THEORY OF COMPUTING, Volume 14 (9), 2018, pp. 1–55
www.theoryofcomputing.org

SPECIAL ISSUE: CCC 2016

Average-Case Lower Bounds and Satisfiability Algorithms for Small Threshold Circuits

Ruiwen Chen   Rahul Santhanam   Srikanth Srinivasan

Received June 16, 2016; Revised May 10, 2018; Published June 5, 2018

Abstract: We show average-case lower bounds for explicit Boolean functions against bounded-depth threshold circuits with a superlinear number of wires. We show that for each integer $d > 1$, there is a constant $\varepsilon_d > 0$ such that the Parity function on n bits has correlation at most $n^{-\varepsilon_d}$ with depth-d threshold circuits which have at most $n^{1+\varepsilon_d}$ wires, and the Generalized Andreev function on n bits has correlation at most $\exp(-n^{\varepsilon_d})$ with depth-d threshold circuits which have at most $n^{1+\varepsilon_d}$ wires. Previously, only worst-case lower bounds in this setting were known (Impagliazzo, Paturi, and Saks (SICOMP 1997)).

We use our ideas to make progress on several related questions. We give satisfiability algorithms beating brute force search for depth-d threshold circuits with a superlinear number of wires. These are the first such algorithms for depth greater than 2. We also show that Parity on n bits cannot be computed by polynomial-size AC^0 circuits with $n^{o(1)}$ *general* threshold gates. Previously no lower bound for Parity in this setting could handle more than

A conference version of this paper appeared in the Proceedings of the 31st Conference on Computational Complexity, 2016 [8].

ACM Classification: F.1.2, F.2.3

AMS Classification: 68Q17, 68Q32

Key words and phrases: complexity theory, circuit complexity, correlation bounds, threshold functions, random restrictions, learning, SAT

DOI: 10.4086/toc.2018.v014a009

$\log(n)$ gates. This result also implies subexponential-time learning algorithms for AC^0 with $n^{o(1)}$ threshold gates under the uniform distribution. In addition, we give almost optimal bounds for the number of gates in a depth-d threshold circuit computing Parity on average, and show average-case lower bounds for threshold formulas of *any* depth.

Our techniques include adaptive random restrictions, anti-concentration and the structural theory of linear threshold functions, and bounded-read Chernoff bounds.

1 Introduction

One of the main goals in complexity theory is to prove circuit lower bounds for explicit functions in P or NP. We seem quite far from being able to prove that there is a problem in NP that requires superlinear Boolean circuits. We have some understanding, via formulations such as the relativization barrier [5], the "natural proofs" barrier [46] and the algebrization barrier [1], of why current techniques are inadequate for this purpose.

However, the community has had more success proving explicit lower bounds against bounded-depth circuits of various kinds. Thanks to pioneering work of Ajtai [2], Furst-Saxe-Sipser [15], Yao [56] and Håstad [21], we know that the Parity and Majority functions require bounded-depth unbounded fan-in circuits of exponential size if only AND, OR, and NEGATION gates are allowed. Later, Razborov [45] and Smolensky [53] showed that Majority requires exponential size even when MOD_p gates are allowed in addition to AND and OR gates, for any prime p.[1] The case of bounded-depth circuits with AND, OR and MOD_m gates, where m is a composite, has been open for nearly thirty years now, even though Majority is conjectured to be hard for such circuits. Williams [55] recently made significant progress by showing that nondeterministic exponential time does not have super-polynomial size circuits with AND, OR and MOD_m gates, for any m.

For all the bounded-depth circuit classes above, Majority is either known or conjectured to be hard. How about circuit classes which incorporate majority gates, or more generally, gates that are arbitrary linear threshold functions? Note that such gates generalize AND and OR, though not MOD_p. In the 90s, there was some work on studying the power of bounded-depth threshold circuits. Paturi and Saks [41] showed that depth-2 circuits with majority gates computing Parity require $\widetilde{\Omega}(n^2)$ wires; there is also a nearly matching upper bound for Parity. Impagliazzo, Paturi and Saks [24] considered bounded-depth threshold circuits with arbitrary linear threshold gates, and showed that for each depth d, there is a constant $\varepsilon_d > 0$ such that Parity requires $n^{1+\varepsilon_d}$ wires to compute with depth-d threshold circuits.

These lower bounds are *worst-case* lower bounds—they show that for any sequence of small circuits, there exist inputs of every length on which the circuits fail to compute Parity. There are several reasons to be interested in *average-case* lower bounds under the uniform distribution, or equivalently, in *correlation* upper bounds. For one, average-case lower bounds show that a randomly chosen input is likely to be hard, and thus give a way to generate hard instances efficiently. Second, average-case lower bounds are closely tied to pseudorandom generators via the work of Nisan and Wigderson [33, 35], and are indeed a prerequisite for obtaining pseudorandom generators with non-trivial seed length for a circuit class. Third, recent work on satisfiability algorithms [49, 22, 7] indicates that the design and analysis of non-trivial

[1] Recall that for an integer $m \geq 2$, a MOD_m gate accepts exactly those inputs whose Hamming weight is not divisible by m.

satisfiability algorithms is closely tied to proving average-case lower bounds, though there is no formal connection. Fourth, the seminal paper by Linial, Mansour, and Nisan [30] shows that average-case lower bounds for Parity against a circuit class are tied to non-trivially learning the circuit class under the uniform distribution.

With these different motivations in mind, we systematically study average-case lower bounds for bounded-depth threshold circuits. Our first main result shows correlation upper bounds for Parity and another explicit function known as the Generalized Andreev function with respect to threshold circuits with few wires. No correlation upper bounds for explicit functions against bounded-depth threshold circuits with superlinear wires was known before our work.

Theorem 1.1. *For each depth $d \geq 1$, there is a constant $\varepsilon_d > 0$ such that for all large enough n, no threshold circuit of depth d with at most $n^{1+\varepsilon_d}$ wires agrees with Parity on more than $1/2 + 1/n^{\varepsilon_d}$ fraction of inputs of length n, and with the Generalized Andreev function on more than $1/2 + 1/2^{n^{\varepsilon_d}}$ fraction of inputs of length n.*

Theorem 1.1 captures the content of Theorem 4.2 and Theorem 4.13 in Section 4.

Quite often, ideas behind lower bound results for a circuit class $\mathcal{C}$ result in satisfiability algorithms (running in better than brute-force time) for $\mathcal{C}$, and vice-versa (see, e. g., [40]). Here, we leverage the ideas behind our correlation bounds against Threshold circuits to improved satisfiability algorithms for these circuits. More precisely, we constructivize the ideas of the proof of the strong correlation upper bounds for the Generalized Andreev function to get non-trivial satisfiability algorithms for bounded-depth threshold circuits with few wires. Previously, such algorithms were only known for depth-2 circuits, due to Impagliazzo–Paturi–Schneider [23] and Tamaki (unpublished).

Theorem 1.2. *For each depth $d \geq 1$, there is a constant $\varepsilon_d > 0$ such that the satisfiability of depth-d threshold circuits with at most $n^{1+\varepsilon_d}$ wires can be solved in randomized time $2^{n-n^{\varepsilon_d}}\mathrm{poly}(n)$.*

Theorem 1.2 is re-stated and proved as Theorem 6.4 in Section 6.

Using our ideas, we also show correlation bounds against AC^0 circuits with a few threshold gates, as well as learning algorithms under the uniform distribution for such circuits.

Theorem 1.3. *For each constant d, there is a constant $\gamma > 0$ such that Parity has correlation at most $1/n^{\Omega(1)}$ with AC^0 circuits of depth d and size at most $n^{\log(n)^{0.4}}$ augmented with at most n^{γ} threshold gates. Moreover, the class of AC^0 circuits of size at most $n^{\log(n)^{0.4}}$ augmented with at most n^{γ} threshold gates can be learned to constant error under the uniform distribution in time $2^{n^{1/4+o(1)}}$.*

Theorem 1.3 captures the content of Corollary 8.4 and Theorem 8.2 in Section 7.

Having summarized our main results, we now describe related work and our proof techniques in more detail.

1.1 Related work

There has been a large body of work proving upper and lower bounds for constant-depth threshold circuits. Much of this work has focused on the setting of small gate complexity, which seems to be the somewhat easier case to handle. A distinction must also be drawn between work that has focused on the setting where the threshold gates are assumed to be *majority gates* (i. e., the linear function sign representing the gate has integer coefficients that are bounded by a polynomial in the number of variables) and work that focuses on general threshold gates, since analytic tools such as *rational approximation* that are available for majority gates do not work in the setting of general threshold gates.

