak pigeonhole principle PHP^m_n mapping m pigeons to n holes.

▶ **Theorem 4** (Constructive pseudowidth lower bound – Informal, cf. Theorem 20)**.** *There is a poly(m)-size randomized* Res*-adversary C claiming that $\neg\mathsf{PHP}^m_n$ with $m = 2^{n^{1/5}}$ is satisfiable such that for each $2^{\Omega(n^{1/4})}$-size* Res*-refutation π of any set of clauses, with probability $\geq 9/10$ over the randomness of C, π does not catch C.*

Our proof of Theorem 4 differs substantially from the proofs in [54, 61, 62, 63, 20] and might be of independent interest. The strategy of the adversary from Theorem 4 is guided by random partial multi-matchings. The notion of pseudowidth arises here naturally from the observation that the partial multi-matchings satisfy certain clauses with high probability – these clauses can be thus treated as pseudo-axioms.

Finally, we look at the crown jewel of proof complexity, the lower bound on the length of proofs of the pigeonhole principle PHP_n (mapping $n+1$ pigeons to n holes) in depth d Frege systems (AC^0_d-Frege), which is obtained via k-evaluations [1, 38, 51].

Unfortunately, it is not known how to prove a lower bound for tt-tautologies in AC^0-Frege. The closest result of this kind was obtained in [27], where it was shown that a particular kind of $\mathsf{NP} \cap \mathsf{coNP}$ lower bounds (similar to formulas $\neg\tau_W(A, b)$), formalized in a suitable way, does not admit short proofs in AC^0-Frege. The lower bound from [27] proceeds by a reduction to the PHP_n lower bound.

▶ **Theorem 5** (Constructive AC^0-Frege lower bound – Informal, cf. Theorem 21)**.** *Let $\epsilon < 1/9^d$. There is a poly(n^{n^ϵ})-size randomized AC^0_d-Frege-adversary C claiming that $\neg\mathsf{PHP}_n$ is satisfiable such that for each 2^{n^ϵ}-size AC^0_d-Frege-refutation π of any set of clauses, with probability $\geq 9/10$ over the randomness of C, π does not catch C.*

The proof of Theorem 5 follows directly by inspection of the original proof. On high level, the reason why the theorem holds is that a random partial matching "kills" short AC^0-Frege proofs of PHP_n with high probability.

Unification. Theorems 3-5 imply the corresponding proof complexity lower bounds for $\mathsf{Res}(k)$, Res and AC^0-Frege. Moreover, our presentation of lower bound methods of width, pseudowidth and k-evaluation in terms of adversaries provides a more unifying perspective on these methods by showing how they arise from natural choices of strategies of adversaries.

1.2 Related work

There are several results in proof complexity that could be interpreted as barriers for deriving proof complexity lower bounds.

Incompleteness. Already Gödel's second incompleteness theorem implies that, e.g., if ZFC is consistent, then ZFC cannot prove superlinear lower bounds for the propositional proof system obtained from ZFC in a natural way (otherwise ZFC would prove its own soundness). For similar reasons, a result of Krajíček [31, Corollary 3.2] implies that Buss's theory S^1_2 cannot prove superlinear lower bounds on the so called Implicit EF. These results, however, do not present an actual barrier for proving proof complexity lower bounds because in proof complexity we prove lower bounds under the assumption that the proof system in question is sound.

Cook's translation. A more substantial proof complexity barrier was obtained by Krajíček-Pudlák [36], who showed that Cook's theory PV_1 cannot prove that there exists a tautology which requires superpolynomial, say $n^{\log n}$-size, EF proofs (Buss [11] and Cook-Urquhart [18] derived the same unprovability result for the intuitionistic PV_1). The result crucially relies on the formalization of the statement of the lower bound, which is unfortunately not natural from the perspective of PV_1 in the sense that $n^{\log n}$ is not the length of a number. Consequently, PV_1 cannot use induction over the proofs of size $n^{\log n}$. Informally, the theory cannot see (and analyze) the whole $n^{\log n}$-size proof properly.

Nisan-Wigderson generators and the sum-check protocol. This drawback was eliminated in a proof complexity barrier obtained in [49] (building on [32, 46] and extended in [39, 26]), where it was shown that subexponential average-case lower bounds for nondeterministic circuits are not provable in PV_1 (and certain extensions of PV_1). In a related work using completely different methods, Chen, Rothblum and Tell [15] constructed a non-deterministic algorithm P such that under a derandomization assumption S^1_2 cannot prove that P is not a p-bounded propositional proof system.[4] Both [49] and [15] use a formalization which allows the theory to perform induction over objects of even exponential size $2^{O(n)}$. Unfortunately, the complexity of these formalizations is Σ^p_2 and while theories like PV_1 and S^1_2 can prove a

[4] [15] presents the unprovability result for APC_1 but it works for S^1_2 as well.

lot of complexity theory and are generally very powerful when reasoning with P-time objects and deriving coNP statements, they are not well-designed for reasoning with NP statements[5], which is why these results do not present a significant barrier in the practice of proving lower bounds.

This issue was partially addressed in [28]. We believe that [28] implicitly shows that there is a concrete MA proof system R such that, if it is hard to derive superpolynomial circuit lower bounds in a sufficiently strong propositional proof system P, then the theory T_P corresponding to P cannot prove that R does not admit p-size proofs of all tautologies of length n. Here, the lower bound for R assumes that $2^{O(n)}$ is the length of a number and it is formulated as a coNP statement, so the T_P-provability of the lower bound for R does not necessarily imply that P simulates R and T_P should be in principle very powerful when proving lower bounds for R. The result in [28] would therefore give us a remarkable proof complexity barrier if we could only prove the assumption or at least replace it by a standard complexity-theoretic conjecture.

Krajíček [33] used the methods of [32] to show that under standard hardness assumptions, there is no efficient algorithm which given an algorithm P (potentially defining a propositional proof system) outputs a tautology hard for P or a witness that P does not define a consistent proof system.

Natural properties. In an attempt to mimic the natural proofs barrier, [48] defined a version of natural properties against proof systems roughly as the efficient provability of EF lower bounds for formulas $\mathsf{tt}(f, n^{O(1)})$, for many functions f, and showed that if super-bits exist (an assumption about the existence of pseudorandom generators safe against nondeterministic circuits), then such natural properties against EF do not exist or EF does not prove efficiently any formula $\mathsf{tt}(f, n^{O(1)})$. If we knew how to avoid the second disjunct about the EF-hardness of $\mathsf{tt}(f, n^{O(1)})$, this would present a close analogue of the natural proofs barrier in proof complexity.

A combination of the results in [28] and [48] shows that under the assumption of the existence of super-bits, either there is no natural proof against P or P cannot prove proof complexity lower bounds for a concrete MA system. The conclusion is interesting, but it does not tell us much about the hardness of deriving lower bounds for systems like EF.

Self-provability. More recently it has been shown that EF lower bounds imply circuit lower bounds such as $\mathsf{SAT} \notin \mathsf{P/poly}$ [50, 3] under various assumptions such as that EF proves efficiently that E is hard on average for subexponential-size circuits. If we could avoid this assumption, the resulting strong connection between EF lower bounds and circuit lower bounds could explain a part of the difficulty of deriving EF lower bounds.

A problem with interpreting the results from [48] or [50, 3] as proof complexity barriers is that they require the assumption that EF proves efficiently a sufficiently strong circuit lower bound, which is the very statement we are trying to refute when proving proof complexity lower bounds (in Razborov's conjecture).

TFNP, automatability & algebraic complexity. Li, Li and Ren [40] initiated a systematic investigation of proof complexity of proof complexity lower bounds through the lens of TFNP, but they restrict their attention to Resolution lower bounds, see also [19]. Similarly, the

[5] For example, it follows from Buss's witnessing theorem [10] that S^1_2 cannot prove that a TFNP problem is total unless the problem is in P.

NP-hardness results for the automatability of various weak proof systems are based on the hardness of deriving proof complexity lower bounds in systems such as Resolution [4, 2]. Lu, Santhanam and Tzameret [42] showed that certain lower bounds for a constant-depth version of IPS are hard to prove in $\mathsf{AC}^0[2]$-Frege. The result builds on a work of Santhanam and Tzameret [67] where IPS lower bounds for formulas expressing IPS lower bounds for VP $\neq$ VNP were shown to follow from VP $\neq$ VNP. A drawback of [67] when interpreted as a barrier result is that the conclusion holds only for infinitely many input lengths, [67] does not rule out the possibility that IPS proves efficiently lower bounds for IPS for VP $\neq$ VNP on an infinite and dense set of input lengths. Nevertheless, the potential sensitivity to the input length is a valuable information about the possibility of deriving proof complexity lower bounds for strong proof systems.

Upper bounds. Ultimately, the prospect of proving proof complexity lower bounds is limited by the possible existence of surprising upper bounds. In this context, it is worth mentioning an MA protocol of Williams which breaks a version of SETH for MA [70]. A recent result of Mazor and Pass [43] and Hirahara, Ilango and Williams [24] constructs subexponentials-size (deterministic) circuits for more specific problems like MCSP, but so far all these upper bounds remain in the regime of SETH.

All in all, while there have been several previous results that can be seen as some kind of proof complexity barriers, they are significantly less impactful than natural proofs in circuit complexity. In our view, the concept of efficient adversaries introduced in the present paper comes the closest to the potential "natural proofs barrier" in proof complexity so far despite the fact that the second assumption in Theorem 1 is similar to the assumption that EF proves efficiently some circuit lower bound. Theorem 2 could turn out as the main barrier, if the existing lower bounds such as the $\mathsf{Res}(k)$ lower bound for $\mathsf{tt}(f,t)$ can be made more constructive by designing adversaries who are efficient w.r.t. the queries of the prover, not w.r.t. the formula for which the lower bound is proved.

1.3 Open problems

Constructivity of proof complexity lower bounds. The main problem we leave open is the question of constructing efficient adversaries for all the remaining major proof complexity lower bounds. As the proofs of Theorems 3-4 show, it is not always obvious how to constructivize an existing lower bound. For example, in Theorem 4 we only obtained efficient adversaries for the weak pigeonhole principle, but the question of constructing such adversaries for Res lower bounds for the *functional* weak pigeonhole principle [62, 63] remains open. Further, it is not clear how to get efficient adversaries from lower bounds exploiting "deterministic" restrictions such as the recent super-linear lower bound for constant-depth Frege systems by Gryaznov and Talebanfard [23]. In general, lower bounds obtained via the width-size relation of Ben-Sasson and Wigderson [7] use deterministic restrictions as well and we do not know how to make them constructive. That is, we do not have a general strategy for replacing deterministic restricitions whose existence is shown non-constructively by random restrictions which would work with high probability. A prominent example we leave as an open problem is the constructivization of lower bounds based on feasible interpolation such as Pudlák's lower bound for Cutting Planes CP [52] – notably, it is still an open problem to show that $\mathsf{tt}(f,s)$ tautologies are hard for CP. Nevertheless, a significant barrier for using feasible interpolation to derive lower bounds for strong proof systems has been established already in [37, 9, 8].

▶ **Problem 6.** *Does the* CP *lower bound of Pudlák [52] admit efficient adversaries?*

If there are lower bounds which cannot be made constructive, can we rule out the possibility of a more substantial proof complexity barrier by showing at least that under a standard complexity-theoretic assumption some proof complexity lower bound does not admit an efficient adversary?

If we cannot construct an efficient adversary for some lower bound, if for each efficient P-adversary claiming that $\neg\phi$ is satisfiable, there is a short P-refutation (not necessarily of $\neg\phi$) catching the adversary, we can say that the refutations constitute a short *P-pseudo-proof* of ϕ. Is there a proof system P with p-size pseudo-proofs of all tautologies? To formulate the question more formally we consider only proof systems EF_P obtained by extending EF with axioms postulating the soundness of a proof system P. EF_P simulates P [34]. An EF_P-adversary is an EF-adversary who is additionaly required to assign 1s to all instantiations of the axiom postulating the soundness of P.

▶ **Problem 7** (NP = coNP w.r.t. efficient adversaries)**.** *Is there a proof system P such that for each tautology ϕ, for each $(|\phi|+n)^3$-size EF_P-adversary claiming that $\neg\phi$ is satisfiable there is a poly$(|\phi|)$-size EF_P-refutation catching the adversary?*

Avoiding anticheckers. Is it possible to remove the assumption about the provably constructive anticheckers from Theorem 1? In order to do so, it might be helpful to interpret efficient adversaries as algorithms for the search problems of finding errors in supposed short refutations of hard formulas, cf. [40].

Efficient adversaries w.r.t. tt-tautologies and other formulas. Theorem 2 does not rule out the possibility of proving constructively EF lower bounds for $\mathsf{tt}(f,s)$ formulas with adversaries which are efficient (not w.r.t. the queries of the prover but) w.r.t. the size of the $\mathsf{tt}(f,s)$ formula, i.e. with adversaries of size $2^{O(n)}$. Is it possible to make Theorem 2 work in this case as well? Can we further extend the theorem to all formulas instead of just formulas expressing circuit lower bounds? The latter question is related to the problem of proving that MCSP is NP-complete. Another relevant problem in this direction is to make Theorem 3 work for $\mathsf{tt}(f,t)$ with better than $2^{O(n)}$-size adversaries, i.e. to construct $\mathsf{Res}(k)$-adversaries for $\neg\mathsf{tt}(f,t)$ which are efficient w.r.t. the queries of the prover.

Simulations preserving efficient adversaries. The notion of efficient adversaries appears to be sensitive to the specific choice of proof system in question. Do p-simulations preserve efficient adversaries? That is, is it possible to show that if there is an efficient P-adversary for Φ and the proof system P is p-equivalent to a proof system R, then there is an efficient R-adversary for Φ? Is there another natural notion of simulation between proof systems which preserves efficient adversaries?

Connection to automatability. Automatability of a proof system P should yield a close-to-optimal P-adversary because the optimal strategy for the adversary proceeds by deciding in each step if a given query has a short derivation from formulas that have been already assigned 1. Is it possible to use this connection to shed new light on the question of automatability of standard proof systems?

2 From efficient adversaries to algorithms

In this section we formally introduce efficient adversaries and prove Theorem 1.

Standard proof systems. We define adversaries for standard propositional proof systems from the set SYS consisting of:

- Extended Frege EF, Frege, AC^0_d-Frege and Circuit Frege CF systems;
- Resolution Res; Cutting Planes CP; k-DNF Resolution Res(k);
- tree-like versions of the systems listed above.[6]

The definitions of these systems can be found in [34], but we will use the definition of Res(k) from [64, Definition 2.6] and the definition of CF from [25]. Each instantiation of EF, Frege, AC^0_d-Frege and CF system is represented in SYS separately. This is because we do not know whether the existence of an efficient adversary for a system P guarantees the existence of an efficient adversary for a system P' if P' is p-equivalent to P. We consider Frege, EF and CF systems for all finite and complete languages for propositional logic. SYS^- will stand for SYS without EF systems.