We discuss the work on wire complexity first, followed by the results on gate complexity.

Wire complexity. Paturi and Saks [41] considered depth-2 *majority* circuits and showed an $\widetilde{\Omega}(n^2)$ lower bound on the wire complexity required to compute Parity; this nearly matches the upper bound of $O(n^2)$. They also showed that there exist majority circuits of size $n^{1+\Theta(\varepsilon_1^d)}$ and depth d computing Parity; here $\varepsilon_1 = 2/(1+\sqrt{5})$. Impagliazzo, Paturi, and Saks [24] showed a depth-d lower bound for *general* threshold circuits computing Parity: namely, that any such circuit must have wire complexity at least $n^{1+\varepsilon_2^d}$ where $\varepsilon_2 < \varepsilon_1$.

The proof of [24] proceeds by induction on the depth d. The main technical lemma shows that a circuit of depth d can be converted to a depth-$(d-1)$ circuit of the same size by setting some of the input variables. The fixed variables are set in a random fashion, but *not* according to the uniform distribution. In fact, this distribution has statistical distance close to 1 from the uniform distribution and furthermore, depends on the circuit whose depth is being reduced. Therefore, it is unclear how to use this technique to prove a correlation bound with respect to the uniform distribution. In contrast, we are able to reduce the depth of the circuit by setting variables uniformly at random (though the variables that we restrict are sometimes chosen in a way that depends on the circuit), which yields the correlation bounds we want.

Gate complexity. The aforementioned paper by Paturi and Saks [41] also proved a near optimal $\widetilde{\Omega}(n)$ lower bound on the number of gates in any depth-2 majority circuits computing Parity.

Siu, Roychowdhury, and Kailath [52] considered majority circuits of bounded depth and small gate complexity. They showed that Parity can be computed by depth-d majority circuits with $O(dn^{1/(d-1)})$ gates. Building on the ideas of [41], they also proved a near matching lower bound of $\widetilde{\Omega}(dn^{1/(d-1)})$. Further, they also considered the problem of correlation bounds and showed that there exist depth-d majority circuits with $O(dn^{1/2(d-1)})$ gates that compute Parity almost everywhere and that majority circuits of significantly smaller size have $o(1)$ correlation with Parity (i. e., these circuits cannot compute Parity on more than a $1/2+o(1)$ fraction of inputs; recall that $1/2$ is trivial since a constant function computes Parity correctly on $1/2$ of its inputs). Impagliazzo, Paturi, and Saks [24] extended the worst-case lower bound to general threshold gates, where they proved a slightly weaker lower bound of $\Omega(n^{1/2(d-1)})$. As discussed above, though, it is unclear how to use their technique to prove a correlation bound.

Beigel [6] extended the result of Siu et al. to the setting of AC^0 augmented with a few majority gates. He showed that any subexponential-sized depth-d AC^0 circuit with significantly less than some $k = n^{\Theta(1/d)}$ majority gates has correlation $o(1)$ with Parity. The techniques of all the above results with the exception of [24] were based on the fact that majority gates can be well-approximated by low-degree *rational functions*. However, this is not true for general threshold functions [51] and hence, these techniques do not carry over to the case of general threshold gates.

A lower bound technique that does carry over to the setting of general threshold gates is that of showing that the circuit class has low-degree polynomial *sign-representations*. Aspnes, Beigel, Furst and Rudich [3] used this idea to prove that AC^0 circuits augmented with a single general threshold *output* gate—we refer to these circuits as TAC^0 circuits as in [17]—of subexponential-size and constant-depth have correlation $o(1)$ with Parity. More recently, Podolskii [43] used this technique along with a trick due to Beigel [6] to prove similar bounds for subexponential-sized AC^0 circuits augmented with general threshold gates. However, this trick incurs an exponential blow-up with the number of threshold gates and hence, in the setting of the Parity function, we cannot handle $k > \log n$ threshold gates.

Another technique that has proved useful in handling general threshold gates is *Communication Complexity*, where the basic idea is to show that the circuit—perhaps after restricting some variables—has low communication complexity in some suitably defined communication model. We can then use results from communication complexity to infer lower bounds or correlation bounds. Nisan [34] used this technique to prove exponential correlation bounds against general threshold circuits (not necessarily even constant-depth) with $n^{1-\Omega(1)}$ threshold gates. Using Beigel's trick and multiparty communication complexity bounds of Babai, Nisan and Szegedy [4], Lovett and Srinivasan [31] (see also [47, 19]) proved exponential correlation bounds against any polynomial-sized AC^0 circuits augmented with up to $n^{\frac{1}{2}-\Omega(1)}$ threshold gates.

We do not use this technique in our setting for many reasons. Firstly, it cannot be used to prove lower bounds or correlation bounds for functions such as Parity (which has small communication complexity in most models). In particular, these ideas do not yield the noise sensitivity bounds we get here. Even more importantly, it is unclear how to use these techniques to prove any sort of superlinear lower bound on wire complexity, since there are functions that have threshold circuits with linearly many wires, but large communication complexity even after applying restrictions (take a generic read-once depth-2 majority formula for example).

Perhaps most closely related to our work is that of Gopalan and Servedio [17] who use analytic techniques to prove correlation bounds against AC^0 circuits augmented with a few threshold gates. Their idea is to use noise sensitivity bounds (as we do as well) to obtain correlation bounds for Parity against TAC^0 circuits and then extend these results in the same way as in the paper by Podolskii [43] mentioned above. As a result, though, the result only yields non-trivial bounds when the number of threshold gates is bounded by $\log n$, whereas our result yields correlation bounds for up to $n^{1/2(d-1)}$ threshold gates.

1.2 Proof techniques

In recent years, there has been much work exploring the analytic properties (such as noise sensitivity) of linear threshold functions (LTFs) and their generalizations polynomial threshold functions (PTFs) (e. g., [50, 38, 11, 20, 12, 32, 26]). We show here that these techniques can be used in the context of constant-depth threshold circuits as well. In particular, using these techniques, we prove a qualitative refinement of Peres' theorem (see below) that may be of independent interest.

Our first result (Theorem 3.2 in Section 3) is a tight correlation bound for Parity with threshold circuits of depth d and gate complexity much smaller than $n^{1/2(d-1)}$. This generalizes both the results of Siu et al. [52], who proved such a result for *majority* circuits, and Impagliazzo, Paturi, and Saks [24], who proved a worst-case lower bound of the same order. The proof uses a fundamental theorem of Peres [42] on the noise sensitivity of LTFs; Peres' theorem has also been used by Klivans, O'Donnell, and Servedio [28] to obtain learning algorithms for functions of a few threshold gates. We use Peres' theorem to prove a noise sensitivity upper bound on small threshold circuits of constant depth.

The observation underlying the proof is that the noise sensitivity of a function is exactly the expected variance of the function after applying a suitable random restriction (see also [36]). Seen in this light, Peres' theorem says that, on application of a random restriction, any threshold function becomes quite biased in expectation and hence is well approximated by a constant function. Our analysis of the threshold circuit therefore proceeds by applying a random restriction to the circuit and replacing all the threshold gates at depth $d-1$ by the constants that best approximate them to obtain a circuit of depth $d-1$. A straightforward union bound tells us that the new circuit is a good approximation of the original circuit after the restriction. We repeat this procedure with the depth-$(d-1)$ circuit until the entire circuit becomes a constant, at which point we can say that after a suitable random restriction, the original circuit is well approximated by a constant, which means its variance is small. Hence, the noise sensitivity of the original circuit must be small as well and we are done.

This technique is expanded upon in Section 8, where we use a powerful noise sensitivity upper bound for low degree PTFs due to Kane [26] along with standard switching arguments [21] to prove similar results for AC^0 circuits augmented with almost $n^{1/2(d-1)}$ threshold gates. This yields Theorem 1.3 by some standard results.

In Section 4, we consider the problem of extending the above correlation bounds to threshold circuits with small (slightly superlinear) wire complexity. The above proof breaks down even for depth-2 threshold circuits with a superlinear number of wires, since such circuits could have a superlinear number of gates and hence the union bound referred to above is no longer feasible.

In the case of depth-2 threshold circuits, we are nevertheless able to use Peres' theorem, along with ideas of [3] to prove correlation bounds for Parity against circuits with nearly $n^{1.5}$ wires. (This result was independently obtained by Kane and Williams in a recent paper [27].) This result is tight, since by a result of Siu et al. [52], Parity can be well approximated by depth-2 circuits with $O(\sqrt{n})$ gates and hence $O(n^{1.5})$ wires. This argument is in Section 4.1.

Unfortunately, however, this technique requires us to set a large number of variables, which renders it unsuitable for larger depths. The reason for this is that, if we set a large number of variables to reduce the depth from some large constant d to $d-1$, then we may be in a setting where the number of wires is much larger than the number of surviving variables and hence correlation bounds for Parity may no longer be possible at all.

We therefore use a different strategy to prove correlation bounds against larger constant depths. The linchpin in the argument is a qualitative refinement of Peres' theorem (Lemma 4.4) that says that on application of a random restriction to an LTF, with good probability, the variance of the LTF becomes *negligible* (even exponentially small for suitable parameters). The proof of this argument is via anticoncentration results based on the Berry-Esseen theorem and the analysis of general threshold functions via a *critical index* argument as in many recent papers [50, 38, 11, 32].

The above refinement of Peres' theorem allows us to proceed with our argument as in the gates case. We apply a random restriction to the circuit and by the refinement, with good probability (say $1-n^{-\Omega(1)}$) most gates end up exponentially close to constants. We can then set these "imbalanced" gates to constants and still apply a union bound to ensure that the new circuit is a good approximation to the old one. For the small number of gates that do not become imbalanced in this way, we set *all* variables feeding into them. Since the number of such gates is small, we do not set too many variables. We now have a depth-$(d-1)$ circuit. Continuing in this way, we get a correlation bound of $n^{-\Omega(1)}$ with Parity. This gives part of

Theorem 1.1.

We then strengthen this correlation bound to $\exp(-n^{\Omega(1)})$ for the *Generalized Andreev function*, which, intuitively speaking, has the following property: even after applying any restriction that leaves a certain number of variables unfixed, the function has exponentially small correlation with any LTF on the surviving variables. To prove lower bounds for larger depth threshold circuits, we follow more or less the same strategy, except that in the above argument, we need most gates to become imbalanced with very high probability $(1-\exp(-n^{\Omega(1)}))$. To ensure this, we use a *bounded read Chernoff bound* due to Gavinsky, Lovett, Saks, and Srinivasan [16]. We can use this technique to reduce depth as above as long as the number of threshold gates at depth $d-1$ is "reasonably large." If the number of gates at depth $d-1$ is very small, then we simply guess the values of these few threshold gates and move them to the top of the circuit and proceed. This gives the other part of Theorem 1.1.

This latter depth-reduction result can be completely constructivized to design a satisfiability algorithm that runs in time $2^{n-n^{\Omega(1)}}$. The algorithm proceeds in the same way as the above argument, iteratively reducing the depth of the circuit. A subtlety arises when we replace imbalanced gates by constants, since we are changing the behaviour of the circuit on some (though very few) inputs. Thus, a circuit which was satisfiable only at one among these inputs might now end up unsatisfiable. However, we show that there is an efficient algorithm that enumerates these inputs and can hence check if there are satisfiable assignments to the circuits from among these inputs. This gives Theorem 1.2.

In Section 7, we prove correlation bounds for the Generalized Andreev function against threshold *formulas* of any arity and any depth. The proof is based on a retooling of the argument of Nečiporuk for formulas of constant arity over any basis and yields a correlation bound as long as the wire complexity is at most $n^{2-\Omega(1)}$.

Independent work of Kane and Williams [27] and connections to our results. Independent of our work, Kane and Williams [27] proved some very interesting results on threshold circuit lower bounds. Their focus is on threshold circuits of depth 2 and (a special case of) depth 3 only, but in this regime they are able to show stronger superlinear gate lower bounds and superquadratic wire lower bounds on the complexity of an explicit function (which is closely related to the Generalized Andreev function referred to above).

The techniques of [27] are closely related to ours, since they also analyze the effect of random restrictions on threshold gates. The statement of the random restriction lemma in our paper is different from that in [27]. While we are only able to prove that a threshold gate becomes highly biased with high probability under a random restriction, [27] prove that with high probability, the threshold gate actually becomes a *constant*.

However, to obtain this stronger conclusion, the random restrictions from [27] have to set *most* variables, which makes it unsuitable for proving lower bounds for larger depths as mentioned above. This technique can recover our tight average-case gate lower bounds (Section 3) and wire lower bounds (Section 4.1) for depth 2.

While our random restriction lemmas are weaker (in that they don't make the threshold gate a constant), it makes sense to ask if we can use them to recover the threshold circuit lower bounds of [27].

In Section 9, we show that this is true by using a natural strengthening of our main restriction lemma to prove a lower bound for threshold circuits that matches the lower bound of [27] upto $n^{o(1)}$ factors. The results of this section were obtained subsequent to the work of Kane and Williams.

This result should not be considered an "alternate proof," since we use many other ideas from [27] in the proof (we have also used one of these ideas, namely Theorem 2.9, to improve our lower bounds for Threshold formulas from [8] in Section 7). Our interest is in understanding the relative power of the random restriction lemmas in the two results.

2 Preliminaries

Throughout, all logarithms will be taken to the base 2.

2.1 Basic Boolean function definitions

A Boolean function on n variables will be a function $f : \{-1,1\}^n \to \{-1,1\}$. We use the standard inner product on functions $f,g : \{-1,1\}^n \to \mathbb{R}$ defined by

$$\langle f,g\rangle = \mathop{\mathbf{E}}_{x\sim\{-1,1\}^n}[f(x)g(x)].$$

(Unless specifically mentioned otherwise, we use $a \sim A$ for a finite set A to denote that a is a uniform random sample from the set A.)

Given Boolean functions f,g on n variables, the *Correlation* between f and g—denoted $\mathrm{Corr}(f,g)$—is defined as

$$\mathrm{Corr}(f,g) := |\langle f,g\rangle| = \left|\mathop{\mathbf{E}}_{x\sim\{-1,1\}^n}[f(x)g(x)]\right| = \left|2\Pr_x[f(x)=g(x)]-1\right|.$$

Also, we use $\delta(f,g)$ to denote the fractional distance between f and g: i. e., $\delta(f,g) = \Pr_x[f(x) \neq g(x)]$. Note that for Boolean f,g, we have $\mathrm{Corr}(f,g) = |1-2\delta(f,g)|$. We say that f is δ-approximated by g if $\delta(f,g) \leq \delta$.

We use Par_n to denote the parity function on n variables. I.e., $\mathrm{Par}_n(x_1,\ldots,x_n) = \prod_{i=1}^n x_i$.

Definition 2.1 (Restrictions)**.** A restriction on n variables is a function $\rho : [n] \to \{-1,1,*\}$. A random restriction is a distribution over restrictions. We use $\mathcal{R}_p^n$ to denote the distribution over restrictions on n variables obtained by setting each $\rho(x) = *$ with probability p and to 1 and -1 with probability $(1-p)/2$ each. We will often view the process of sampling a restriction from $\mathcal{R}_p^n$ as picking a pair (I,y) where $I \subseteq [n]$ is obtained by picking each element of $[n]$ to be in I with probability p and $y \in \{-1,1\}^{n-|I|}$ is chosen uniformly at random.

Definition 2.2 (Restriction trees and Decision trees)**.** A *restriction tree* T on $\{-1,1\}^n$ of depth h is a binary tree of depth h all of whose internal nodes are labeled by one of n variables, and the outgoing edges from an internal node are labeled $+1$ and -1; we assume that a node and its ancestor never query the same variable. Each leaf ℓ of T defines a restriction ρ_ℓ that sets all the variables on the path from the root of the decision tree to ℓ and leaves the remaining variables unset. A random restriction tree $\mathcal{T}$ of depth h is a distribution over restriction trees of depth h.

Given a restriction tree T, the process of choosing a random edge out of each internal node generates a distribution over the leaves of the tree (note that this distribution is not uniform: the weight it puts on leaf ℓ at depth d is 2^{-d}). We use the notation $\ell \sim T$ to denote a leaf ℓ of T picked according this distribution.

A *decision tree* is a restriction tree all of whose leaves are labeled either by $+1$ or -1. We say a decision tree has size s if the tree has s leaves. We say a decision tree computes a function $f : \{-1,1\}^n \to \{-1,1\}$ if for each leaf ℓ of the tree, $f|_{\rho_\ell}$ is equal to the label of ℓ.