We denote by $O(P)$ the set of "objects" a proof system $P \in \mathsf{SYS}$ operates with. EF and Frege systems operate with formulas built from the connectives in the respective language. AC^0_d-Frege operates with formulas of depth d defined as in [34, Chapter 15], CF systems operate with circuits in the respective language, Res operates with disjunctions of literals (clauses), CP operates with integer inequalities, Res(k) with k-DNFs.

In order to unify the presentation, we treat all systems $P \in \mathsf{SYS}$ as refutational systems, refer to the objects of $O(P)$ as formulas (in the case of CF this is an abuse of terminology but it should not cause any confusion) and denote by 0 the last formula of a refutation.

Pudlák-Buss game. Our definition of efficient adversaries is inspired by the Pudlák-Buss game [53], which can be used to characterize short Frege proofs. The game is played between Prover and Liar. In each round Prover sends a formula to Liar who responds by 0 or 1. Denote the response of Liar on formula ψ by $C(\psi)$, not displaying the history of the game for simplicity. Intuitively, we interpret $C(\psi) = 0/1$ as claims that Liar possesses a falsifying/satisfying assignment of ψ. In each round, Liar has to satisfy $C(0) = 0$ and $C(\phi) = 1$, if we want to estimate the size of the refutation of a formula ϕ.

Given an s-size Frege-refutation of ϕ, there is a strategy of Prover "catching" Liar in $K \log s$ rounds by querying formulas of size $\leq Ks^K$, where $K \in \mathbb{N}$ depends only on the Frege system: Using binary search Prover can find in $K \log s$ rounds formulas $\phi_1, \ldots, \phi_{k+1}$ such that $C(\phi_i) = 1$, for $i \in [k]$, $C(\phi_{k+1}) = 0$ and ϕ_{k+1} follows from $\phi_1, \ldots, \phi_k$ by a single application of a Frege-rule.

In the opposite direction, such a strategy of Prover that catches Liar in r rounds with queries of size $\leq 2^{O(r)}$ yields a $2^{O(r)}$-size tree-like Frege-refutation of ϕ, if the logical connectives in the language of the system include $0, 1, \neg, \vee, \wedge$, cf. [34, Lemma 2.2.3].

▶ **Definition 8** (Efficient adversary). *For $P \in \mathsf{SYS}^-$, an s-size P-adversary is a sequence $C := \{C_i\}_{i=1}^{\infty}$, where C_i computes the responses of the adversary in the i-th round as follows. Each C_i is a sequence of s-size circuits $C_i := \{C_{i,j}\}$ which for each $\langle a_1, \ldots, a_i \rangle \in \mathbb{N}^i$ contains a single circuit $C_{i,j}$ with i blocks of inputs partitioned so that the k-th block, $k \leq i-1$, consists of $a_k + 1$ inputs and the i-th block consists of a_i inputs. Given a history of the adversary's responses $h := \{\langle \phi_k, b_k \rangle\}_{k=1}^{i-1}$, where $\phi_k \in O(P)$ and $b_k \in \{0, 1\}$ is a response of the adversary to ϕ_k in the k-th round, the response of the adversary to $\phi_i \in O(P)$ in the i-th round is $C_i(h, \phi_i) := C_{i,j}(h, \phi_i)$, where $C_{i,j}$ is the circuit in C_i with inputs partitioned in accord with*

[6] Strictly speaking, it does not make much sense to consider tree-like EF as we want to allow the system to reuse extension axioms.

h, ϕ_i. The P-adversary is honest *if it responds consistently with its responses in previous rounds, i.e. for each $\phi \in O(P)$ and each history $h = \{\langle \phi_k, b_k \rangle\}_{k=1}^{i-1}$, $C_i(h, \phi) = b_\ell$ if $\phi = \phi_\ell$ for some $\langle \phi_\ell, b_\ell \rangle \in h$ with $\ell < i$.*

If the adversary responds to $\phi \in O(P)$ with a fixed $b \in \{0, 1\}$ in each round independently of the history, we write $C(\phi) = b$. If the P-adversary satisfies $C(0) = 0$ and $C(\phi) = 1$, for all $\phi \in \Phi \subseteq O(P)$, we say that the adversary is for Φ*. We will consider only finite sets Φ.*

A randomized s-size P-adversary *(*for Φ*) is an s-size P-adversary $\{C_i\}_{i=1}^{\infty}$ (*for Φ*) such that for each i, each $C_{i,j} \in C_i$ has access to $\leq s$ bits of a finite random string of bits shared among all circuits $C_{i,j}$, for all i, j. The random strings are distributed according to some arbitrary distribution.*

The definition of EF-adversaries requires a small twist which will become apparent after Definition 9. For $P \in \mathsf{SYS}^-$, a P-prover is a function mapping histories $\{\langle \phi_k, b_k \rangle\}_{k=1}^{i}$, where $\phi_k \in O(P)$, $b_k \in \{0, 1\}$, $i \in \mathbb{N}$, to new formulas (queries) in $O(P)$.

Switching quantifiers and the role of randomness. If for a given prover with $\leq s$-size queries, there is an adversary who stays consistent for r rounds, then there is a $poly(sr)$-size adversary who achieves that as well. To make the notion of efficient adversaries interesting we will thus switch quantifiers and consider s-size adversaries staying consistent for r rounds against all possible provers. A randomized adversary will be slightly more flexible: for each prover, the randomized adversary will choose a random string and then proceed as the deterministic adversary specified by the random string. The definition ensures that circuits in C_2 responding in the second round can access the random bits used by circuits C_1 responding in the first round.

Next we introduce a notion that allows us to generalize the characterization of proof-size from Pudlák-Buss games to all systems in SYS^-.

▶ **Definition 9** (Evaluation of adversary). *Given a refutation π of $\Phi \subseteq O(P)$ in a proof system $P \in \mathsf{SYS}^-$ and an (s-size) P-adversary C for some $\Phi' \subseteq O(P)$, define the* evaluation of C on π *by the following process. In round 0, let $h_0 := \emptyset$ and start at the last formula of π, i.e. 0. If in the i-th round, the current formula $\phi \notin \Phi$ is derived in π by applying a single derivation rule on (a possibly empty set of) formulas $\phi_j, j \in [k]$, proceed to the first $\phi_{j'}, j' \in [k]$, such that $C_{i+1}(h_i, \phi_{j'}) = 0$ and let $h_{i+1} := h_i \cup \langle \phi_{j'}, 0 \rangle$. If there is no such $\phi_{j'}, j' \in [k]$, output 0. If $\phi \in \Phi$ is reached, output 1. If π evaluated C to 0, we say that π* caught *C. Otherwise, C* stayed consistent on *π.*

Obviously, there is an s-size refutation of Φ in a proof system $P \in \mathsf{SYS}^-$ if and only if there is an s-size P-refutation of Φ which catches all P-adversaries for Φ.

EF adversaries. Definitions 8-9 can be straightforwardly adapted to the case of EF systems, but the extension axioms need to be treated carefully: It would be unfair to ask an EF adversary to assign a value to an extension variable e and only subsequently reveal that e is equivalent to 1 or 0 – the adversary would be caught immediately. The issue does not appear in the case of the p-equivalent CF systems. This also suggests the right way to define EF adversaries: We adapt Definitions 8-9 so that whenever an EF adversary evaluated on an EF refutation π receives a formula ϕ, they receive also all extension axioms in π defining the extension variables appearing in ϕ and the extension axioms of all the extension variables in the received extension axioms. In other words, the adversary gets to know the whole circuit defining the extension variables in ϕ.

To justify the definition, following the proof of the simulation of EF by CF [25, Lemma 2.4], it can be shown that for each EF system P and each CF system R such that P and R have the same language and Frege rules, if each $poly(s)$-size P-adversary is caught by some t-size P-refutation, then each s-size R-adversary is caught by some $poly(t)$-size R-refutation.

Proposition 10 points out that for Frege systems, an honest adversary for Φ stays consistent against all provers with $poly(n)$-size queries for $O(\log n)$ rounds if and only if it stays consistent on all $poly(n)$-size $O(\log n)$-height refutations (not necessarily of Φ). (Note that an s-size Frege proof can be transformed to a $poly(s)$-size $O(\log s)$-height Frege proof of the same formula [34].) A height of a Frege derivation is the maximal number of edges on a directed path in the derivation graph.

▶ **Proposition 10.** *Let P be a Frege system containing modus ponens rule (which derives ψ from ϕ and $\neg\phi \vee \psi$).*

1. *If an s-size r-height P-refutation (not necessarily of Φ) catches a P-adversary for Φ, then some P-prover with s-size queries catches the adversary in $O(r)$ rounds.*
2. *If a P-prover with s-size queries catches an honest P-adversary for Φ in r rounds, then some $poly(s, r)$-size $(r + 2)$-height P-refutation (not necessarily of Φ) catches the adversary.*

Proof. To prove Item 1, we define a strategy of the prover by querying formulas on the path in the P-refutation going from 0 to an axiom where the adversary assigns 0s.

Item 2 is obtained by treating each query as a resolvent ϕ in the modus ponens rule deriving ψ from ϕ and $\neg\phi \vee \psi$. The formula ψ is initially 0. If ϕ is assigned 0, we proceed with extending the refutation we have built so far by building a derivation of ϕ and leaving $\neg\phi \vee \psi$ as an additional axiom. Otherwise, we proceed with building a derivation of $\neg\phi \vee \psi$ and leaving ϕ is an axiom. The evaluation of the adversary on the resulting refutation π ends up in one of the following two cases. In Case 1, the last queried formula ϕ_r is assigned 0 but ϕ_r follows by a single application of a Frege rule from formulas assigned 1. In Case 2, the last queried formula is assigned 1 and again a "direct contradiction" is found in which some formula ϕ_r is assigned 0. In Case 1, we extend π by a single derivation step deriving ϕ_r from the formulas assigned 1. In Case 2, we extend π by a modus ponens rule resolving on ϕ_r and then adding the derivation rule deriving ϕ_r from the formulas assigned 1. Since the adversary has to stay consistent with its previous responses, in both cases, we obtain an $\leq (r + 2)$-height P-refutation of size $poly(s, r)$ which catches them. ◀

Proposition 10 holds also for AC^0_d-Frege systems with a minor adjustment of parameters. We do not know how to improve the conclusion in Proposition 10 (Item 2) into a P-refutation of Φ catching the adversary, but we will show that a typical proof complexity lower bound for a formula Φ yields a randomized efficient (possibly dishonest) adversary for Φ who stays consistent on all small refutations (not necessarily of Φ) with probability $\geq 9/10$ over the randomness of the adversary, if the adversary is not required to assign 1s to formulas outside Φ. This should be contrasted with the next theorem showing that under certain assumptions it is not possible to construct such EF-adversaries for formulas encoding circuit lower bounds for SAT.

In order to state the theorem we need to introduce some notation. Let $\neg\mathsf{lb}_A(\mathsf{SAT}, s)[B]$ be a propositional formula encoding the $poly(s, |A|)$-size predicate $R_1(B) :=$ "variables B encode an s-size circuit with n inputs computing SAT on inputs from $A \subseteq \{0, 1\}^n$". The formula can use other hidden auxiliary variables as well and is satisfiable iff $\exists B, R_1(B) = 1$. The values of SAT on A are "hardwired" bits. Sometimes we write $\mathsf{lb}_A(\mathsf{SAT}, s)$ without displaying $[B]$. We

define $\mathsf{tt}(\mathsf{SAT}, s)$ as $\mathsf{lb}_{\{0,1\}^n}(\mathsf{SAT}, s)$. The notation SAT should not be confused with $\mathsf{SAT}(x)$ which we use as a shortcut for saying that "x is a satisfiable formula" and with $\mathsf{SAT}(x, y)$ used to denote a propositional formula encoding the predicate $R_2(x, y) :=$ "x encodes a propositional formula satisfied by assignment y". For concreteness, we assume that $\mathsf{SAT}(x, y)$ uses extension variables encoding the gates of a $poly(|x|, |y|)$-size circuit computing $R_2(x, y)$ and is a tautology for an assignment a, b of x, y iff $R_2(a, b) = 1$. Similarly, given a circuit F with inputs y, the predicate $R_3(x, y) :=$ "x is a propositional formula satisfied by $F(y)$" can be encoded by a propositional formula $\mathsf{SAT}(x, F(y))$ using extension variables for the gates of a $poly(|F|, |x|)$-size circuit computing $R_3(x, y)$ and is a tautology for an assignment a, b of x, y iff $R_3(a, b) = 1$. We do not specify the encoding of these formulas further as the results in this section work for any natural encoding.

▶ **Theorem 11.**
For each EF *system* P *there is* $K \in \mathbb{N}$ *such that the following holds.*

Let $v : \mathbb{N} \mapsto \mathbb{N}$, $n, s, t, u, w \in \mathbb{N}$ *be such that* $n < s \leq t \leq u$ *and assume 1 - 2:*

1. (The canonical pair of P is hard.) *For each* $(n \cdot v(t^K))^K$*-size circuit* D *with* n *inputs, there is* z *such that* $\mathsf{SAT}(z) \wedge D(z) = 0$ *or* $D(z) = 1$ *while there is a* w*-size* P*-refutation of (the formula encoded by)* z*.*
2. (Provably constructive anticheckers.) *There is a* t*-size set* $A \subseteq \{0,1\}^n$ *and a* t*-size circuit* F *such that for each* $z \in \{0,1\}^n$ *there is a* u*-size* P*-proof of*

$$\neg\mathsf{lb}_A(\mathsf{SAT}, s)[B] \wedge \mathsf{SAT}(z, y) \to \mathsf{SAT}(z, F(z, B)).$$

Then,

a) *For each (possibly dishonest) randomized* v*-size* P*-adversary* $C = \{C_i\}_{i=1}^{\infty}$ *for* $\neg\mathsf{lb}_A(\mathsf{SAT}, s)$*, there is a* $(w + u)^K$*-size* P*-refutation catching* C *with probability* $\geq 1/2 - 1/n$ *over the randomness of* C*.*

b) *For each randomized* v*-size* P*-adversary* C *for* $\neg\mathsf{tt}(\mathsf{SAT}, s)$*, there is a* $(2^n + w + u)^K$*-size* P*-refutation catching* C *with probability* $\geq 1/2 - 1/n$ *over the randomness of* C*.*

Moreover, Item b) holds even without the assumption in Item 2, if $s \leq 2^n$*.*

Typical parameters of interest in Theorem 11 are $s, t, u = poly(n)$ and v being a polynomial function in n and the size of the input, where the size of the input might be $\ll n$ or $\gg n$. In Item a), we typically consider $w = poly(n)$ while in Item b) it is best to set $w = 2^{O(n)}$.