Fact 2.3 (Facts about correlation)**.** *Let $f,g,h : \{-1,1\}^n \to \{-1,1\}$ be arbitrary.*

1. $\mathrm{Corr}(f,g) \in [0,1]$.
2. *If* $\mathrm{Corr}(f,g) \le \varepsilon$ *and* $\delta(g,h) \le \delta$, *then* $\mathrm{Corr}(f,h) \le \varepsilon + 2\delta$.
3. *Let $\mathcal{T}$ be any random restriction tree. Then*

$$\mathrm{Corr}(f,g) \le \mathop{\mathbf{E}}_{T\sim\mathcal{T},\ell\sim T}\left[\mathrm{Corr}(f|_{\rho_\ell}, g|_{\rho_\ell})\right].$$

Definition 2.4 (Noise sensitivity and Variance [37])**.** Given a Boolean function $f : \{-1,1\}^n \to \{-1,1\}$ and a parameter $p \in [0,1]$, we define the *noise sensitivity of f with noise parameter p*—denoted $\mathrm{NS}_p(f)$—as follows. Pick $x \in \{-1,1\}^n$ uniformly at random and $y \in \{-1,1\}^n$ by negating (i. e., flipping) each bit of x independently with probability p; we define

$$\mathrm{NS}_p(f) = \Pr_{(x,y)}[f(x) \ne f(y)].$$

The variance of f—denoted $\mathrm{Var}(f)$—is defined to be $2\mathrm{NS}_{1/2}(f)$.

Fact 2.5. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be any Boolean function. Let*

$$p = \min\{\Pr_x[f(x) = 1], \Pr_x[f(x) = -1]\}$$

where x is chosen uniformly from $\{-1,1\}^n$. Then, $\mathrm{Var}(f) = \Theta(p)$.

Proposition 2.6. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be any Boolean function. Then,*

1. *for $p \le 1/2$,* $\mathrm{NS}_p(f) = (1/2)\,\mathbf{E}_{\rho\sim\mathcal{R}^n_{2p}}\left[\mathrm{Var}(f|_\rho)\right]$*;*
2. *if $p \ge 1/n$, then* $\mathrm{Corr}(f, \mathrm{Par}_n) \le O(\mathrm{NS}_p(f))$.

The above proposition seems to be folklore, but we couldn't find explicit proofs in the literature. For completeness we present the proof below.

Proof. For point 1, we know that

$$\mathrm{NS}_p(f) = \Pr_{(x,y)}[f(x) \ne f(y)]$$

where x and y are sampled as in Definition 2.4. Alternately, we may also think of sampling (x,y) in the following way: choose $\rho = (I,z) \sim \mathcal{R}^n_{2p}$ and for the locations indexed by I we choose $x', y' \in \{-1,1\}^{|I|}$ independently and uniformly at random to define strings x and y respectively. Hence, we have

$$\mathrm{NS}_p(f) = \Pr_{x,y}[f(x) \neq f(y)] = \mathop{\mathbf{E}}_{\rho}\left[\Pr_{x',y'}\left[f|_\rho(x') \neq f|_\rho(y')\right]\right] = \mathop{\mathbf{E}}_{\rho}\left[\frac{1}{2}\mathrm{Var}(f|_\rho)\right].$$

We now proceed with point 2. As $\mathrm{NS}_p(f)$ is a decreasing function of p [37], we may assume that $p = 1/n \leq 1/2$ and hence we have

$$\mathrm{NS}_{1/n}(f) = \frac{1}{2}\mathop{\mathbf{E}}_{\rho\sim\mathcal{R}^n_{2/n}}\left[\mathrm{Var}(f|_\rho)\right].$$

Note that for $\rho = (I,y)$ chosen as above, the probability that $I \neq \emptyset$ is $\Omega(1)$. Hence we have

$$\mathrm{NS}'_{1/n}(f) := \frac{1}{2}\mathop{\mathbf{E}}_{\rho\sim\mathcal{R}^n_{2/n}}\left[\mathrm{Var}(f|_\rho) \mid I \neq \emptyset\right] \leq \frac{\mathrm{NS}_{1/n}(f)}{\Pr_I[I \neq \emptyset]} = O(\mathrm{NS}_{1/n}(f)).$$

Further, note that for any $m \geq 1$ and any Boolean function $g : \{-1,1\}^m \to \{-1,1\}$, its distance from either the constant function 1 or the constant function -1 is at most $\mathrm{Var}(g)/2$. Since Par_m has correlation 0 with any constant function, using Fact 2.3, we have $\mathrm{Corr}(\mathrm{Par}_m, g) \leq \mathrm{Var}(g)/2$.

Using Fact 2.3 again, we get

$$\begin{aligned}\mathrm{Corr}(\mathrm{Par}_n, f) &\leq \mathop{\mathbf{E}}_{\rho\sim\mathcal{R}^n_{2/n}}\left[\mathrm{Corr}(\mathrm{Par}_n|_\rho, f|_\rho) \mid I \neq \emptyset\right] = \mathop{\mathbf{E}}_{\rho\sim\mathcal{R}^n_{2/n}}\left[\mathrm{Corr}(\mathrm{Par}_{|I|}, f|_\rho) \mid I \neq \emptyset\right] \\ &\leq \mathop{\mathbf{E}}_{\rho\sim\mathcal{R}^n_{2/n}}\left[\frac{1}{2}\mathrm{Var}(g) \mid I \neq \emptyset\right] = \mathrm{NS}'_{1/n}(f) = O(\mathrm{NS}_{1/n}(f)).\end{aligned}$$ □

2.2 Threshold functions and circuits

Definition 2.7 (Threshold functions and gates)**.** A *Threshold gate* is a gate ϕ labeled with a pair (w,θ) where $w \in \mathbb{R}^m$ for some $m \in \mathbb{N}$ and $\theta \in \mathbb{R}$. The gate computes the Boolean function $f_\phi : \{-1,1\}^m \to \{-1,1\}$ defined by $f_\phi(x) = \mathrm{sgn}(\langle w,x\rangle - \theta)$ (we define $\mathrm{sgn}(0) = -1$ for the sake of this definition). The fan-in of the gate ϕ—denoted fan-in(ϕ)—is m. A *Linear Threshold function* (LTF) is a Boolean function that can be represented by a Threshold gate. More generally, a Boolean function $f : \{-1,1\}^n \to \{-1,1\}$ is a degree-D *Polynomial Threshold function* (PTF) if there is a degree-D polynomial $p(x)$ such that $f(x) = \mathrm{sgn}(p(x))$ for all $x \in \{-1,1\}^n$.

Definition 2.8 (Threshold circuits)**.** A *Threshold circuit* C is a Boolean circuit whose gates are all threshold gates. There are designated output gates, which compute the functions computed by the circuit. Unless explicitly mentioned, however, we assume that our threshold circuits have a unique output gate. The *gate complexity* of C is the number of (non-input) gates in the circuit, while the *wire complexity* is the sum of all the fan-ins of the various gates.

A *Threshold map* from n to m variables is a depth-1 threshold circuit C with n inputs and m outputs. We say that such a map is *read-k* if each input variable is an input to at most k of the threshold gates in C.

The following theorem bounds the number of Threshold functions of a few other arbitrary functions on n bits.

Theorem 2.9 ([48]). *Let $f_1, \dots, f_s$ be a collection of Boolean functions on n variables. Then there are at most $2^{O(ns)}$ distinct functions g of the form*

$$g(x_1, \dots, x_n) = \operatorname{sgn}\left(\sum_{i=1}^{s} w_i f_i(x_1, \dots, x_n) - \theta\right),$$

where $w_1, \dots, w_s, \theta \in \mathbb{R}$.

Applying the above theorem in the case that $s = n$ and the f_is are just the input bits, we obtain the following.

Corollary 2.10. *The number of distinct LTFs on n bits is at most $2^{O(n^2)}$.*

Definition 2.11 (Restrictions of threshold gates and circuits)**.** Given a threshold gate ϕ of fan-in m labeled by the pair (w, θ) and a restriction ρ on m variables, we use $\phi|_\rho$ to denote the threshold gate over the variables indexed by $\rho^{-1}(*)$ obtained in the natural way by setting variables according to ρ.

We will also need Peres' theorem, which bounds the noise sensitivity of threshold functions.

Theorem 2.12 (Peres' theorem [42, 37])**.** *Let $f : \{-1,1\}^n \to \{-1,1\}$ be any LTF. Then,*

$$\mathop{\mathbf{E}}_{\rho \sim \mathcal{R}_p^n}\left[\operatorname{Var}(f|_\rho)\right] = \mathrm{NS}_{\frac{p}{2}}(f) = O(\sqrt{p}).$$

Using the above for $p = 1/n$ and Proposition 2.6, we obtain

Corollary 2.13. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be any threshold function. Then $\operatorname{Corr}(f, \mathrm{Par}_n) \le O(n^{-1/2})$.*

2.3 Description lengths and Kolmogorov complexity

Definition 2.14 (Kolmogorov Complexity)**.** The Kolmogorov complexity of an n-bit Boolean string x is the length of the shortest bit string of the form (M, w) where M is the description of a Turing Machine and w an input to M such that $M(w) = x$. We use $K(x)$ to denote the Kolmogorov complexity of x.

Fact 2.15. *For any $\alpha \in (0,1)$, the fraction of n-bit strings x satisfying $K(x) \le (1-\alpha)n$ is at most $2^{-\alpha n+1}$.*

2.4 The Generalized Andreev function

We state here the definition of a generalization of Andreev's function, due to Komargodski and Raz, and Chen, Kabanets, Kolokolova, Shaltiel, and Zuckerman [29, 7]. This function will be used to give strong correlation bounds against constant-depth threshold circuits with slightly superlinear wire complexity.

We first need some definitions.

Definition 2.16 (Bit-fixing extractor)**.** A *subcube* $C \subseteq \{-1,1\}^n$ is a set obtained by restricting some subset of the n Boolean variables to some fixed values. A function $E : \{-1,1\}^n \to \{-1,1\}^m$ is a (n,k,m,ζ) *bit-fixing extractor* if for every random variable X that is uniform on a subcube of $\{-1,1\}^n$ of dimension at least k, the function $E(X)$ is ζ-close to uniform on $\{-1,1\}^m$.

We have the following explicit construction of a bit-fixing extractor.

Theorem 2.17 ([44])**.** *There is an absolute constant $c \geq 1$ so that the following holds. There is a polynomial-time computable function $E : \{-1,1\}^n \to \{-1,1\}^m$ that is an (n,k,m,ζ)-bit fixing extractor for any $k \geq (\log n)^c$, $m = 0.9k$, and $\zeta \leq 2^{-k^{\Omega(1)}}$.*

Also recall that a function $\mathrm{Enc} : \{-1,1\}^a \to \{-1,1\}^b$ defines (α,L)-*error-correcting code* for parameters $\alpha \in [0,1]$ and $L \in \mathbb{N}$ if for any $z \in \{-1,1\}^b$, the number of elements in the image of Enc that are at relative Hamming distance at most α from z is bounded by L.

The following theorem is folklore and is stated explicitly in the paper by Chen et al. [7].

Theorem 2.18 ([7], Theorem 6.4)**.** *Let $r = n^\beta$ for any fixed $0 < \beta < 1$. There exists an (α,L)-error correcting code with $\mathrm{Enc} : \{-1,1\}^{4n} \to \{-1,1\}^{2^r}$ where $\alpha = 1/2 - O(2^{-r/4})$ and $L = O(2^{r/2})$. Further, there is a $\mathrm{poly}(n)$ time algorithm, which when given as input $x \in \{-1,1\}^{4n}$ and $i \in [2^r]$ in binary, outputs $\mathrm{Enc}(x)_i$, the ith bit of $\mathrm{Enc}(x)$.*

Now we can define the Generalized Andreev function as in [7]. The function is $F : \{-1,1\}^{5n} \times \{-1,1\}^n \to \{-1,1\}$ and is defined as follows. Let $\gamma > 0$ be a constant parameter. The parameter will be fixed later according to the application at hand.

Let E be *any* $(n, n^\gamma, m = 0.9n^\gamma, 2^{-n^{\Omega(\gamma)}})$ extractor (we can obtain an explicit one using Theorem 2.17). We interpret the output of E as an integer from $[2^m]$ in the natural way. Let $\mathrm{Enc} : \{-1,1\}^{4n} \to \{-1,1\}^{2^m}$ define a $(1/2 - O(2^{-m/4}), 2^{m/2})$-error correcting code as in Theorem 2.18. Then, we define $F(x_1,x_2)$ by

$$F(x_1,x_2) = \mathrm{Enc}(x_1)_{E(x_2)}. \tag{2.1}$$

Given $a \in \{-1,1\}^{4n}$, we use $F_a(\cdot)$ to denote the resulting sub-function on n bits obtained by fixing $x_1 = a$.

The following lemma was proved as part of Theorem 6.5 in [7].

Lemma 2.19 ([7], Theorem 6.5)**.** *Let C be any circuit on n^γ variables with binary description length $\sigma(C) \leq n$ according to some fixed encoding scheme. Let ρ be any restriction of n variables leaving n^γ variables unfixed. Let $f(y) := F_a|_\rho(y)$ for $a \in \{-1,1\}^{4n}$ satisfying $K(a) \geq 3n$. Then*

$$\mathrm{Corr}(f,C) \leq \exp(-n^{\Omega(\gamma)}).$$

2.5 Concentration bounds

We state a collection of concentration bounds that we will need in our proofs. The proofs of Theorems 2.20 and 2.22 may be found in the excellent book by Dubhashi and Panconesi [13].

Theorem 2.20 (Chernoff bound)**.** *Let $w \in \mathbb{R}^n$ be arbitrary and x is chosen uniformly from $\{-1,1\}^n$. Then*

$$\Pr_x[|\langle w,x\rangle| \geq t \cdot \|w\|_2] \leq \exp(-\Omega(t^2)).$$

Definition 2.21 (Imbalance)**.** We say that a threshold gate ϕ labeled by (w,θ) is t-imbalanced if $|\theta| \geq t \cdot \|w\|_2$ and t-balanced otherwise.

We also need a multiplicative form of the Chernoff bound for sums of Boolean random variables.

Theorem 2.22 (Multiplicative Chernoff bound)**.** *Let $Y_1,\dots,Y_m$ be independent Boolean random variables such that $\mathbf{E}[Y_i] = p_i$ for each $i \in [m]$. Let p denote the average of the p_i. Then, for any $\varepsilon > 0$*

$$\Pr\left[|\sum_i Y_i - pm| \geq \varepsilon pm\right] \leq \exp(-\Omega(\varepsilon^2 pm)).$$

We will need the following variant of the Chernoff bound that may be found in the survey of Chung and Lu [9].

Theorem 2.23 (Theorem 3.4 in [9])**.** *Fix any parameter $p \in (0,1)$. Let $X_1,\dots,X_n$ be independent Boolean valued random variables such $\mathbf{E}[X_i] = p$ for each $i \in [n]$. Let $X = \sum_{i\in[n]} a_i X_i$ where each $a_i \geq 0$. Then,*

$$\Pr\left[X \geq \mathbf{E}[X] + \lambda\right] \leq \exp(-\lambda^2/(2(\nu + a\lambda))),$$

where $a = \max\{a_1,\dots,a_n\}$ and $\nu = p \cdot \left(\sum_i a_i^2\right)$. In the case that $\lambda \geq \mathbf{E}[X]$, we have $\nu \leq pa \cdot \sum_i a_i = a\mathbf{E}[X] \leq a\lambda$ and hence we obtain

$$\Pr\left[X \geq \mathbf{E}[X] + \lambda\right] \leq \exp(-\lambda/(4a)).$$

Let $Y_1,\dots,Y_m$ be random variables defined as functions of independent random variables $X_1,\dots,X_n$. For $i \in [m]$, let $S_i \subseteq [n]$ index those random variables among $X_1,\dots,X_n$ that influence Y_i. We say that $Y_1,\dots,Y_m$ are read-k random variables if any $j \in [n]$ belongs to S_i for at most k different $i \in [m]$.

The notation $D(p||q)$ represents the KL-divergence (see, e. g., [10]) between the two probability distributions on $\{0,1\}$ where the probabilities assigned to 1 are p and q respectively.

Theorem 2.24 (A read-k Chernoff bound [16])**.** *Let $Y_1,\dots,Y_m$ be $\{0,1\}$-valued read-k random variables such that $\mathbf{E}[Y_i] = p_i$. Let p denote the average of $p_1,\dots,p_m$. Then, for any $\varepsilon > 0$,*

$$\Pr\left[\sum_i Y_i \geq pm(1+\varepsilon)\right] \leq \exp\bigl(-D(p(1+\varepsilon) \parallel p)m/k\bigr).$$

It was pointed out to us by an anonymous reviewer that the above actually follows easily from an older, stronger result of Janson [25]. We cite the result from [16] since it is stated in the form above, which is useful for us.

Using standard estimates on the KL-divergence, Theorem 2.24 implies the following.