▶ Remark 12. For $w = 2^{O(n)}$, which is the setting suitable for the "moreover" part of Theorem 11, Item 1 in the assumptions can be replaced by $\mathsf{SAT} \notin \mathsf{Circuit}[(n \cdot v(t^K))^K]$ because each unsatisfiable z has a $2^{O(|z|)}$-size EF-refutation.

Proof of Theorem 11.

Item a). Assume that the modus ponens rule is available in P and let D^r be the n-input $v(poly(t))$-size randomized circuit

$$D^r(x) := C_{1,j}(\mathsf{SAT}(x, F(x, B))),$$

where $C_{1,j}$ is the randomized circuit in C_1 with the appropriate input size, F is the circuit from the second assumption and B are the variables from the displayed formula in the second assumption. In more detail, D^r maps $x \in \{0,1\}^n$ to $\mathsf{SAT}(x, F(x, B))$ and sends the formula to $C_{1,j}$ together with the extension axioms of the extension variables in $\mathsf{SAT}(x, F(x, B))$. Intuitively, D^r attempts to use C to solve SAT. However, there is an input z of D^r on which D^r fails to separate the canonical pair of P (for the given w) with probability

$\geq 1/2 - 1/n$. Otherwise, by an amplification employing Chernoff's bound and the subsequent derandomization, a circuit of size $poly(n) \cdot |D^r| + poly(n)$ would separate the canonical pair of P, contradicting Item 1.

The last step of the refutation catching C derives 0 from $\mathsf{SAT}(z, F(z, B))$ and its negation (using modus ponens).[7] By the choice of z, with probability $\geq 1/2 - 1/n$ over the randomness of C, one of the following two cases happens.

Case I: $D^r(z) = 1$ and there is a w-size P-refutation of z. A w-size P-refutation of z yields a $poly(w + t)$-size P-derivation π_0 of $\neg\mathsf{SAT}(z, F(z, B))$ from empty assumptions, cf. [34, Chapter 12.4]. Since extension variables cannot appear in the initial formulas of a P-refutation containing π_0 (they would not be allowed to be introduced in π_0), we also need to use a $poly(t)$-size P-derivation π_2 of $\mathsf{SAT}(z, F(z, B))$ from 0. The refutation catching C is then constructed by composing π_0, π_2 with the application of modus ponens on $\mathsf{SAT}(z, F(z, B))$ and its negation. It is a refutation of 0.

Case II: $D^r(z) = 0 \wedge \mathsf{SAT}(z)$. Since $\mathsf{SAT}(z)$ and we assume Item 2, there is a $poly(u)$-size P-derivation π_1 of $\mathsf{SAT}(z, F(z, B))$ from $\neg\mathsf{lb}_A(\mathsf{SAT}, s)$. Again, in order to deal with the extension axioms, we construct a $poly(t)$-size P-derivation π_2' of $\neg\mathsf{SAT}(z, F(z, B))$ from 0 and then the P-refutation catching C analogously as in Case I. It is a refutation of $\neg\mathsf{lb}_A(\mathsf{SAT}, s)$ and 0.

Since only one of the cases I and II can happen, this yields the desired refutation.

If modus ponens is not available in P, we proceed similarly. Given $x \in \{0,1\}^n$, D^r constructs a $poly(t)$-size P-derivation π_x of 0 from $\mathsf{SAT}(x, F(x, B))$ and its negation (expressed in the language of the system). Then, D^r simulates the evaluation of C on π_x (it suffices to store finitely many $v(poly(t))$-size circuits from C to simulate the evaluation for all $x \in \{0,1\}^n$). If the evaluation ends up in the formula $\mathsf{SAT}(x, F(x, B))$, D^r outputs 0. If the evaluation ends up in the formula $\neg\mathsf{SAT}(x, F(x, B))$, D^r outputs 1. Otherwise, D^r outputs an arbitrary bit, say, 0. The size of $D^r(x)$ is $poly(v(poly(t)))$. The existence of $z \in \{0,1\}^n$ such that Case I or Case II happens follows as before. The refutation catching C is constructed as before but the final application of modus ponens, i.e. the last step of the refutation, is replaced by π_z. The refutation catches C because in Case I, the evaluation of C on π_z cannot end up in $\mathsf{SAT}(z, F(z, B))$ – in such case D^r would output 0 on z – and in Case II the evaluation on π_z cannot end up in $\neg\mathsf{SAT}(z, F(z, B))$ – in such case D^r would output 1 on z. This completes the proof of Item a).

Item b). We proceed as in Item a) but instead of the assumption from Item 2 we use a formalization of the search-to-decision reduction for SAT. In more detail, let F be a $poly(s)$-size circuit which solves the search version of SAT given an s-size circuit B witnessing that $\neg\mathsf{tt}(\mathsf{SAT}, s)$. As before, the last step of the refutation resolves on $\mathsf{SAT}(z, F(z, B))$, for a suitable z (or this step is simulated). The only difference from Item a) is that in Case II we use a $2^{O(n)}$-size P-derivation of $\mathsf{SAT}(z, F(z, B))$ from $\neg\mathsf{tt}(\mathsf{SAT}, s)$. The existence of the P-derivation follows from the S^1_2-formalization of the search-to-decision reduction for SAT, see [46, Proposition 3.1] or [17], and the fact that Π^b_1-consequences of S^1_2 have p-size proofs in EF [29]. ◀

While the assumption in Item 1 in Theorem 11 is standard (roughly corresponding to the non-automatability of P), Item 2 deserves a discussion.

[7] If we formulated the proof in terms of prover-adversary games, $\mathsf{SAT}(z, F(z, B))$ would be the first formula sent by the prover.

Plausibility of provably constructive anticheckers. A theorem of Lipton and Young [41] about "anti-checkers" guarantees that if SAT is hard for circuits of size $poly(s)$, there is a set A of size $poly(s)$ such that $\mathsf{lb}_A(\mathsf{SAT}, s)$ is a tautology. In fact, a simple case analysis (considering the cases that $\mathsf{SAT} \in \mathsf{Circuit}[poly(s)]$ and $\mathsf{SAT} \notin \mathsf{Circuit}[poly(s)]$) yields a circuit F with $t = poly(s)$ such that the displayed formulas in Item 2 are tautologies, but we do not have the guarantee that the tautologies are efficiently provable. The existence of such F with the efficiently provable property from Item 2 follows, if we assume that EF proves efficiently that a concrete Boolean function is hard on average and that EF proves efficiently that a concrete efficient circuit given a circuit breaking a one-way permutation (OWP) produces a circuit for SAT, see [50, Lemma 2]. In other words, if EF can prove a circuit lower bound and reduce OWPs to $\mathsf{SAT} \notin \mathsf{P/poly}$, Item 2 holds. In fact, it holds with z being a free variable in the displayed formula.

Item 2 is similar to the assumption used in [50] to establish the implication that lower bounds on lengths of proofs in concrete propositional proof systems imply explicit circuit lower bounds. Indeed, as pointed out to us by Noel Arteche, Item 2 (with $s, t, u = poly(n)$) implies that $\mathsf{NP} \not\subseteq \mathsf{Circuit}[n^\ell]$, for each $\ell \in \mathbb{N}$ – $\mathsf{SAT} \in \mathsf{P/poly}$ combined with Item 2 implies that EF is p-bounded and $\mathsf{NP} = \mathsf{coNP}$ implies the desired lower bound using Kannan's lower bound. It is possible that Item 2 implies that EF proves efficiently $\mathsf{tt}(f, n^{O(1)})$ for some f, but it is not immediate: The outlined proof of $\mathsf{NP} \not\subseteq \mathsf{Circuit}[n^\ell]$ involves a case analysis yielding $\mathsf{tt}(\mathsf{SAT}_{n^k}, n^\ell)$ or $\mathsf{tt}(f', n^k)$ where f' has inputs of length n. Such disjunctions have, however, trivial proofs of polynomial size in the size of the formula (i.e. in $2^{O(n^k)}$). Nevertheless, Item 2 implies $poly(u)$-size EF-proofs of $\mathsf{lb}_A(\mathsf{SAT}, s) \vee \mathsf{tt}(h, s')$, for each $h \notin \mathsf{Circuit}[s']$, where the truth-table of h has length $< n$. Further, by truncating the inputs from A we get $poly(u)$-size EF-proofs of $\mathsf{tt}(h', s) \vee \mathsf{tt}(h, s')$, for some h' with the truth-table of size $\sim t$.[8]

Importantly, our theorem holds for stronger systems than EF as well – specifically, for systems EF_P from Section 1.3 – in case these are needed to satisfy Item 2. One way to interpret Theorem 11 is thus to say that if we can construct efficient adversaries for lower bounds in a sufficiently strong proof system which can prove constructively the existence of anti-checkers (e.g. by proving some circuit lower bound and reducing OWPs to $\mathsf{SAT} \notin \mathsf{P/poly}$), then we can separate the canonical pair of the system efficiently.

▶ Remark 13. Theorem 11 remains valid if the formula $\neg\mathsf{lb}_A(\mathsf{SAT}, s)$ in its conclusion is replaced by an arbitrary formula ϕ such that P has a $poly(u)$-size derivation of $\neg\mathsf{lb}_A(\mathsf{SAT}, s)|_\rho$ from ϕ, where $\psi|_\rho$ is a formula obtained from ψ by applying a substitution ρ which replaces some (possibly all) variables of ψ by $poly(s)$-size formulas built on variables of ϕ. In this case, the P-refutation is constructed as in the proof of Theorem 11 but applying ρ on formulas $\mathsf{lb}_A(\mathsf{SAT}, s)$, $\mathsf{SAT}(x, F(x, B))$ and using the fact that the assumption from Item 2 remains valid under ρ. Moreover, in Case II, we use the short P-derivation of $\neg\mathsf{lb}_A(\mathsf{SAT}, s)|_\rho$ from ϕ. Similarly, $\neg\mathsf{tt}(\mathsf{SAT}, s)$ in Theorem 11 can be replaced by an arbitrary formula ϕ such that P has a $2^{O(n)}$-size derivation of $\neg\mathsf{tt}(\mathsf{SAT}, s)|_\rho$ from ϕ, where ρ substitutes some variables by $poly(s)$-size (not $2^{O(n)}$-size) formulas built on variables of ϕ.

As noted by Moritz Müller, Theorem 11 holds also if we substitute $\neg\mathsf{lb}_A(\mathsf{SAT}, s)$ both in the conclusion and in the assumption by an arbitrary formula.

---

[8] If we assume the existence of efficient EF-proofs of $\mathsf{tt}(h', s) \vee \mathsf{tt}(h, s')$, for some h', h and suitable t, s, s', then a version of EF-natural proofs from [48] (defined as the existence of efficient EF-proofs of the formulas encoding that "EF does not prove efficiently $\mathsf{tt}(h', s) \vee \mathsf{tt}(f, s'/3)$", for many functions f) breaks superbits. In order to see this, note that short EF-proofs of $\mathsf{tt}(h', s) \vee \mathsf{tt}(h, s')$ yield short EF-proofs of $\mathsf{tt}(h', s) \vee \mathsf{tt}(h \oplus f, s'/3)$, for every $f \in \mathsf{Circuit}[s'/3]$.

3 Efficient adversaries for known lower bounds

3.1 Width and expansion

The strongest known lower bound on the length of proofs of tt-tautologies is due to Razborov [64, Theorem 2.13], who showed that (a concrete natural encoding of) $\mathsf{tt}(f_n,t)$, $n^2 \leq t \leq 2^n$, and in fact also $\mathsf{lb}_W(f_n,t)$, for any $W \subseteq \{0,1\}^n$ and any f_n, requires $\mathsf{Res}(k)$-proofs of size $2^{t^{\Omega(1)}}$, if $k = \epsilon \log t$ and $\epsilon > 0$ is a sufficiently small constant. The lower bound combines a small-restriction switching lemma, which turns short proofs into narrow ones, with a width lower bound which is obtained using expansion properties of the lb-formula (similarly as in the influential work of Ben-Sasson and Wigderson [7]). We show how to adapt the proof so that it yields efficient adversaries for formulas $\tau_W(A,b)$ described below. On the way, we will refer to the details of the proof (and definitions) from [64].

The reason why we focus on formulas τ instead of lb is that Razborov's lower bound for $\mathsf{lb}_W(f_n,t)$ is obtained by showing that every S-size $\mathsf{Res}(k)$-refutation of $\neg\mathsf{lb}_W(f_n,t)$ can be turned into an $S2^{O(k)}$-size $\mathsf{Res}(2k)$-refutation of $\tau_W(A,b)$ and then proving that $\tau_W(A,b)$ does not have such refutations. In fact, as we will observe, $\mathsf{Res}(2)$ admits $poly(|W|,t)$-size derivations of each clause of $\neg\mathsf{lb}_W(f_n,t)|_\rho$ from $\tau_W(A,b)$, where ρ substitutes some variables of $\neg\mathsf{lb}_W(f_n,t)$ by a function of at most two variables of $\tau_W(A,b)$. On the other hand, by Remark 13, Theorem 11 holds even if $\neg\mathsf{lb}_W(\mathsf{SAT},s)$ is replaced by an arbitrary formula ϕ such that EF has a $poly(s)$-size derivation of $\neg\mathsf{lb}_W(\mathsf{SAT},s)|_\rho$ from ϕ.[9] The focus on τ-formulas allows us to avoid the problem of encoding the transformation of refutations of $\neg$lb to refutations of τ in the strategy of the adversary.

In the following paragraphs defining formulas $\tau_W(A,b)$ we refer mainly to [64, Sections 6 and 8], where the reduction of lower bounds for $\neg\mathsf{lb}_W(f_n,t)|_\rho$ to lower bounds for $\tau_W(A,b)$ is described for the case $W := \{0,1\}^n$. We start with specifying formulas

$$\tau_{\leq}(A,b) =: \tau_{\{0,1\}^n}(A,b)$$

from [64], for A,b such that the desired short derivation of $\neg\mathsf{tt}(f_n,t)|_\rho$ from $\tau_{\leq}(A,b)$ exists. Formulas $\tau_{\leq}(A,b)$ are defined generally in [64, Definition 2.2] but we will consider only their particular instantiations obtained from $\tau(C^{n+1}_{A,\leq},b)$ in [64, Section 8] with two caveats:

1. We adopt the convention from the beginning of [64, Section 5] introducing axioms $\neg y^i_\emptyset$ and $y^i_{\{j\}} \vee \neg x_j, \neg y^i_{\{j\}} \vee x_j$ (instead of identifying x_j with $y^i_{\{j\}}$);
2. A,b in our $\tau_{\leq}(A,b)$ is not the same as A,b in $\tau(C^{n+1}_{A,\leq},b)$ from [64, Section 8]. In order to specify our A,b in $\tau_{\leq}(A,b)$, we need to express $\tau(C^{n+1}_{A,\leq},b)$ from [64, Section 8] as in the proof of [64, Theorem 2.12] which in turn refers to the proof of [64, Theorem 2.10], where $\tau(C^{n+1}_{A,\leq},b)$ is interpreted as a linear system $\hat{A}X = \hat{b}$ with constant right-hand side – this $\hat{A},\hat{b}$ is our A,b in $\tau_{\leq}(A,b)$ and our formula $\tau_{\leq}(A,b)$ is the formula $\tau_{\leq}(\hat{A},\hat{b})$ in the proof of [64, Theorem 2.10] obtained from $\tau(C^{n+1}_{A,\leq},b)$ considered in [64, Section 8].