Corollary 2.25. *Let $Y_1,\dots,Y_m$ be as in the statement of Theorem 2.24 and assume $\mathbf{E}[\sum_i Y_i] \leq \mu$. Then,*

$$\Pr\left[\sum_i Y_i \geq 2\mu\right] \leq \exp(-\Omega(\mu/k)).$$

3 Correlation bounds against threshold circuits with small gate complexity

This section serves as a warm up to our main results in the following sections. Here we use a simple version of our technique to show that constant-depth threshold circuits with a small number of gates cannot correlate well with the Parity function. The following is a consequence of our main result.

Corollary 3.1. *Fix any $d \geq 2$. Assume that C is a depth-d threshold circuit over n variables with $k \leq n^{1/(2(d-1))}$ threshold gates and let $\delta = k/n^{1/(2(d-1))}$. Then,* $\mathrm{Corr}(C, \mathrm{Par}_n) \leq O(\delta^{(1-1/d)})$. *In particular, for any constant d,* $\mathrm{Corr}(C, \mathrm{Par}_n) = o(1)$ *unless $k = \Omega(n^{1/2(d-1)})$.*

It should be noted that Nisan [34] already proved stronger correlation bounds for the *Inner Product* function against *any* threshold circuit (not necessarily constant-depth) with a sub-linear (much smaller than $n/\log n$) number of threshold gates. The idea of his proof is to first show that each threshold gate on n variables has a δ-error randomized communication protocol with complexity $O(\log(n/\delta))$ [34, Theorem 1]. One can use this to show that any threshold circuit as in the theorem can be written as a decision tree of depth n/k querying threshold functions and hence has an $\exp(-\Omega(k))$-error protocol of complexity at most $n/10$. Standard results in communication complexity imply that any such function can have correlation at most $\exp(-\Omega(k))$ with inner product.

However, such techniques cannot be used to obtain lower bounds or correlation bounds for the parity function, since the parity function has low communication complexity, even in the deterministic setting. An even bigger disadvantage to this technique is that it cannot be used to obtain *any* superlinear lower bound on the wire complexity, since threshold circuits with a linear number of wires can easily compute functions with high communication complexity, such as the Disjointness function.

The techniques we use here can be used to give nearly tight [52] correlation bounds for the parity function (and can also be strengthened to the setting of small wire complexity, as we will show later). In fact, we prove something stronger: we upper bound the noise sensitivity of small constant-depth threshold circuits, which additionally implies the existence of non-trivial learning algorithms [28, 17] for such circuits. Further, our techniques also imply noise sensitivity bounds for AC^0 circuits augmented with a small number of threshold gates. For the sake of exposition, we postpone these generalizations to Section 8 and prove only the noise sensitivity result for constant-depth threshold circuits in this section.

3.1 Correlation bounds via noise sensitivity

The main result of this section is the following.

Theorem 3.2. *Let C be a depth-d threshold circuit with at most k threshold gates. Then, for any parameters $p, q \in [0,1]$, we have*

$$\mathrm{NS}_{p^{d-1}q}(C) \leq O(k\sqrt{p} + \sqrt{q}).$$

Proof. We assume that $q \leq 1/2$, since otherwise we always have $\mathrm{NS}_{p^{d-1}q}(C) \leq 1 \leq 2\sqrt{q}$.

We prove below that for $p_d := 2p^{d-1}q \in [0,1]$ and $\rho_d \sim \mathcal{R}^n_{p_d}$ (n is the number of input variables to C), we have

$$\mathop{\mathbf{E}}_{\rho_d}\left[\mathrm{Var}(C|_{\rho_d})\right] \leq O(k\sqrt{p} + \sqrt{q}). \tag{3.1}$$

This will imply the theorem, since by Proposition 2.6, we have

$$\mathrm{NS}_{p^{d-1}q}(C) = \frac{1}{2}\mathop{\mathbf{E}}_{\rho_d}\left[\mathrm{Var}(C|_{\rho_d})\right].$$

The proof of (3.1) is by induction on the depth d of the circuit. The base case $d=1$ is just Peres' theorem (Theorem 2.12).

Now assume that C has depth $d>1$. Let k_1 be the number of threshold circuits at depth $d-1$ in the circuit. We choose a random restriction $\rho \sim \mathcal{R}_p^n$ and consider the circuit $C|_\rho$. It is easy to check that

$$\mathop{\mathbf{E}}_{\rho_d}\left[\mathrm{Var}(C|_{\rho_d})\right] = \mathop{\mathbf{E}}_{\rho}\left[\mathop{\mathbf{E}}_{\rho_{d-1}}\left[\mathrm{Var}((C|_\rho)|_{\rho_{d-1}})\right]\right] \tag{3.2}$$

and hence to prove (3.1), it suffices to bound the expectation of $\mathrm{Var}((C|_\rho)|_{\rho_{d-1}})$.

Let us first consider the circuit $C|_\rho$. Peres' theorem tells us that on application of the restriction ρ, each threshold gate at depth $d-1$ becomes quite biased on average. Formally, by Theorem 2.12 and Fact 2.5, for each threshold gate ϕ at depth $d-1$, there is a bit $b_{\phi,\rho} \in \{-1,1\}$ such that

$$\mathop{\mathbf{E}}_{\rho}\left[\Pr_{x\in\{-1,1\}^{|\rho^{-1}(*)|}}\left[\phi|_\rho(x) \neq b_{\phi,\rho}\right]\right] \leq O(\sqrt{p}).$$

In particular, replacing $\phi|_\rho$ by $b_{\phi,\rho}$ in the circuit $C|_\rho$ yields a circuit that differs from $C|_\rho$ on only an $O(\sqrt{p})$ fraction of inputs (in expectation). Applying this replacement to each of the k_1 threshold gates at depth $d-1$ yields a circuit C'_ρ with $k-k_1$ threshold gates and depth $d-1$ such that

$$\mathop{\mathbf{E}}_{\rho}\left[\delta(C|_\rho, C'_\rho)\right] \leq O(k_1\sqrt{p}), \tag{3.3}$$

where $\delta(C|_\rho, C'_\rho)$ denotes the fraction of inputs on which the two circuits differ. On the other hand, we can apply the inductive hypothesis to C'_ρ to obtain

$$\mathop{\mathbf{E}}_{\rho_{d-1}}\left[\mathrm{Var}((C'_\rho)|_{\rho_{d-1}})\right] \leq O((k-k_1)\sqrt{p}+\sqrt{q}). \tag{3.4}$$

Therefore, to infer (3.1), we put the above together with (3.3) and the following elementary fact.

Proposition 3.3. *Say $f,g:\{-1,1\}^m \to \{-1,1\}$ and $\delta = \delta(f,g)$. Then, for any $r\in[0,1]$, we have*

$$\mathop{\mathbf{E}}_{\rho\sim\mathcal{R}_r^n}\left[\mathrm{Var}(f|_\rho)\right] \leq \mathop{\mathbf{E}}_{\rho\sim\mathcal{R}_r^n}\left[\mathrm{Var}(g|_\rho)\right] + 4\delta.$$

Proof of Proposition 3.3. By Proposition 2.6, we know that

$$\mathop{\mathbf{E}}_{\rho\sim\mathcal{R}_r^n}\left[\mathrm{Var}(f|_\rho)\right] = 2\mathrm{NS}_{r/2}(f),$$

and similarly for g. By definition of noise sensitivity, we have

$$\mathrm{NS}_{r/2}(f) = \Pr_{(x,y)}\left[f(x)\neq f(y)\right]$$

where $x \in \{-1,1\}^m$ is chosen uniformly at random and y is chosen by flipping each bit of x with probability $r/2$. Note that each of x and y is individually uniformly distributed over $\{-1,1\}^m$ and hence, both $f(x) = g(x)$ and $f(y) = g(y)$ hold with probability at least $1-2\delta$. This yields

$$\mathrm{NS}_{r/2}(f) = \Pr_{(x,y)}[f(x) \neq f(y)] \leq \Pr_{(x,y)}[g(x) \neq g(y)] + 2\delta = \mathrm{NS}_{r/2}(g) + 2\delta,$$

which implies the claimed bound. □

□

The above theorem yields the correlation bound for Parity stated above (Corollary 3.1) as we now show.

Proof of Corollary 3.1. We apply Theorem 3.2 with the following optimized parameters:

$$p = \frac{1}{n^{1/d}} \cdot \frac{1}{k^{2/d}}$$

and $q \in [0,1]$ such that $p^{d-1}q = 1/n$. It may be verified that for this setting of parameters, Theorem 3.2 gives us

$$\mathrm{NS}_{1/n}(C) \leq O\left(\frac{k^{1-1/d}}{n^{1/(2d)}}\right) \leq O(\delta^{1-\frac{1}{d}}).$$

As noted in Proposition 2.6, we have $\mathrm{Corr}(C, \mathrm{Par}_n) \leq O(\mathrm{NS}_{1/n}(C))$. This completes the proof. □

Remark 3.4. It is instructive to compare the above technique with the closely related work of Gopalan and Servedio [17]. The techniques of [17] applied to the setting of Theorem 3.2 show that $\mathrm{NS}_p(C) \leq O(k2^k\sqrt{p})$, which gives a better dependence on the noise parameter p, but a much worse dependence on k. Indeed, this is not surprising since in this setting, the technique of Gopalan and Servedio does not use the fact that the circuit is of depth d. The threshold circuit is converted to a decision tree of depth k querying threshold functions and it is this tree that is analyzed.

We believe that the right answer should incorporate the best of both bounds:

$$\mathrm{NS}_p(f) \leq O_d(k^{d-1} \cdot \sqrt{p}).$$

As in Corollary 3.1, this would show that $\mathrm{Corr}(C, \mathrm{Par}_n) = o(1)$ if $k = o(n^{1/2(d-1)})$, but additionally, we would also get $\mathrm{Corr}(C, \mathrm{Par}_n) \leq n^{-1/2+o(1)}$ as long as $k = n^{o(1)}$ and d is a constant.

It is known from the work of Siu, Roychowdhury and Kailath [52, Theorem 7] that Corollary 3.1 is tight in the sense that there do exist circuits of gate complexity roughly $n^{1/2(d-1)}$ that have significant correlation with Par_n. More formally,

Theorem 3.5 (Theorem 7 in [52])**.** *Let $\varepsilon > 0$ be an arbitrary constant. Then, there is a threshold circuit of depth d with $O(d) \cdot (n \log(1/\varepsilon))^{1/2(d-1)}$ gates that computes Par_n correctly on a $1-\varepsilon$ fraction of inputs.*

4 Correlation bounds against threshold circuits with small wire complexity

In this section, we prove correlation bounds against threshold circuits of small *wire complexity.* We prove three results of this kind.

The first result, proved in Section 4.1, is a near-optimal correlation bound against depth-2 circuits computing the Parity function. This result was obtained independently by Kane and Williams [27].

Theorem 4.1 (Correlation bounds against depth-2 threshold circuits). *Fix any constant $\varepsilon < 1/2$. Let $\gamma = (1/2) - \varepsilon$. Any depth-2 threshold circuit on n variables with at most $n^{1+\varepsilon}$ wires has correlation at most $n^{-\Omega(\gamma)}$ with the Parity function on n variables.*

Note that the above theorem is nearly tight, since by Theorem 3.5, there is a depth-2 circuit with $O(\sqrt{n})$ gates (and hence $O(n^{3/2})$ wires) that computes Parity on n variables correctly with high probability. The proof is a simple argument based on Peres' theorem along with a correlation bound against low-degree polynomial threshold functions due to Aspnes, Beigel, Furst and Rudich [3].

The advantage of the above technique is that it yields a near-optimal lower bound in the depth-2 case. Unfortunately, however, it does not directly extend to depths larger than 2. The main results of the section, stated below, show how to obtain correlation bounds against all constant depths.

The second result, proved in Section 4.2 via a more involved argument, yields a correlation bound for the parity function against all constant depths.

Theorem 4.2 (Correlation bounds for parity). *For any $d \geq 1$, there is an $\varepsilon_d = 1/2^{O(d)}$ such that any depth-d threshold circuit C with at most $n^{1+\varepsilon_d}$ wires satisfies $\mathrm{Corr}(C, \mathrm{Par}_n) \leq O(n^{-\varepsilon_d})$ where the $O(\cdot)$ hides absolute constants (independent of d and n).*

For comparison, we note that a result of Paturi and Saks [41] shows that the Parity function can be computed by threshold circuits of depth d with $n^{1+2^{-\Omega(d)}}$ many wires. A *worst-case* lower bound of a similar form was proved by Impagliazzo et al. [24].

Finally, we are also able to extend the techniques used in the proof of the above theorem to prove exponentially small correlation bounds. This kind of correlation bound cannot be proved for the Parity function since the Parity function on n variables has correlation $\Omega(1/\sqrt{n})$ with the Majority function on n variables (see, e. g., [37, Section 5.3]), which clearly has a threshold circuit with only n wires.

We prove such a correlation bound in Section 4.3 for the Generalized Andreev function from Section 2.4. For some technical reasons, we prove a slightly stronger result (Theorem 4.13) from which we obtain the following consequence.

Corollary 4.3 (Correlation bounds for Andreev's function). *For any constant $d \geq 1$, there is an $\varepsilon_d = 1/2^{O(d)}$ such that the following holds. Let F be the Generalized Andreev function on $5n$ variables as defined in Section 2.4 for any constant $\gamma < 1/6$. Any depth-d threshold circuit C of wire complexity at most $n^{1+\varepsilon_d}$ satisfies $\mathrm{Corr}(C, F) \leq \exp(-n^{\Omega(\varepsilon_d)})$ where the $\Omega(\cdot)$ hides constants independent of d and n.*

We now state a key lemma that will be used in the proofs of our correlation bounds in Sections 4.2 and 4.3. The lemma will be proved in Section 5.

Recall that a threshold gate with label (w, θ) is t-balanced if $|\theta| \leq t \cdot \|w\|_2$.

Lemma 4.4 (Main Structural lemma for threshold gates)**.** *The following holds for some absolute constant $p_0 \in [0,1]$. For any threshold gate ϕ over n variables with label (w,θ) and any $p \in [0,p_0]$, we have*[2]

$$\Pr_{\rho \sim \mathcal{R}_p^n}\left[\phi|_\rho \text{ is } \frac{1}{p^{\Omega(1)}}\text{-balanced}\right] \leq p^{\Omega(1)}.$$

4.1 Correlation bounds against depth-2 threshold circuits computing Parity

In this section, we prove Theorem 4.1. The proof is based on the following two subclaims.

Theorem 4.5 (Aspnes, Beigel, Furst, and Rudich [3])**.** *Any degree-t polynomial threshold function (PTF)*[3] *has correlation at most $O(t/\sqrt{m})$ with the parity function on m variables.*

Lemma 4.6. *Let ε, γ be as in the statement of Theorem 4.1 and let α denote $\gamma/3$. Say C denotes a depth-2 threshold circuit of wire complexity $n^{1+\varepsilon}$ and let $f_1,\ldots,f_t$ be the LTFs computed by C at depth-1. Under a random restriction ρ with $*$-probability $p = 1/n^{1-\alpha}$, with probability at least $1-n^{-\Omega(\gamma)}$, the circuit $C|_\rho$ is $n^{-\Omega(\gamma)}$-approximated by a circuit $\tilde{C}_\rho$ which is obtained from C by replacing each of the $f_i|_\rho$s by an $O(n^{\alpha/2-\Omega(\gamma)})$-junta g_i.*

Assuming the above two claims, we can finish the proof of Theorem 4.1 easily as follows.

Let C be a circuit of wire complexity $n^{1+\varepsilon}$. We apply a random restriction ρ with $*$-probability $p = 1/n^{1-\alpha}$ as in Lemma 4.6. Call the restriction *good* if there is a circuit $\tilde{C}_\rho$ as in the lemma that $n^{-\Omega(\gamma)}$-approximates $C|_\rho$ and $|\rho^{-1}(*)| \geq n^\alpha/2$. The probability that the first of these events does not occur is at most $n^{-\Omega(\gamma)}$ by Lemma 4.6 and the probability of the second is at most $\exp(\Omega(n^{-\alpha}))$ by a Chernoff bound (Theorem 2.22). Thus, the probability that ρ is not good is at most

$$n^{-\Omega(\gamma)} + \exp(-\Omega(n^{-\alpha})) \leq n^{-\Omega(\gamma)}.$$

Say ρ is a good restriction. Note that each restricted LTF at depth-1 in the circuit $\tilde{C}_\rho$ is an $O(n^{\alpha/2-\Omega(\gamma)})$-junta and hence can be represented *exactly* by a polynomial of the same degree. This implies that $\tilde{C}_\rho$ is a $O(n^{\alpha/2-\Omega(\gamma)})$-degree PTF and hence, by Theorem 4.5, has correlation at most $n^{-\Omega(\gamma)}$ with the Parity function (on the remaining $n^\alpha/2$ variables). Moreover, then $C|_\rho$ is well-approximated by $\tilde{C}_\rho$ and hence has correlation at most $n^{-\Omega(\gamma)} + n^{-\Omega(\gamma)}$ with parity.