The formulas $\neg\tau_{\leq}(A,b)$ express that a bit-string b of the length $|b| = 2t_0(2^{n+1}-1)$, $t_0 = \Omega((t/n)^{1/2})$, is not in the range of an NW-generator based on a $|b| \times t_0(2^{n+1}-1)$ matrix A and the parity function. In other words, $\tau_{\leq}(A,b)$ is a set of clauses asserting that $A\overline{x} = b$ is a system of linear equations solved by a $(|b|/2)$-bit vector $\overline{x}$. The matrix

[9] If $\neg\phi$ is to be hard for a strong proof system, it should be different from $\neg\tau_W(A,b)$, because systems simulating Gaussian elimination can prove $\neg\tau_W(A,b)$ efficiently.

A is an $(r, d+1)$-lossless expander, with $r = t^\delta$ for an arbitrary sufficiently small $\delta > 0$ and a sufficiently big $d = O(1)$, corresponding to an iterated composition of a function $H : \{0,1\}^{t_0} \mapsto \{0,1\}^{2t_0}$ (iteratively applied on the first t_0 and the last t_0 of its output bits) as described in [64, Section 6] (H denotes a different object in [64]). Additionally, A (is constructed from an (r,d)-lossless expander in [64, Section 8] and so it) satisfies $|J_i(A)| \geq t^\delta$ for all $i \in [|b|]$.[10] The iterated composition of functions H forms a tree T_H of $2^{n+1}-1$ copies of H. The indices of b are in one-to-one correspondence with outputs of the copies of H in T_H. The string b has 0s on all positions corresponding to outputs of the "inner" copies of H. In order to get a short derivation of $\neg\mathsf{tt}(f_n,t)|_\rho$ from $\tau_{\leq}(A,b)$ one defines $b_{a,j} := f_n(a)$, for $a \in \{0,1\}^n, j \in [2t_0]$, where $b_{a,j}$ corresponds to the jth output of the ath "outer" copy of H, i.e. the copy of H forming the ath leaf in T_H. As explained in [64, Section 8], there is a t-size circuit $D_{n,x_1,\ldots,x_{t_0}}(a)$ computing $b_{a,1} = f_n(a)$ on $a \in \{0,1\}^n$. The circuit $D_{n,x_1,\ldots,x_{t_0}}(a)$ gradually computes all copies of H in T_H that lead to $b_{a,1}$. In particular, the ath outer copy of H is computed by $D_{n,x_1,\ldots,x_{t_0}}(a)$ as well.

Given $W \subseteq \{0,1\}^n$, the formula $\tau_W(A,b)$ is obtained from restricting $\tau_{\leq}(A,b)$ to (the axioms encoding) the equations from $A\overline{x} = b$ which are used by the computation of $D_{n,x_1,\ldots,x_0}(a)$, for $a \in W$. Consequently, the size of $\tau_W(A,b)$ is $poly(t, |W|)$.

By inspecting the reduction in [64, Section 8], it can be verified that each clause of formula $\neg\mathsf{lb}_W(f_n,t)|_\rho$ has a $\mathsf{Res}(2)$ derivation of size $O(1)$ from $\tau_W(A,b)$, where ρ substitutes some variables of $\neg\mathsf{lb}_W(f_n,t)$ by a function of at most two variables of $\tau_W(A,b)$ and b is defined from f_n as above.

In the rest of this section we give a proof of the following theorem in which we construct a $poly(t^{t^\delta}, m')$-size (randomized circuits defining the responses of an) adversary for $\tau_W(A,b)$.

▶ **Theorem 14** (Efficient adversary for Razborov's $\mathsf{Res}(k)$ lower bound)**.** *Let $\gamma < 1$, $a \in \mathbb{N}$, let $\delta > 0$ be sufficiently small, $d \in \mathbb{N}$ be sufficiently big and $\epsilon > 0$. There is n_0 such that for all $n \geq n_0$ the following holds. Suppose A, b, t are as above (i.e. $n^2 \leq t \leq 2^n$, f_n specifying b is arbitrary, A is an $(r, d+1)$-lossless expander with $r = t^\delta$ and dimensions $|b| \times t_0(2^{n+1}-1)$), $k = \epsilon \log t$ and $W \subseteq \{0,1\}^n$ has size $|W| \leq t^a$. There is a $poly(t^{rad}, m')$-size randomized $\mathsf{Res}(k)$-adversary C for $\tau_W(A,b)$ such that for each $\frac{1-\gamma}{8}2^{r/(2^{O(k(d+1))})}$-size $\mathsf{Res}(k)$-refutation π of any set of clauses Φ, with probability $\geq \gamma$ over the randomness of C, π does not catch C. Here, m' refers to the input length of circuits from C.*

Let $A^-\overline{x} = b^-$ be the system of $m := |b^-|$ equations encoded by $\tau_W(A,b)$. That is, A^- results from A by erasing some rows and b^- is the corresponding restriction of b. Since $m \leq |\tau_W(A,b)| = poly(t, |W|)$ and $|W| \leq t^a$, such A^- satisfies

$$|J_i(A^-)| = t^\delta \geq C'_\gamma(d+1)(k + \log m), \text{ for all } i \in [m], \tag{1}$$

for arbitrary constant C'_γ, if n is sufficiently big.

On high level, the strategy of our adversary is to transform π using a transformation R and a random restriction ρ into a new $\mathsf{Res}(O(k))$-refutation which is practically equivalent to π restricted by ρ, but which contains only DNFs with small height and thus can be simulated by a Res-refutation of a small width. The small height/width will allow the adversary to find a partial solution of $A^-\overline{x} = b^-$ and respond to any formula from the original proof π so that each axiom of $\tau_W(A,b)$ is assigned 1 and the consistency w.r.t. $\mathsf{Res}(k)$-derivation rules is maintained (implying that π will not catch the adversary). We proceed with a more formal description.

[10] The existence of such A follows from the proof of [64, Theorem 2.5] and an application of Chernoff's bound, if $t \geq n_0$ for some n_0 depending on δ and d.

Moving to protected DNFs. Given a k-DNF F from π, the adversary first sets all variables in F which do not appear in $\tau_W(A,b)$ to 0 and applies the transformation R from the proof of [64, Claim 5.8] which produces a "protected" $O(k(d+1))$-DNF $R(F)$. However, our adversary adjusts the parameters in order to accommodate a more "sparse" ρ. Specifically, in [64, Definition 5.3] in Item 2, $3d$ is increased to $5d$ and in Item 3, $20kd$ is increased to $32kd$.[11] Further, in [64, Definition 5.4], $6d$ is increased to $10d$.[12] Consequently, $14kd$ in [64, Claim 5.5 a)] becomes $22kd$. In the proof of [64, Claim 5.7], $3d$ becomes $5d$ and $6kd$ becomes $10kd$. Recall also that d in [64, Definitions 5.3 & 5.4] is $d+1$ in our setting of parameters.

R (together with assigning 0's outside τ) transforms π into a set of $O(k(d+1))$-DNFs which can be completed into a $|\pi|2^{O(k(d+1))}$-size $\mathsf{Res}(O(k(d+1)))$-refutation of some Φ' such that each axiom of $\tau_W(A,b)$ gets converted to a DNF which logically follows from $\tau_W(A,b)$. The completion exists by the proof of [64, Claim 5.8] and by the proof of [64, Claim 5.7] (which guarantees that for each axiom F' in $\tau_W(A,b)$, we have $F' \Leftrightarrow R(F')$ assuming $\tau_W(A,b)$), but our adversary will construct only a part of it – we postpone the details. The definition of R exploits the property (1), which as we already explained can be assumed to hold for the relevant submatrix A^- of A.

The transformation R can be implemented by $poly(t, 2^{k(d+1)}, |W|, m')$-size circuits. This follows straightforwardly by inspection of its definition. We provide A^- as a non-uniform advice of the adversary. The occurrence of $|W|$ is enforced by the necessity of generating literals $(y^i_{J_i(A)})^{b_i}$ (referred to as "cosmetic") as the only way for our adversary to access bits b_i is to have the bits $b_{u,1}$ with $u \in W$ stored in a non-uniform advice.[13]

Applying a random restriction. Next, the adversary applies the random restriction ρ from [64, Definition 5.9], but the parameter $2d$ in [64, Definition 5.9] is increased to $4d$ – again, note that in our setting d in [64, Definition 5.9] becomes $d+1$. In more detail, if $R(F)$ contains a variable y^i_Σ, the adversary constructs $\tilde{Y}_i$ and a $poly(t_0)$-size part of J needed to determine Y_i (the part of J in $J_i(A)$). After determining Y_i, the adversary constructs the part of ρ needed to determine $R(F)|_\rho$ (using bits $b_{u,1}$ again to assign the cosmetic variable $y^i_{J_i(A)}$). Similarly, if $R(F)$ contains a variable x_j. The fact that circuits $C_{i,j}$ from our adversary share a random string is used to ensure that in future rounds our adversary will re-produce the same ρ which was used in previous rounds.

In the definition of sparsity, [64, Definition 5.1], $2d$ is increased to $4d$. Then, by [64, Claim 5.10], the probability that ρ is sparse is $\geq 1/2$, but the same proof shows that the probability is $\geq 1-(1-\gamma)/2$, if C'_γ is big enough. Here we use (1) again. To see that the proof of [64, Claim 5.10] goes through even with our adjusted parameters, it suffices to implement the following changes in it: $9d$ becomes $18d$, $s \geq C \log m/9$ becomes $s \geq C'_\gamma \log m/18$, $4d$ becomes $8d$ and $|\Delta^c_v| = d$ becomes $|\Delta^c_v| = 2d$.

As explained in the proof of [64, Claim 5.2], for any sparse ρ, $\tau_W(A,b)|_\rho$ can be transformed using a variable substitution ρ' into clauses encoding a linear system of equations $A^+\overline{x} = b^+$, where A^+ is an $(r, d+1)$-lossless expander with $|J_i(A^+)| \geq 4(d+1)$. (If we did not adjust the parameters above, we would get only $|J_i(A^+)| \geq 2(d+1)$, which would not suffice to prove Lemma 15 guarantying the existence of suitable partial solutions of $A^+\overline{x} = b^+$.)

[11] For clarification, we remark that $L_{i,1}, R_{i,1}$ in Item 1 of [64, Definition 5.3] denote L_{i1}, R_{i1} and $Conv(L_{i\nu}, R_{i\nu})$ in Item 3 is a subset of $J_i(A)$.

[12] Again, in order to avoid a potential ambiguity, we stress that intervals $(r_{i,\nu}, r_{i,\nu+1}]$ in [64, Definition 5.4] are subsets of $J_i(A)$.

[13] For this reason we would get at best only $poly(t^{t^\delta}, 2^n, m')$-size $\mathsf{Res}(k)$-adversaries for $\neg\mathsf{tt}(f_n, t) = \neg\mathsf{lb}_W(f_n, t)$, if $W = \{0,1\}^n$.

[64, Claim 5.14] remains valid for our adjusted parameters as well. To see that, it suffices to increase $6d$ to $10d$ in the proof of [64, Claim 5.12] and $2d$ to $4d$ in the proof of [64, Claim 5.13]. Therefore, by [64, Claim 5.14], ρ collapses the height of $R(F)$ with high probability,

$$\Pr_{\rho}[h(R(F)|_{\rho}) \geq h] \leq 1/e^{h/2^{O(k(d+1))}},$$

where $h := r/(C'_h k(d+1))$, for a big enough constant C'_h. Since $4|\pi| \leq (1-\gamma)2^{h/2^{O(k(d+1))}}/2$, with probability $\geq 1-(1-\gamma)/2$, every $F \in \pi \cup \pi'$ was transformed to a formula $R(F)|_{\rho}$ with height $\leq h$. Here, π' is a set of k-DNFs of size $|\pi'| \leq 3|\pi|$ which will be specified later. The adversary will produce formulas in π' from the formulas in π it receives and the transformation R will be defined on π'.

The last three paragraphs imply that with probability $\geq \gamma$, the applications of R (with assigning 0s outside τ) and ρ transform π into (a set of formulas that can be completed to) a $\mathsf{Res}(O(k(d+1)))$-refutation of some Φ'' such that each axiom of $\tau_W(A,b)|_{\rho}$ is converted to a DNF which follows from axioms which can be further interpreted (using ρ') as axioms encoding $A^+\overline{x} = b^+$, where A^+ is an $(r, d+1)$-lossless expander satisfying $|J_i(A^+)| \geq 4(d+1)$, for all $i \in [|b^+|]$. Additionally, $\forall F \in \pi \cup \pi'$, $h(R(F)|_{\rho}) \leq h$.

In the rest of the proof we define the responses of the adversary in the case that R and ρ ensured the properties above. If this does not happen and the adversary is led to respond to an axiom of $\tau_W(A,b)$ by 0, it recognizes that the transformation failed and (despite failing in this case) overwrites its response to 1 in order to ensure that it is *for* $\tau_W(A,b)$.

Finding partial solutions. Our adversary already produced $R(F)|_{\rho}$, for a given k-DNF F. To complete the proof of Theorem 14 it remains to explain how the adversary solves the system of equations corresponding to $R(F)|_{\rho}$ in a consistent fashion (so that it responds to each axiom of $\tau_W(A,b)$ by 1 and preserves the $\mathsf{Res}(k)$-derivation rules). For this purpose we adapt a proof of a width lower bound presented by Krajíček [34, Lemma 13.4.5]. Our proof differs from Krajíček's presentation in that instead of using $(r,\cdot)$-expansion we use $(r, d+1)$-lossless expansion.

For a set of rows $I \subseteq [|b^+|]$, let A^+_I be the matrix obtained from A^+ by deleting all rows in I and all columns in $\bigcup_{i\in I} J_i$, where $J_i := J_i(A^+)$.

▶ **Lemma 15.** *Let A^+ be an $(r, d+1)$-lossless expander as above satisfying $|J_i| \geq 4(d+1)$, for all $i \in [|b^+|]$. For any set of rows $I \subseteq [|b^+|]$ of size $|I| \leq r/2$, there is $\hat{I} \supseteq I$, $|\hat{I}| \leq 2|I|$ such that*

$$\text{For each } i \notin \hat{I}, \quad \left| J_i \setminus \bigcup_{u\in\hat{I}} J_u \right| \geq |J_i| - 2(d+1). \tag{2}$$

Moreover, for any $\hat{I}$ satisfying (2), for any nonempty set K of $\leq r$ rows in $A^+_{\hat{I}}$, $|\partial_{A^+_{\hat{I}}}(K)| > 0$.

Proof. Assume $|I| \leq r/2$. Put $I_0 := I$. Add in consecutive steps $t = 0, \ldots$ to I_t any one row i as long as

$$\left| J_i \cap \bigcup_{u\in I_t} J_u \right| > 2(d+1). \tag{3}$$

We claim that the process stops before t reaches $|I|$. Assume not, and let $I' := I_{|I|}$. Then by (3),

$$\sum_{i\in I'} |J_i| - \partial_{A^+}(I') > \sum_{i\in I'} |J_i| - \left(\sum_{i\in I} |J_i| + \sum_{i\in I'\setminus I} (|J_i| - 2(d+1))\right) = \sum_{i\in I'\setminus I} 2(d+1) = (d+1)|I'|.$$

This contradicts the expansion property of A^+.

Let $\hat{I}$ be the last I_t in the process, so $t < r/2$, $|\hat{I}| \le 2|I|$ and $\hat{I}$ has the property (2).

For the "moreover" part, note that

$$\partial_{A^+_{\hat{I}}}(K) = \partial_{A^+}(K) \setminus \bigcup_{i\in\hat{I}} J_i.$$

As for all $i \in K\setminus\hat{I}$ we have $|J_i \cap \bigcup_{u\in\hat{I}} J_u| \le 2(d+1)$, this implies that

$$|\partial_{A^+_{\hat{I}}}(K)| \ge |\partial_{A^+}(K)| - \sum_{i\in K} 2(d+1) \ge \sum_{i\in K} |J_i| - (d+1)|K| - 2(d+1)|K| > 0.$$

The last inequality uses $|J_i| \ge 4(d+1)$. ◀

Any $I \subseteq [|b^+|]$ satisfying (2) from Lemma 15 will be referred to as a *safe set of rows*. A partial assignment α to $\overline{x}$ assigning to 0 or 1 variables from a set $dom(\alpha)$ is *safe with support* I if $\{j \mid x_j \in dom(\alpha)\} \subseteq \bigcup_{i\in I} J_i$ for some safe I. A safe partial assignment α with support I is a *safe partial solution* of $A^+x = b^+$ *with support* I if and only if $\bigoplus_{j\in J_i} \alpha(x_j) = b^+_i$, for all $i \in I$.

▶ **Lemma 16.** *Assume $I \subseteq I' \subseteq [|b^+|]$ are safe sets and $|I'\setminus I| \le r$. Assume further that α is a safe assignment with support I, and that $c_i \in \{0,1\}$, for $i \in I'\setminus I$.*

Then the assignment α can be extended to a safe assignment α' with support I' such that

$$\bigoplus_{j\in J_i} \alpha'(x_j) = c_i,$$

for all $i \in I'\setminus I$.

Proof. By Lemma 15, each subset of $I'\setminus I$ has a non-empty border in A^+_I. In particular, no subset of $I'\setminus I$ can be a linearly dependent set of rows of A^+_I. Hence the system

$$\bigoplus_{j\in J_i\setminus dom(\alpha)} x_j = c_i \oplus \bigoplus_{j\in J_i\cap dom(\alpha)} \alpha(x_j), \text{ for } i \in I'\setminus I,$$

has a solution θ. Put $\alpha' := \alpha \cup \theta$. ◀

The last lemma we need is a result of Segerlind, Buss and Impagliazzo [68, Theorem 5.1] which translates Res(k)-derivations with lines of small height to Res-derivations of small width. Our formulation is slightly more general but it follows directly from the proof presented in [68]. Given a k'-DNF M of height $h(M) \le h'$, denote by $\mathcal{M}^{h'}_M$ the set of clauses M' of width h' built on variables of M such that $\neg M'$ falsifies all terms in M.

▶ **Lemma 17** ([68])**.** *If there is a* Res(k')*-derivation of a k'-DNF M_0 that has height $h(M_0) \le h'$ from a set of k'-DNFs $\mathcal{M}$ such that each $M_1 \in \mathcal{M}$ has height $h(M_1) \le h'$ and each k'-DNF M in the derivation has height $h(M) \le h''$, where $h'' \ge h'$, then every clause $M'_0 \in \mathcal{M}^{h'}_{M_0}$ can be derived from $\bigcup_{M_1\in\mathcal{M}} \mathcal{M}^{h'}_{M_1}$ by a* Res*-derivation of width $\le h''k'$ (possibly using the weakening rule).*

We are finally ready to finish the proof of Theorem 14.

In each round $j \geq 1$, given the jth k-DNF F, our adversary will construct an h-CNF $\overline{F} := \mathcal{M}^h_{R(F)|_\rho}$ equivalent to $R(F)|_\rho$ by going through all possible clauses M of width h and adding M to $\overline{F}$ if $\neg M$ falsifies all terms in $R(F)|_\rho$. The h-CNF exists and the construction succeeds because $h(R(F)|_\rho) \leq h$. Since there are only $poly(t, |W|)$ variables in $R(F)|_\rho$ (recall that all variables outside $\tau_W(A, b)$ are automatically set to 0), $\overline{F}$ can be constructed by a $poly((t + |W|)^h, |R(F)|_\rho|)$-size circuit. Further, the adversary will produce $\alpha_1, I_1, \ldots, \alpha_j, I_j$, where $\alpha_{j'}$, $j' \in [j]$, is a safe partial solution of $A^+\overline{x} = b^+$ with support $I_{j'}$ of size $|I_{j'}| \leq 2h$. The partial solution α_j will be used to evaluate F: either a clause in $\overline{F}$ will be falsified by $\overline{\rho \circ \alpha_j}$, in which case the response of the adversary to F is 0, or otherwise the response to F is 1. Here (and in the rest of the proof) we abuse the notation and use ρ' to identify variables of $R(F)$ with variables encoding $A^+\overline{x} = b^+$ – note that the relevant part of ρ' can be computed by $poly(t, |R(F)|, |W|, m')$-size circuits from $R(F)$ (the access to m' ensures that the adversary can use new random bits when constructing ρ). The partial assignment $\overline{\rho \circ \alpha_j}$ is the extension of $\rho \circ \alpha_j$ which assigns also the variables y^i_Σ whose values are determined by $\rho \circ \alpha_j$ and $A^+\overline{x} = b^+$.

Response to axioms. Let us first observe that the response of such an adversary to each axiom F in $\tau_W(A, b)$ in each round $j \geq 1$ will be 1. Denote the system of equations of $A^+\overline{x} = b^+$ corresponding to rows from a set $I \subseteq [|b^+|]$ by $A^+\overline{x} = b^+|_I$. Let I^+ be a minimal set of rows in $[|b^+|]$ such that all variables of $R(F)|_\rho$ are in the clauses encoding the equations $A^+\overline{x} = b^+|_{I^+}$. As F has a constant width, [64, Claim 5.7 a)] implies that $|I^+| \leq O(k(d+1))$ and $|I_j \cup I^+| \leq 2h + O(k(d+1)) \leq r/2$. By Lemma 15, there is thus an extension of $I_j \cup I^+$ to a safe set of size $\leq r$ and by Lemma 16 there is a solution $\alpha \supseteq \alpha_j$ of $A^+\overline{x} = b^+|_{I_j \cup I^+}$. A closer look at the definition reveals that $R(F)$ results from rewriting some variables $y^i_{\Sigma^i_\nu}$ in F by linear forms in x_ℓ's which these variables $y^i_{\Sigma^i_\nu}$ represent according to the i'th equation in $\tau_W(A, b)$ and the sign of the "anchor" variable. It is not hard to check that $A^+\overline{x} = b^+|_{I^+}$ implies that the $y^i_{\Sigma^i_\nu}$'s (whose assignment is now determined by $\rho \circ \alpha$) are equivalent to these linear forms. More precisely, if $y^i_{\Sigma^i_{\nu_{i\beta}}}$ is the anchor variable and $\nu > \nu_{i\beta}$, then $A^+\overline{x} = b^+|_{I^+}$ implies

$$y^i_{\Sigma^i_\nu} \oplus \bigoplus_{j \in \Sigma^i_{\nu_+} \setminus \Sigma^i_\nu} x_j = \rho(y^i_{\Sigma^i_{\nu_+}}) = \rho(y^i_{\Sigma^i_{\nu_-}}) \oplus \bigoplus_{j \in (\Sigma^i_{\nu_+} \setminus \Sigma^i_{\nu_-}) \cap J_x(\rho)} \rho(x_j) \oplus \bigoplus_{v \in (\nu-, \nu+]} (b|_\rho)_{i,v}$$
$$= y^i_{\Sigma^i_{\nu_{i\beta}}} \oplus \bigoplus_{j \in \Sigma^i_{\nu_+} \setminus \Sigma^i_{\nu_{i\beta}}} x_j,$$

where $\Sigma^i_\nu \subseteq \Sigma^i_{\nu_+}$ is the minimal initial segment of $J_i(A)$ such that $y^i_{\Sigma^i_{\nu_+}} \in sup(\rho)$ and $\Sigma^i_{\nu_-} \subseteq \Sigma^i_{\nu_{i\beta}}$ is the maximal initial segment of $J_i(A)$ such that $y^i_{\Sigma^i_{\nu_-}} \in sup(\rho)$, so $y^i_{\Sigma^i_\nu} = y^i_{\Sigma^i_{\nu_{i\beta}}} \oplus \bigoplus_{j \in \Sigma^i_\nu \setminus \Sigma^i_{\nu_{i\beta}}} x_j$. If $\nu < \nu_{i\beta}$, we get the same equations but $\Sigma^i_{\nu_+}$ is defined so that $\Sigma^i_{\nu_{i\beta}} \subseteq \Sigma^i_{\nu_+}$ and $\Sigma^i_{\nu_-}$ so that $\Sigma^i_{\nu_-} \subseteq \Sigma^i_\nu$, which yields $y^i_{\Sigma^i_\nu} = y^i_{\Sigma^i_{\nu_{i\beta}}} \oplus \bigoplus_{j \in \Sigma^i_{\nu_{i\beta}} \setminus \Sigma^i_\nu} x_j$. Since $\overline{\rho \circ \alpha}$ satisfies F, $R(F)|_{\overline{\rho \circ \alpha_j}}$ is satisfiable and the adversary responds to F by 1.

Round 1 and Weakening. The first k-DNF our adversary receives is the empty clause and $\alpha_1 := \emptyset, I_1 := \emptyset$. Recall the derivation rules of $\mathsf{Res}(k)$ from [64, Definition 2.6]. If the jth DNF F is a premise of an application of the weakening rule whose conclusion is a DNF $F_{j'}$ to which the adversary responded by 0 in round $j' < j$, the adversary recognizes that $F_{j'}$ is derived using the weakening rule or the CUT-rule. Even if the rule is CUT, it pretends that it is Weakening and sets $\alpha_j := \alpha_{j'}, I_j := I_{j'}$. As $\overline{F_{j'}} \subseteq \overline{F}$, this guarantees the consistency w.r.t. the rule.

Round $j > 1$ for CUT, AND-elimination and AND-introduction. Suppose that the jth k-DNF $F \vee T$ is a premise of the CUT-rule, AND-elimination or AND-introduction whose conclusion is a DNF $F \vee G$ to which the adversary responded by 0 in round $j' < j$. Further, assume that the situation is not consistent with an application of the weakening rule. If the situation is consistent with an application of AND-elimination, the adversary pretends that the rule is AND-elimination. If this is not the case and $T = G$ is a single literal, the adversary pretends that the rule is AND-introduction with a single premise. In the remaining cases, the adversary does not necessarily recognize if $F \vee G$ is derived using CUT or AND-introduction. If this happens, it simulates both alternatives. For AND-introduction, the adversary knows all premises of the derivation step, which is not necessarily the case for CUT. When simulating CUT, the adversary pretends that the premises are $F \vee G \vee T$ and $F \vee G \vee \neg T$. The formulas $F \vee G \vee T$ and $F \vee G \vee \neg T$ – which are such that the transformation R is defined on them – are added to π'. This adds to π' at most two new formulas for each CUT in π, i.e. $\leq 2|\pi|$ formulas in total. Denote the set of the premises of the simulated rule (as pretended by the adversary) by Γ.

As explained in the proof of [64, Claim 5.8], the R-image of the $\mathsf{Res}(k)$-derivation step concluding $F \vee G$ can be completed to a $\mathsf{Res}(O(k(d+1)))$-derivation Π using the axioms of $\tau_W(A, b)$. [14] Now we finish the specification of π' by including in it also the non-principal formulas of each derivation step in π except CUTs (the non-principal formulas of simulated CUTs are already in π'). Here, a non-principal formula of a derivation step is the biggest subformula in its conclusion not intersecting the principal formula. This adds to π' at most $|\pi|$ formulas, making the total size of π' at most $|\pi'| \leq 3|\pi|$. Given that the height of the $R|_\rho$-image of the non-principal formula of the simulated rule is $\leq h$, it is easy to check that height of each formula in $\Pi|_\rho$ is $\leq h + O(k(d+1))$. Hence, by Lemma 17, the clause in $\overline{F \vee G}$ falsified by $\overline{\rho \circ \alpha_{j'}}$ can be derived from $\overline{\Phi_\Gamma}$'s, for $\Phi_\Gamma \in \Gamma$, and from $\tau_W(A, b)|_\rho$ by a Res-derivation Π' of width $w \leq O(k(d+1)h + (k(d+1))^2) \leq O(k(d+1)h)$. The adversary can construct Π' by listing all possible clauses of width $\leq w$ (there are $poly((t + |W|)^w)$ of them) and gradually deriving every clause of width $\leq w$ that can be derived from $\overline{\Phi_\Gamma}$'s, for $\Phi_\Gamma \in \Gamma$, and from $\tau_W(A, b)|_\rho$ in $poly((t + |W|)^w)$ steps.

Next, the adversary simulates its evaluation on Π'. In each step i of the evaluation, it stands on a clause $M_i \in \Pi'$ of width $\leq w$ and holds a safe partial solution $\alpha_{j',i}$ of $A^+\overline{x} = b^+$ with a support $I_{j',i}$ of size $|I_{j',i}| \leq 2w$ such that $\overline{\rho \circ \alpha_{j',i}}$ falsifies M_i. Initially $\alpha_{j',0} := \alpha_{j'}$, $I_{j',0} := I_{j'}$ and M_0 is the clause in $\overline{F \vee G}$ falsified by $\overline{\rho \circ \alpha_{j'}}$. If the adversary wants to move to a predecessor of M_i in Π', it identifies the variable ℓ which was resolved in the step in Π' concluding M_i and a row r_0 in $A^+\overline{x} = b^+$ such that ℓ occurs in the clauses encoding $A^+\overline{x} = b^+|_{\{r_0\}}$. Since $|I_{j',i}| + 1 \leq r/2$, for a sufficiently big C'_h, by Lemma 15, there exists a safe set of rows $\hat{I}_{j',i} \supseteq I_{j',i} \cup \{r_0\}$ of size $|\hat{I}_{j',i}| \leq r$. As $|J_u(A^+) \cap J_{u'}(A^+)| > 1$ only if u, u' are from the same copy of A_0 (in the construction of A^+), the algorithm from the proof of Lemma 15 producing $\hat{I}_{j',i}$ can be implemented by a circuit of size $poly(t_0)$. By Lemma 16 there exists a solution $\hat{\alpha}_{j',i} \supseteq \alpha_{j',i}$ of $A^+\overline{x} = b^+|_{\hat{I}_{j',i}}$. Our adversary finds $\hat{\alpha}_{j',i}$ efficiently using Gaussian elimination (and the bits $b_{u,1}$ with $u \in W$ stored in the non-uniform advice). The adversary then proceeds to the predecessor $M_{i+1} \in \Pi'$ of M_i falsified by $\overline{\rho \circ \hat{\alpha}_{j',i}}$.

[14] Note that terms $R((y^i_{J_i(A)})^{b_i})$, which are by definition empty (by the proof of [64, Claim 5.8] R is defined on literals and by the proof of [64, Claim 5.7 a)] cosmetic literals are removed even if they appeared in the original term), are treated as valid and efficiently derivable. We prefer to set $R((y^i_{J_i(A)})^{b_i}) := (y^i_{J_i(A)})^{b_i}$ and $R(\neg y^i_\emptyset) := \neg y^i_\emptyset$. Formally, this is a new transformation R' which might not output a protected DNF, but whenever $R(F) \neq R'(F)$, $R'(F)|_\rho$ is trivialized because $\rho((y^i_{J_i(A)})^{b_i}) = \rho(\neg y^i_\emptyset) = 1$.

Since M_{i+1} has width $\leq w$, there exists a set $I^w_{j',i} \subseteq \hat{I}_{j',i}$ of w rows such that $\overline{\rho \circ \beta'}$ falsifies M_{i+1}, where β' is the restriction of $\hat{\alpha}_{j',i}$ to the variables with indices in $\bigcup_{i \in I^w_{j',i}} J_i$. The adversary can find $I^w_{j',i}$ efficiently. Finally, the adversary constructs the safe set of rows $I_{j',i+1} \supseteq I^w_{j',i}$, $I_{j',i+1} \subseteq \hat{I}_{j',i}$, of size $|I_{j',i+1}| \leq 2w$ using the algorithm from the proof of Lemma 15 again and sets $\alpha_{j',i+1}$ to be the restriction of $\hat{\alpha}_{j',i}$ to the variables with indices in $\bigcup_{i \in I_{j',i+1}} J_i$. If M_{i+1} is an initial clause of Π', it sets $\alpha_j := \alpha_{j',i+1}$ and $I_j := I_{j',i+1}$. Note that in such case, $|I_j| \leq 2h$ can be ensured.

The adversary cannot arrive in an axiom of $\tau_W(A, b)$, it arrives in a clause of $\overline{\Phi_\Gamma}$, for some $\Phi_\Gamma \in \Gamma$. Hence, if the simulated rule was AND-elimination or AND-introduction with a single premise, the adversary responds by 0 to $F \vee T$ and remains consistent. If the rule was CUT (but not Weakening) or AND-introduction with more premises and in at least one of these cases the simulation ended up in a clause of $\overline{F \vee T}$, the adversary outputs 0. The simulation of a rule is independent of the premise received by the adversary and in the case of CUT we have $\overline{F \vee G \vee T} \subseteq \overline{F \vee T}$. Therefore, the adversary responds by 0 to at least one of the premises and remains consistent w.r.t. every rule.

Efficiency. The least inefficient part of the algorithm implemented by our adversary was the construction of the Res-derivation Π' which took $poly(t^{rad}, m')$ steps. In each round j, the adversary uses a circuit of size $j \cdot poly(t^{rad}, m') \leq m' \cdot poly(t^{rad}, m')$ to produce $\alpha_{j'}, I_{j'}$, for all $j' \in [j]$. The complexity of the adversary is thus $poly(t^{rad}, m')$ and the proof of Theorem 14 is completed.

3.2 Pseudowidth

We present a constructive version of the lower bound for the weak pigeonhole principle PHP^m_n based on the method of pseudowidth, cf. [54, 61, 62, 63, 20]. The lower bound proceeds in two steps. In the first step, a supposed short Res refutation of PHP^m_n is transformed into a Res refutation of small pseudowidth. In the second step, it is shown that the resulting small pseudowidth Res refutation cannot exist. In order to make the lower bound constructive, we modify the second step.

$\neg\mathsf{PHP}^m_n$ is a CNF formula using variables $x_{i,j}$, $i \in [m], j \in [n]$. The intended meaning of $x_{i,j}$ is that the pigeon i is mapped to the hole j. $\neg\mathsf{PHP}^m_n$ consists of the following clauses:

- (pigeon axioms) $\bigvee_{j \in [n]} x_{i,j}$, for $i \in [m]$;
- (hole axioms) $\neg x_{i,j} \vee \neg x_{i',j}$, for $i \neq i' \in [m], j \in [n]$.

For a clause C and pigeon i,

$$N_C(i) := \{j \in [n] \mid C \text{ is satisfied if } x_{i,j} = 1 \wedge \bigwedge_{j' \neq j} x_{i,j'} = 0\}.$$

Further, define $\Delta_C(i) := |N_C(i)|$.

Let $d = (d_1, \ldots, d_m)$ be a vector such that $d_i \in \mathbb{N} \setminus \{0\}, \delta \in \mathbb{R}$ and $d_i > \delta$ for all $i \in [m]$. Then

$$P_d(C) := \{i \in [m] \mid \Delta_C(i) \geq d_i\},$$
$$P_{d,\delta}(C) := \{i \in [m] \mid \Delta_C(i) \geq d_i - \delta\}.$$

The pseudowidth of C is $w_{d,\delta}(C) := |P_{d,\delta}(C)|$. For a set of clauses $\mathcal{A}$ and a Res refutation π, the pseudowidth of $\pi \setminus \mathcal{A}$ is $w_{d,\delta}(\pi \setminus \mathcal{A}) := max_{C \in \pi \wedge C \notin \mathcal{A}} w_{d,\delta}(C)$. A clause C such that $|P_d(C)| \geq w_0$ is referred to as a (w_0, d)-*subaxiom*.

In this section, we include the weakening rule in Res, i.e. a clause C can be derived from a clause D if $D \subseteq C$. We denote the number of clauses in a Res refutation π by L_π.

The first step of the lower bound based on the pseudowidth method is captured in the following lemma. We formulate it in a stronger way than usual: Instead of concluding the existence of a vector d, we claim that a random vector has the desired properties. This is, in fact, what the standard proof of the lemma shows, cf. [20, Lemma 4.2].

▶ **Lemma 18** (Pseudowidth reduction). *Let $\delta := n/\log m$, $3 < m, n \in \mathbb{N}$. There is a distribution of m-dimensional integer vectors $\boldsymbol{d}$ depending only on m, n such that each vector $d = (d_1, \ldots, d_m)$ sampled according to $\boldsymbol{d}$ satisfies $\delta < d_i \leq n$, for all $i \in [n]$, and such that for each Res refutation π of a set of clauses Φ built on variables of PHP^m_n, for each $w_0 \in [m]$ and a random vector $d = (d_1, \ldots, d_m)$ sampled according to $\boldsymbol{d}$, with probability $\geq 1 - 1/L_\pi$, we have a set $\mathcal{A}$ of (w_0, d)-subaxioms with $|\mathcal{A}| \leq L_\pi$ and a Res refutation $\pi' \subseteq \pi$ of $\Phi \cup \mathcal{A}$ satisfying*

$$w_{d,\delta}(\pi' \setminus \mathcal{A}) = O(w_0 + \log L_\pi).$$

▶ **Lemma 19** (Pseudowidth lower bound). *Let $m, n, r \in \mathbb{N} \setminus \{0\}$ and $4 \leq \delta \in \mathbb{R}$. Suppose that $d = (d_1, \ldots, d_m)$ is an integer vector such that $\delta < d_i \leq n$, for all $i \in [m]$. Further, let $\mathcal{A}$ be an arbitrary set of $(1, d)$-subaxioms such that $\mathcal{A} \cap \neg\mathsf{PHP}^m_n = \emptyset$. Then for every Res refutation π of $\neg\mathsf{PHP}^m_n \cup \mathcal{A}$ satisfying $L_\pi < 2^{r\delta/8n}$, we have*

$$w_{d,\delta}(\pi \setminus \mathcal{A}) > \delta/6r.$$

Proof. For the sake of contradiction assume that π is a Res refutation of $\neg\mathsf{PHP}^m_n \cup \mathcal{A}$ with $w_{d,\delta}(\pi \setminus \mathcal{A}) \leq \delta/6r$. W.l.o.g. $|\mathcal{A}| \leq L_\pi$ since all unused $(1, d)$-subaxioms can be erased from $\mathcal{A}$. For the simplification of the notation we assume that for all $C \in \mathcal{A}$ we have $P_{d,\delta}(C) = P_d(C)$ and $P_d(C)$ consists of a single pigeon. (If there are more pigeons in $P_{d,\delta}(C)$, we choose one from $P_d(C)$ in an arbitrary way.)

To reach a contradiction we trace π from its last clause to an axiom. During the trace, when we are at a clause C, we maintain a partial multi-matching ρ_C on $P_{d,\delta}(C)$ which does not satisfy C. For each pigeon i from the domain of ρ_C, ρ_C sets $x_{i,j} = 1$ for r not necessarily different j's and $x_{i,j} = 0$ for all the remaining j's. Additionally, for all $i \neq i'$ in the domain of ρ_C, for all j, ρ_C sets $x_{i,j} = 0$ or $x_{i',j} = 0$. Each ρ_C will be generated using a randomized procedure.

If we reach an axiom C, we are in one of the following three cases:

1. C is a hole axiom $\neg x_{i,j} \vee \neg x_{i',j}$: Then $i, i' \in P_{d,\delta}(C)$, so ρ_C maps both i, i' into different holes and C is satisfied.
2. C is a pigeon axiom $\bigvee_j x_{i,j}$: Then $i \in P_{d,\delta}(C)$, so ρ_C maps i into some hole and C is satisfied.
3. $C \in \mathcal{A}$.

The desired contradiction will be obtained by showing that with high probability over the randomness used to generate ρ_C, $C \in \mathcal{A}$ is satisfied by ρ_C.

Denote the assignment that we have at the ℓth step by ρ_ℓ. Initially, ρ_1 is empty. Assume that in step $\ell > 1$ we have $\rho_\ell = \rho_C$ and that C_0, C_1 are the predecessors of C in π, i.e. C is derived by resolving on C_0, C_1. First, we extend ρ_ℓ to ρ'_ℓ in the following way:

- For each $i \in (P_{d,\delta}(C_0) \cup P_{d,\delta}(C_1)) \setminus P_{d,\delta}(C)$, choose r (not necessarily different) holes $j_1, \ldots, j_r$ uniformly at random from $[n] \setminus (N_C(i) \cup R)$, map i to $j_1, \ldots, j_r$ (i.e. set $x_{i,j_1}, \ldots, x_{i,j_r}$ to 1 and the remaining $x_{i,j}$'s to 0) and add $j_1, \ldots, j_r$ to R. Here, R is initially the set of holes occupied by pigeons assigned by ρ_ℓ.

During the construction of ρ'_ℓ, the size of R does not exceed $3wr \leq \delta/2$, where $w := w_{d,\delta}(\pi \setminus \mathcal{A}) \geq 1$.[15] Moreover, as $i \notin P_{d,\delta}(C)$, we have $\Delta_C(i) < d_i - \delta$, which leaves at least $n - 3wr - d_i + \delta \geq 1$ free holes for each pigeon. So the construction of ρ'_ℓ is well-defined. Since i was mapped to j's such that $j \notin N_C(i)$, the resulting ρ'_ℓ does not satisfy C and thus it does not satisfy at least one of the clauses C_0, C_1. We proceed to the clause which is not satisfied, w.l.o.g. it is C_0, and define $\rho_{\ell+1} = \rho_{C_0}$ as ρ'_ℓ restricted to $P_{d,\delta}(C_0)$. If C is derived via the weakening rule, the construction of $\rho_{\ell+1}$ is analogous. The partial multi-matchings ρ_C satisfy the required properties except that we need to show that with high probability, $C \in \mathcal{A}$ is satisfied by $\rho_\ell = \rho_C$ for $\ell \leq L_\pi$.

Suppose $D \in \mathcal{A}$. We want to bound the probability that $\exists t < \ell, E_t^D$ and D is not satisfied by $\rho_\ell = \rho_C$ at a step $\ell \leq L_\pi$, where E_t^D is the event that the pigeon $i \in P_d(D)$ is assigned in the process of generating ρ_{t+1} from ρ_t.

Consider ρ_t such that E_t^D takes place. During the construction of ρ_{t+1}, while assigning the pigeon i the r'th hole $j_{r'}$, $r' \leq r$, we have at least $n - 3wr - d_i + \delta$ possible choices among the free holes. Denote the set of free holes by $H^{r'}_{\text{free}}$. On the other hand, since $i \in P_d(D)$, there are at least $d_i - 1$ different holes such that sending i to one of these holes will satisfy D (If sending i to a hole j_i satisfies D because $x_{i,j'} = 0$ for some $j' \neq j_i$, the hole j_i becomes "blocked" and sending i to the remaining free holes will still satisfy D even as a multi-matching). Denote the set by $H^{r'}_{\text{sat}}$. Therefore, the probability of the event $B^D_{\rho_t}$ that assigning the pigeon i in the process of generating ρ_{t+1} from ρ_t does not satisfy any literal in D is

$$\begin{aligned}\Pr[B^D_{\rho_t}] &= \prod_{r' \leq r} \left(1 - \frac{|H^{r'}_{\text{free}} \cap H^{r'}_{\text{sat}}|}{|H^{r'}_{\text{free}}|}\right) \leq \prod_{r' \leq r} \left(1 - \frac{|H^{r'}_{\text{free}}| + |H^{r'}_{\text{sat}}| - n}{n}\right) \\ &\leq \left(1 - \frac{\delta - 3wr - 1}{n}\right)^r \leq \left(1 - \frac{\delta}{4n}\right)^r.\end{aligned}$$

The estimation of $\Pr[B^D_{\rho_t}]$ is independent of ρ_t and the assignment of other pigeons than i in the process of constructing ρ_{t+1} from ρ_t. Hence, for each $\ell \leq L_\pi$,

$$\Pr\left[\exists t < \ell, E_t^D \cap B_t^D\right] \leq \sum_{t \leq \ell} \left(1 - \frac{\delta}{4n}\right)^r \leq L_\pi \left(1 - \frac{\delta}{4n}\right)^r \leq L_\pi 2^{-r\delta/4n},$$

where B_t^D is defined like $B^D_{\rho_t}$ but in the probability space of all ρ_ℓ's, and

$$\Pr[C \in \mathcal{A} \wedge C \text{ is not satisfied by } \rho_C] \leq \sum_{D \in \mathcal{A}} \Pr[\exists t < \ell, E_t^D \cap B_t^D] \leq |\mathcal{A}| L_\pi / 2^{r\delta/4n} < 1. \blacktriangleleft$$

Lemmas 18-19, with $m = \lfloor 2^{n^{1/5}} \rfloor$, $r = \lfloor n^{1/2} \rfloor$ and $\Phi = \neg\mathsf{PHP}^m_n$, imply that there is no Res refutation of PHP^m_n of size $2^{\epsilon n^{1/4}}$, for a sufficiently small ϵ.

Notably, our pseudowidth lower bound from Lemma 19 crucially relies on the assumption that the refutation π is short, unlike the lower bounds from [61, 62, 63, 20]. An advantage of our Res lower bound in comparison to [54, 61, 62, 63, 20] is that it constructs efficient Res-adversaries for $\neg\mathsf{PHP}^m_n$.

[15] We have $w \geq 1$ because either a cluase of $\neg\mathsf{PHP}^m_n$ is in π or every axiom in π is in $\mathcal{A}$ and there is a clause in π derived from clauses in $\mathcal{A}$ by weakening or by cut. In all cases it follows that $w \geq 1$.

We remark that $\mathsf{PHP}_s^{|W|}$ is closely related to the statement that a Boolean function f_n with n inputs cannot be computed by a perfect DNF (PDNF) of size s on inputs from $W \subseteq \{0,1\}^n$. Here, a PDNF of size s is a disjunction $\bigvee_{j \le s} K_j$, where each K_j is a conjunction of n literals such that each variable from $\{x_1, \ldots, x_n\}$ appears in K_j in the positive or negative form.

To see the connection, interpret the variable $x_{a,j}$ of $\mathsf{PHP}_s^{|W|}$ as the statement $K_j(a) = 1$ for $a \in W := f^{-1}(1)$. If we wanted to construct efficient adversaries for the encoding of PDNF lower bounds from [63], we would need to make our lower bound work for the "onto" version of the "functional" PHP_n^m.

▶ **Theorem 20** (Efficient adversary for the pseudowidth lower bound)**.** *Let $\gamma < 1$ be an arbitrary constant. There is $\epsilon > 0$ and $n_0 \in \mathbb{N}$ such that for all $n \ge n_0$ there is a poly(m, m')-size randomized* Res*-adversary C for $\neg\mathsf{PHP}_n^m$, $m = \lfloor 2^{n^{1/5}} \rfloor$, such that for each $2^{\epsilon n^{1/4}}$-size* Res *refutation π of any set of clauses, with probability $\ge \gamma$ over the randomness of C, π does not catch C. Here, m' is the input length of circuits in C.*

The proof of Theorem 20 follows by combining Lemma 18 with the proof of Lemma 19.

A complication is that when the adversary possesses ρ_C and receives a predecessor C_0 of C, it might not know what C_1 is. In such case the adversary pretends that C_1 is $C \vee \neg(C_0 \setminus C)$. More precisely, the adversary will automatically replace each clause $C_i \in \pi$, which was used in π to derive C from C_0 and C_1, by $C \vee \neg(C_{1-i} \setminus C)$. The transformation maps π into a sequence of clauses which can be completed into a valid Res refutation $\rho^1(\pi)$ of axioms of π: For each clause $C \vee \neg(C_{1-i} \setminus C)$ replacing a clause C_i, $\rho^1(\pi)$ contains a weakening step deriving $C \vee \neg(C_{1-i} \setminus C)$ from C_i. Since $\rho^1(\pi)$ introduces at most two new clauses for each derivation step in π, we have $L_{\rho^1(\pi)} \le 3L_\pi$.

Given a clause $C \in \rho^1(\pi)$, the adversary sets all variables that do not appear in $\neg\mathsf{PHP}_n^m$ to 0. This substitution, denoted ρ^2, transforms $\rho^1(\pi)$ into $\rho^2(\rho^1(\pi))$. Next, the adversary chooses a random vector d according to the distribution $\boldsymbol{d}$ from Lemma 18. Since the circuits defining the adversary share a random string, the adversary can reproduce the same d in each round. By Lemma 18, with probability $\ge 1 - 1/L_{\rho^1(\pi)}$, $\rho^2(\rho^1(\pi))$ contains a Res-refutation π' of $\Phi \cup \mathcal{A}$ such that $w_{d,\delta}(\pi' \setminus \mathcal{A}) \le O(n^{1/4})$, where $\mathcal{A}$ is a set of $(1,d)$-subaxioms and Φ is a subset of axioms of $\rho^2(\rho^1(\pi))$. W.l.o.g. $1 - 1/L_{\rho^1(\pi)} \ge \gamma + (1-\gamma)/2$ as otherwise we can extend π using the weakening rule on the last clause sufficiently many times.

The adversary then simulates the construction of partial multi-matchings ρ_C from the proof of Lemma 19 on π', with $r = \lfloor n^{1/2} \rfloor$, and responds to $C \in \pi$ by 0 if and only if it ends up in C when traversing π'. Since there might be axioms in π' which are not in $\neg\mathsf{PHP}_n^m \cup \mathcal{A}$, we might not have $w \ge 1$, but the size of R is still bounded by $3r \cdot max\{1, w\} \le \delta/2$, for our setting of parameters. The transformation ρ^1 and the shared randomness of the circuits defining the adversary ensure that independently of whether the adversary received a premise C_0 or C_1 of a cut in π, the adversary constructs the same ρ'_ℓ when tracing the corresponding cut in π' and responds consistently on the cut in π. The property that the adversary responds 1 to the axioms of $\neg\mathsf{PHP}_n^m$ is preserved and the probability of responding 0 to the clauses in $\mathcal{A}$ is $\le (1-\gamma)/2$, if n is big enough.

This completes the proof of Theorem 20.

3.3 k-evaluations

One of the most beautiful lower bound methods in proof complexity is the method of k-evaluations which was used to prove a lower bound on the length of proofs of the pigeonhole principle PHP_n (mapping $n+1$ pigeons to n holes) in constant-depth Frege systems [1, 38, 51].

In the rest of the subsection we prove the following theorem constructing efficient AC^0-Frege-adversaries for $\neg\mathsf{PHP}_n$.

▶ **Theorem 21** (Efficient adversary for AC^0-Frege lower bound)**.**
Let P be an AC^0_d-Frege system in the language consisting of $\vee, \neg, 0, 1$. Further let $\delta < 1/9^d$ and $\gamma < 1$. There is $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ there is a poly(n^{n^δ}, m')-size randomized P-adversary C for $\neg\mathsf{PHP}_n$ such that for each 2^{n^δ}-size P-refutation π of any set of formulas, with probability $\geq \gamma$ over the randomness of C, π does not catch C. Here m' is the input length of circuits in C.

We will refer to the presentation of the AC^0_d-Frege lower bound for PHP_n in [34, Chapter 15].

For a sequence of formulas $\{\theta_i\}$, denote by $\Gamma_d\{\theta_i\}$, the set of all depth d formulas in $\{\theta_i\}$ and all their subformulas.

In each round $j \geq 1$, our adversary will produce a k-evaluation (S^j, H^j) of $\Gamma_d\{\phi_i|_{\rho'\circ\rho_d}\}_{i\in[j]}$, where $\phi_i \in \pi$ is the formula received by the adversary in round i and $k = n^\delta$. The partial restriction ρ' sets all variables that do not appear in PHP_n to 0. In order to simplify the notation, we will ignore ρ' in the rest of our proof. The restriction ρ_d will be the restriction ρ_d from the sequence of restrictions $\rho_0 \subseteq \cdots \subseteq \rho_d$ in the proof of [34, Lemma 15.2.3] but with $\epsilon_0^d = \epsilon < 1/9^d$ instead of $\epsilon < 1/5^d$. We assume $\delta < \epsilon$. Since the circuits defining the adversary share a random string, the adversary can re-produce the same ρ_d in each round. The response of the adversary to the j-th formula ϕ_j from a 2^{n^δ}-size P-refutation π is 1 if $S^j_{\phi_j|_{\rho_d}}(H^j_{\phi_j|_{\rho_d}}) = H^j_{\phi_j|_{\rho_d}}$ and 0 otherwise.

The construction will ensure that there exists a k-evaluation (S, H) of $\Gamma_d\pi|_{\rho_d}$ such that for each j, (S^j, H^j) is (S, H) restricted to $\Gamma_d\{\phi_i|_{\rho_d}\}_{i\in[j]}$. In fact, (S, H) will not depend on $\{\phi_i\}_{i\in[j]}$, i.e., independently of whether the adversary traces π along the path $\phi_1, \ldots, \phi_j$ or along a different path $\phi'_1, \ldots, \phi'_j$, it constructs a restriction of a single "global" k-evaluation (S, H). Therefore, by [34, Lemma 15.1.7], the adversary will be consistent w.r.t. all derivation steps in π. Moreover, since ρ_d is a partial bijection, $\neg\mathsf{PHP}_n|_{\rho_d}$ is an instance of $\neg\mathsf{PHP}_{n^\epsilon}$ (possibly with some tautological clauses), so by [34, Lemma 15.1.6] the adversary will assign 1 to each clause of $\neg\mathsf{PHP}_n$. This means that π will not catch the adversary.

It remains to show how the adversary constructs (S^j, H^j).

The adversary will construct gradually $(S^{j,t}, H^{j,t})$, for each $t \in [d]$, where $(S^{j,t}, H^{j,t})$ is a k-evaluation of $\Gamma_t\{\phi_i|_{\rho_t}\}_{i\in[j]}$, and set $(S^j, H^j) := (S^{j,d}, H^{j,d})$. Recall that ρ_t comes from [34, Lemma 15.2.3] with the adjusted parameter.

The definition of k-evaluation gives an explicit algorithm constructing $(S^{j,t}, H^{j,t})$, if t is the depth of constants $1, 0$ or atoms $p_{i,j}$.

Suppose our adversary already constructed $(S^{j,t-1}, H^{j,t-1})$. $(S^{j,t}, H^{j,t})$ is obtained from $(S^{j,t-1}, H^{j,t-1})$ as follows. First, ρ_t is applied on $(S^{j,t-1}, H^{j,t-1})$, which by [34, Lemma 15.2.1] results in a k-evaluation (S', H') of $\Gamma_{t-1}\{\phi_i|_{\rho_t}\}_{i\in[j]}$. Then all formulas in the set $\Gamma_t\{\phi_i|_{\rho_t}\}_{i\in[j]} \setminus \Gamma_{t-1}\{\phi_i|_{\rho_t}\}_{i\in[j]}$ are gradually "k-evaluated" (i.e. added to the domain of (S', H')) in an arbitrary order. The k-evaluation obtained at the end of this process is $(S^{j,t}, H^{j,t})$.

If the adversary wants to k-evaluate $\neg\psi$ for some ψ that has been k-evaluated already, the construction of the corresponding k-tree and its subset follows directly from the definition of k-evaluation. The only case which requires a more sophisticated treatment is when the adversary needs to k-evaluate $\phi := \bigvee_i \psi_i$, for ψ_i's that have been already k-evaluated.

In the remaining case, the adversary already possesses $(S'_{\psi_i}, H'_{\psi_i})$ given by the k-evaluation of ψ_i's. The adversary constructs an $n^{O(k)}$-size k-tree S_ϕ such that $\bigcup_i H'_{\psi_i} \triangleleft S_\phi$ by fixing a strategy I_{fix} on $\bigcup_i H'_{\psi_i}$ and simulating all possible moves of player II in the proof of [34, Lemma 15.2.2]. The set H_ϕ is defined as the projection of $\bigcup_i H'_{\psi_i}$ on S_ϕ.

[34, Lemma 15.2.2], for $\epsilon < 1/9$, does not only show the existence of some ρ such that the desired trees S_i exist, but in fact it shows that these trees exist for a random ρ (chosen from a set of partial matchings) with probability at least $1 - (1-\gamma)/d$. In more detail, in the proof of the claim that for some ρ, the number of critical pairs is $\leq k/2$, we could assume that for $(1-\gamma)/d$ of all ρ's there is a play with at least $k/2$ critical pairs. This would lead to the inequality $(1-\gamma)a \leq dsb(2k^{1/2}n^\epsilon)^k$ which contradicts the setting of the parameters. As a consequence, the probability that the restrictions $\rho_1, \ldots, \rho_d$ guarantee that for each relevant ϕ the desired k-tree S_ϕ exists (and the probability that the adversary will not be caught) is $\geq \gamma$.

W.l.o.g. the construction of the k-evaluation of a formula depends only on the formula itself and on $\rho_0 \subseteq \cdots \subseteq \rho_d$. Since the circuits defining the adversary share randomness, it can be guaranteed that in each round the adversary uses the same $\rho_0, \ldots, \rho_d$, independently of how it traces the proof π, i.e., independently of $\phi_1, \ldots, \phi_j$. Therefore, there exists a k-evaluation (S, H) such that for every path $\phi_1, \ldots, \phi_j$ in π, the adversary constructs (S, H) restricted to $\Gamma_d\{\phi_i|_{\rho_d}\}_{i\in[j]}$.

Finally, generating $(S^{j,t}, H^{j,t})$ requires to k-evaluate $\leq |\pi|$ formulas and for each relevant subformula ϕ, the construction of $(S^{j,t}_\phi, H^{j,t}_\phi)$ can be done by a circuit of size $n^{O(k)}$. The complexity of the adversary is thus $poly(n^{n^\delta}, m')$. This completes the proof of Theorem 21.

References

1 Miklós Ajtai. The Complexity of the Pigeonhole Principle. *Comb.*, 14(4):417–433, 1994. doi:10.1007/BF01302964.

2 Noel Arteche, Albert Atserias, Susanna F. de Rezende, and Erfan Khaniki. The Proof Analysis Problem. In *66th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2025, Sydney, Australia, December 14-17, 2025*, pages 2555–2576, 2025. doi:10.1109/FOCS63196.2025.00133.

3 Noel Arteche, Erfan Khaniki, Ján Pich, and Rahul Santhanam. From Proof Complexity to Circuit Complexity via Interactive Protocols. In *51st International Colloquium on Automata, Languages, and Programming, ICALP 2024, Tallinn, Estonia, July 8-12, 2024*, pages 12:1–12:20, 2024. doi:10.4230/LIPIcs.ICALP.2024.12.

4 Albert Atserias and Moritz Müller. Automating Resolution is NP-Hard. *J. ACM*, 67(5):31:1–31:17, 2020. doi:10.1145/3409472.

5 Per Austrin and Kilian Risse. Sum-Of-Squares Lower Bounds for the Minimum Circuit Size Problem. In *38th Computational Complexity Conference, CCC 2023, Warwick, UK, July 17-20, 2023*, pages 31:1–31:21, 2023. doi:10.4230/LIPIcs.CCC.2023.31.

6 Paul Beame and Toniann Pitassi. Propositional proof complexity: Past, present, and future. In *Current Trends in Theoretical Computer Science, Entering the 21th Century*, pages 42–70. World Scientific, 2001.

7 Eli Ben-Sasson and Avi Wigderson. Short proofs are narrow - resolution made simple. *J. ACM*, 48(2):149–169, 2001. doi:10.1145/375827.375835.

8 Maria Luisa Bonet, Carlos Domingo, Ricard Gavaldà, Alexis Maciel, and Toniann Pitassi. Non-Automatizability of Bounded-Depth Frege Proofs. *Comput. Complex.*, 13(1-2):47–68, 2004. doi:10.1007/S00037-004-0183-5.

9 Maria Luisa Bonet, Toniann Pitassi, and Ran Raz. On Interpolation and Automatization for Frege Systems. *SIAM J. Comput.*, 29(6):1939–1967, 2000. doi:10.1137/S0097539798353230.

10 Samuel R. Buss. *Bounded arithmetic*. Princeton University, 1985.

11 Samuel R. Buss. On model theory for intuitionistic bounded arithmetic with applications to independence results. In *Feasible Mathematics: A Mathematical Sciences Institute Workshop, Ithaca, New York, June 1989*, pages 27–47, 1990.

12 Marco L. Carmosino, Russell Impagliazzo, Valentine Kabanets, and Antonina Kolokolova. Learning Algorithms from Natural Proofs. In *31st Conference on Computational Complexity, CCC 2016, Tokyo, Japan, May 29 - June 1, 2016*, pages 10:1–10:24, 2016. doi:10.4230/LIPIcs.CCC.2016.10.

13 Lijie Chen, Shuichi Hirahara, Igor Carboni Oliveira, Ján Pich, Ninad Rajgopal, and Rahul Santhanam. Beyond Natural Proofs: Hardness Magnification and Locality. *J. ACM*, 69(4):25:1–25:49, 2022. doi:10.1145/3538391.

14 Lijie Chen, Jiatu Li, and Igor C. Oliveira. Reverse Mathematics of Complexity Lower Bounds. In *65th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2024, Chicago, IL, USA, October 27-30, 2024*, pages 505–527, 2024. doi:10.1109/FOCS61266.2024.00040.

15 Lijie Chen, Ron D. Rothblum, and Roei Tell. Fiat-Shamir in the Plain Model from Derandomization (Or: Do Efficient Algorithms Believe that NP = PSPACE?). In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing, STOC 2025, Prague, Czechia, June 23-27, 2025*, pages 977–985, 2025. doi:10.1145/3717823.3718177.

16 Stephen A. Cook. Feasibly Constructive Proofs and the Propositional Calculus. In *Proceedings of the 7th Annual ACM Symposium on Theory of Computing, May 5-7, 1975, Albuquerque, New Mexico, USA*, pages 83–97, 1975.

17 Stephen A. Cook and Jan Krajícek. Consequences of the provability of $NP \subseteq P/poly$. *J. Symb. Log.*, 72(4):1353–1371, 2007. doi:10.2178/JSL/1203350791.

18 Stephen A. Cook and Alasdair Urquhart. Functional Interpretations of Feasibly Constructive Arithmetic. *Ann. Pure Appl. Log.*, 63(2):103–200, 1993. doi:10.1016/0168-0072(93)90044-E.

19 Ben Davis and Robert Robere. Res(log) Proves Bounded-Depth Frege Lower Bounds. *ECCC*, TR26-055, 2026.

20 Susanna F. de Rezende, Jakob Nordström, Kilian Risse, and Dmitry Sokolov. Exponential Resolution Lower Bounds for Weak Pigeonhole Principle and Perfect Matching Formulas over Sparse Graphs. *TheoretiCS*, 4, 2025. doi:10.46298/THEORETICS.25.9.

21 Mika Göös. *Communication Lower Bounds via Query Complexity*. University of Toronto (Canada), 2016.

22 Joshua A. Grochow and Toniann Pitassi. Circuit Complexity, Proof Complexity, and Polynomial Identity Testing: The Ideal Proof System. *J. ACM*, 65(6):37:1–37:59, 2018. doi:10.1145/3230742.

23 Svyatoslav Gryaznov and Navid Talebanfard. Bounded Depth Frege Lower Bounds for Random 3-CNFs via Deterministic Restrictions. *arXiv*, 2024. doi:10.48550/arXiv.2403.02275.

24 Shuichi Hirahara, Rahul Ilango, and Ryan Williams. Beating Brute Force for Compression Problems. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing, STOC 2024, Vancouver, BC, Canada, June 24-28, 2024*, pages 659–670, 2024. doi:10.1145/3618260.3649778.

25 Emil Jerábek. Dual weak pigeonhole principle, Boolean complexity, and derandomization. *Ann. Pure Appl. Log.*, 129(1-3):1–37, 2004. doi:10.1016/J.APAL.2003.12.003.

26 Ondřej Ježil and Dimitrios Tsintsilidas. Parallelism and Adaptivity in Student-Teacher Witnessing. *arXiv*, 2026. arXiv:2602.19934.

27 Erfan Khaniki. Nisan-Wigderson Generators in Proof Complexity: New Lower Bounds. In *37th Computational Complexity Conference, CCC 2022, Philadelphia, PA, USA, July 20-23, 2022*, pages 17:1–17:15, 2022. doi:10.4230/LIPIcs.CCC.2022.17.

28 Erfan Khaniki. Jump Operators, Interactive Proofs and Proof Complexity Generators. In *65th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2024, Chicago, IL, USA, October 27-30, 2024*, pages 573–593, 2024. doi:10.1109/FOCS61266.2024.00044.

29 Jan Krajicek. *Bounded arithmetic, propositional logic and complexity theory*, volume 60. Cambridge University Press, 1995.

30 Jan Krajíček. On the weak pigeonhole principle. *Fundamenta Mathematicae*, 170:123–140, 2001.

31 Jan Krajícek. Implicit proofs. *J. Symb. Log.*, 69(2):387–397, 2004. doi:10.2178/JSL/1082418532.

32 Jan Krajícek. On the Proof Complexity of the Nisan-Wigderson Generator based on a Hard NP ∩ coNP function. *J. Math. Log.*, 11(1), 2011. doi:10.1142/S0219061311000979.

33 Jan Krajíček. On the computational complexity of finding hard tautologies. *Bulletin of the London Mathematical Society*, 46(1):111–125, 2014.

34 Jan Krajíček. *Proof complexity*, volume 170. Cambridge University Press, 2019.

35 Jan Krajíček. *Proof Complexity Generators*. Cambridge University Press, 2025.

36 Jan Krajíček and Pavel Pudlák. Propositional provability and models of weak arithmetic. In *International Workshop on Computer Science Logic*, pages 193–210, 1989.

37 Jan Krajícek and Pavel Pudlák. Some Consequences of Cryptographical Conjectures for S_2^1 and EF. *Inf. Comput.*, 140(1):82–94, 1998.

38 Jan Krajíček, Pavel Pudlák, and Alan Woods. An exponential lower bound to the size of bounded depth Frege proofs of the pigeonhole principle. *Random structures & algorithms*, 7(1):15–39, 1995. doi:10.1002/RSA.3240070103.

39 Jiatu Li and Igor C. Oliveira. Unprovability of Strong Complexity Lower Bounds in Bounded Arithmetic. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing, STOC 2023, Orlando, FL, USA, June 20-23, 2023*, pages 1051–1057, 2023. doi:10.1145/3564246.3585144.

40 Jiawei Li, Yuhao Li, and Hanlin Ren. Finding Bugs in Short Proofs: The Metamathematics of Resolution Lower Bounds . *arXiv*, 2026. arXiv:2411.15515.

41 Richard J. Lipton and Neal E. Young. Simple strategies for large zero-sum games with applications to complexity theory. In *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing, 23-25 May 1994, Montréal, Québec, Canada*, pages 734–740, 1994. doi:10.1145/195058.195447.

42 Jiaqi Lu, Rahul Santhanam, and Iddo Tzameret. $AC^0[p]$-Frege Cannot Efficiently Prove That Constant-Depth Algebraic Circuit Lower Bounds Are Hard. In *17th Innovations in Theoretical Computer Science Conference, ITCS 2026, Bocconi University, Milan, Italy, January 27-30, 2026*, pages 99:1–99:25, 2026.

43 Noam Mazor and Rafael Pass. The Non-Uniform Perebor Conjecture for Time-Bounded Kolmogorov Complexity Is False. In *15th Innovations in Theoretical Computer Science Conference, ITCS 2024, Berkeley, CA, USA, January 30 - February 2, 2024*, pages 80:1–80:20, 2024. doi:10.4230/LIPIcs.ITCS.2024.80.

44 Moritz Müller and Ján Pich. Feasibly constructive proofs of succinct weak circuit lower bounds. *Ann. Pure Appl. Log.*, 171(2), 2020. doi:10.1016/J.APAL.2019.102735.

45 Igor C. Oliveira. SIGACT News Complexity Theory Column 124 Meta-Mathematics of Computational Complexity Theory. *SIGACT News*, 56(1):41–68, 2025. doi:10.1145/3726856.3726862.

46 Ján Pich. Circuit lower bounds in bounded arithmetics. *Ann. Pure Appl. Log.*, 166(1):29–45, 2015. doi:10.1016/J.APAL.2014.08.004.

47 Ján Pich. Logical strength of complexity theory and a formalization of the PCP theorem in bounded arithmetic. *Log. Methods Comput. Sci.*, 11(2), 2015. doi:10.2168/LMCS-11(2:8)2015.

48 Ján Pich and Rahul Santhanam. Why are Proof Complexity Lower Bounds Hard? In *60th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2019, Baltimore, Maryland, USA, November 9-12, 2019*, pages 1305–1324, 2019. doi:10.1109/FOCS.2019.00080.

49 Ján Pich and Rahul Santhanam. Strong co-nondeterministic lower bounds for NP cannot be proved feasibly. In *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*, pages 223–233, 2021. doi:10.1145/3406325.3451117.

50 Jan Pich and Rahul Santhanam. Towards $P \neq NP$ from Extended Frege lower bounds. *J. ACM*, 73(2), 2026.

51 Toniann Pitassi, Paul Beame, and Russell Impagliazzo. Exponential Lower Bounds for the Pigeonhole Principle. *Comput. Complex.*, 3:97–140, 1993. doi:10.1007/BF01200117.

52 Pavel Pudlák. Lower Bounds for Resolution and Cutting Plane Proofs and Monotone Computations. *J. Symb. Log.*, 62(3):981–998, 1997. doi:10.2307/2275583.

53 Pavel Pudlák and Samuel R. Buss. How to Lie Without Being (Easily) Convicted and the Length of Proofs in Propositional Calculus. In *Computer Science Logic, 8th International Workshop, CSL '94, Kazimierz, Poland, September 25-30, 1994, Selected Papers*, pages 151–162, 1994. doi:10.1007/BFB0022253.

54 Ran Raz. Resolution lower bounds for the weak pigeonhole principle. *J. ACM*, 51(2):115–138, 2004. doi:10.1145/972639.972640.

55 Alexander Razborov. Lower bounds on the monotone complexity of some Boolean function. In *Soviet Math. Dokl.*, volume 31, pages 354–357, 1985.

56 Alexander Razborov. Lower bounds on the size of bounded depth circuits over a complete basis with logical addition. *Mathematical Notes of the Academy of Sciences of the USSR*, 41(4):333–338, 1987.

57 Alexander Razborov. On the Method of Approximations. In *Proceedings of the 21st Annual ACM Symposium on Theory of Computing, May 14-17, 1989, Seattle, Washington, USA*, pages 167–176, 1989. doi:10.1145/73007.73023.

58 Alexander Razborov. *On provably disjoint NP-pairs.* Aarhus Universitet. Basic Research in Computer Science [BRICS], 1994.

59 Alexander Razborov. Unprovability of lower bounds on circuit size in certain fragments of bounded arithmetic. *Izvestiya: mathematics*, 59(1):205–227, 1995.

60 Alexander Razborov. Lower Bounds for the Polynomial Calculus. *Comput. Complex.*, 7(4):291–324, 1998. doi:10.1007/S000370050013.

61 Alexander Razborov. Improved Resolution Lower Bounds for the Weak Pigeonhole Principle. *ECCC*, TR01-055, 2001. URL: https://eccc.weizmann.ac.il/eccc-reports/2001/TR01-055/index.html.

62 Alexander Razborov. Resolution lower bounds for the weak functional pigeonhole principle. *Theor. Comput. Sci.*, 303(1):233–243, 2003. doi:10.1016/S0304-3975(02)00453-X.

63 Alexander Razborov. Resolution lower bounds for perfect matching principles. *J. Comput. Syst. Sci.*, 69(1):3–27, 2004. doi:10.1016/J.JCSS.2004.01.004.

64 Alexander Razborov. Pseudorandom generators hard for k-DNF resolution and polynomial calculus resolution. *Annals of Mathematics*, pages 415–472, 2015.

65 Alexander Razborov and Steven Rudich. Natural Proofs. *J. Comput. Syst. Sci.*, 55(1):24–35, 1997. doi:10.1006/JCSS.1997.1494.

66 Robert Robere. *Unified lower bounds for monotone computation.* University of Toronto (Canada), 2018.

67 Rahul Santhanam and Iddo Tzameret. Iterated Lower Bound Formulas: A Diagonalization-Based Approach to Proof Complexity. *SIAM Journal on Computing*, pages STOC21–313, 2025.

68 Nathan Segerlind, Samuel R. Buss, and Russell Impagliazzo. A Switching Lemma for Small Restrictions and Lower Bounds for k-DNF Resolution. *SIAM J. Comput.*, 33(5):1171–1200, 2004. doi:10.1137/S0097539703428555.

69 Roman Smolensky. Algebraic Methods in the Theory of Lower Bounds for Boolean Circuit Complexity. In *Proceedings of the 19th Annual ACM Symposium on Theory of Computing, 1987, New York, New York, USA*, pages 77–82, 1987. doi:10.1145/28395.28404.

70 Ryan Williams. Strong ETH Breaks With Merlin and Arthur: Short Non-Interactive Proofs of Batch Evaluation. In *31st Conference on Computational Complexity, CCC 2016, Tokyo, Japan, May 29 - June 1, 2016*, pages 2:1–2:17, 2016. doi:10.4230/LIPIcs.CCC.2016.2.