Upper bounding the correlation by 1 for bad restrictions, we see that the overall correlation is at most $n^{-\Omega(\gamma)}$.

We now prove Lemma 4.6.

Proof of Lemma 4.6. Let $f_1,\ldots,f_t$ be the LTFs appearing at depth 1 in the circuit. We will divide the analysis based on the fan-ins of the f_is (i. e., the number of variables they depend on).

We denote by β the quantity $3/4 + \varepsilon/2$. It can be checked that we have both

$$\beta = \frac{1}{2} + \varepsilon + \frac{\alpha}{2} + \Omega(\gamma) \qquad \text{and} \qquad 1 - \beta = \frac{\alpha}{2} + \Omega(\gamma). \tag{4.1}$$

[2]Recall from Definition 2.11 that $\phi|_\rho$ is the threshold gate obtained by setting variables outside $\rho^{-1}(*)$ according to ρ.
[3]We refer the reader to Section 2.2 for the definition of PTFs.

Consider any f_i of fan-in at most n^β. When hit with a random restriction with $*$-probability $n^{-(1-\alpha)}$, we see that the expected number of variables of f_i that survive is at most $n^{\beta-(1-\alpha)} = n^{\alpha-(1-\beta)} = n^{\alpha/2-\Omega(\gamma)}$ by (4.1) above. By a Chernoff bound (Theorem 2.22), the probability that this number exceeds twice its expectation is exponentially small. Union bounding over all the gates of small fan-in, we see that with probability $1-\exp(-n^{\Omega(1)})$, all the low fan-in gates depend on at most $2n^{\alpha/2-\Omega(\gamma)}$ many variables after the restriction. We call this high probability event $\mathcal{E}_1$.

Now, we consider the gates of fan-in at least n^β. Without loss of generality, let $f_1,\ldots,f_r$ be these LTFs. Since the total number of wires is at most $n^{1+\varepsilon}$, we have $r \le n^{1+\varepsilon-\beta} = n^{1/2-\alpha/2-\Omega(\gamma)}$ by (4.1).

By Theorem 2.12, we know that for any f_i,

$$\mathop{\mathbf{E}}_{\rho}\left[\mathrm{Var}(f_i|_\rho)\right] \le O(\sqrt{p}) = O\left(\frac{1}{n^{(1-\alpha)/2}}\right).$$

By linearity of expectation, we have

$$E := \mathop{\mathbf{E}}_{\rho}\left[\sum_{i=1}^{r}\mathrm{Var}(f_i|_\rho)\right] \le O\left(r\cdot\frac{1}{n^{(1-\alpha)/2}}\right) = O\left(n^{(1-\alpha)/2-\Omega(\gamma)}\cdot\frac{1}{n^{(1-\alpha)/2}}\right) = O(n^{-\Omega(\gamma)}).$$

By Markov's inequality, we see that the probability that

$$\sum_{i=1}^{r}\mathrm{Var}(f_i|_\rho) > \sqrt{E}$$

is at most $\sqrt{E} = n^{-\Omega(\gamma)}$. We let $\mathcal{E}_2$ denote the event that

$$\sum_{i=1}^{r}\mathrm{Var}(f_i|_\rho) \le \sqrt{E}.$$

Consider the event $\mathcal{E} = \mathcal{E}_1 \wedge \mathcal{E}_2$. A union bound tells us that the probability of $\mathcal{E}$ is at least $1-n^{-\Omega(\gamma)}$. When this event occurs, we construct the circuit $\tilde{C}_\rho$ from the statement of the claim as follows.

When the event $\mathcal{E}$ occurs, the LTFs of low arity are already $n^{\alpha/2-\Omega(\gamma)}$-juntas, so there is nothing to be done for them.

Now, consider the LTFs of high fan-in, which are $f_1,\ldots,f_r$. For each $f_i|_\rho$ ($i\in[r]$), replace f_i by a bit $b_i \in \{-1,1\}$ such that

$$\Pr_x\left[f_i|_\rho(x) = b_i\right] \ge \frac{1}{2}.$$

In the circuit $\tilde{C}_\rho$, these gates thus become constants, which are 0-juntas. The circuit $\tilde{C}_\rho$ now has the required form. We now analyze the error introduced by this operation.

We know that

$$\Pr_x\left[f_i|_\rho(x) \ne b_i\right] \le 2\mathrm{Var}(f_i|_\rho)$$

and thus the overall error introduced is at most

$$2\sum_{i\in[r]}\mathrm{Var}(f_i|_\rho) \le O(\sqrt{E}) = n^{-\Omega(\gamma)}$$

(since $\mathcal{E}_2$ is assumed to occur). Thus, the circuit $\tilde{C}_\rho$ is an $n^{-\Omega(\gamma)}$-approximation to C. □

4.2 Correlation bounds for Parity against constant-depth circuits

In this section, we prove Theorem 4.2, assuming Lemma 4.4. The proof proceeds by iteratively reducing the depth of the circuit. In order to perform this depth-reduction for a depth-d circuit, we need to analyze the threshold map defined by the threshold gates at depth $d-1$. The first observation, which follows from Markov's inequality, shows that we may assume (after setting a few variables) that the map reads each variable only a few times.

Fact 4.7 (Small wire-complexity to small number of reads)**.** *Let C be any threshold circuit on n variables with wire complexity at most cn. Then, there is a set S of at most $n/2$ variables such that each variable outside S is an input variable to at most $2c$ many gates in C.*

The second observation is that if the fan-ins of all the threshold gates are small, then depth-reduction is easy (after setting some more variables).

Proposition 4.8 (Handling small fan-in gates)**.** *Let $C = (\phi_1, \ldots, \phi_m)$ be any read-k threshold map on n variables such that $\max_i \text{fan-in}(\phi_i) \leq t$. Then, there is a set S of n/kt variables such that each ϕ_i depends on at most one variable in S.*

Proof. This may be done via a simple graph theoretic argument. Define an undirected graph whose vertex set is the set of n variables and two variables are adjacent iff they feed into the same threshold gate. We need to pick an S that is an independent set in this graph. Since the graph has degree at most kt, we can greedily find an independent set of size at least n/kt. Let S be such an independent set. □

Let $B > 2$ be a constant real parameter that we will choose to satisfy various constraints in the proofs below. For $d \geq 1$, define

$$\varepsilon_d = B^{-(2d-1)} \qquad \text{and} \qquad \delta_d = B\varepsilon_d. \tag{4.2}$$

Theorem 4.2 (Correlation bounds for Parity (restated))**.** For any $d \geq 1$ and $c \leq n^{\varepsilon_d}$, any depth-d threshold circuit C with at most cn wires satisfies $\mathrm{Corr}(C, \mathrm{Par}_n) \leq O(n^{-\varepsilon_d})$ where the $O(\cdot)$ hides absolute constants (independent of d and n).

Proof idea. The proof will be an induction on d. Assume that C has $n^{1+\varepsilon}$ many wires for a small $\varepsilon > 0$. Our aim is to apply a restriction that leaves plenty of variables alive (i. e., set to $*$) while at the same time reduces the depth of the circuit from d to $d-1$.

We first apply a random restriction ρ with $*$-probability $n^{-\delta}$ for a suitable $\delta > 0$ to the circuit C and analyze the effect of ρ on the threshold gates in C at depth $d-1$ (just above the variables).

By Lemma 4.4, we see that with high probability, each threshold gate ϕ at depth $d-1$ becomes imbalanced and hence (by the Chernoff bound (Theorem 2.20)) highly biased. Such gates can be replaced by constants without noticeably changing the correlation with the parity function. Now the only gates at depth $d-1$ are the balanced gates.

It remains to handle the balanced gates. Consider a gate ϕ of C with initial fan-in k. The expected number of input wires to ϕ that survive the random restriction ρ is $k \cdot n^{-\delta}$. Further by Lemma 4.4, the probability that ϕ remains biased is at most $n^{-\Omega(\delta)}$. While these two events are not independent, we can argue that their effects are somewhat independent in the following sense. We argue that the random variable Y that is equal to the fan-in of $\phi|_\rho$ if ϕ is balanced and 0 otherwise has expectation roughly $k \cdot n^{-\delta} \cdot n^{-\Omega(\delta)}$. Hence, the expected number of wires W that feed into balanced gates is at most $n^{1+\varepsilon} \cdot n^{-\delta} \cdot n^{-\Omega(\delta)}$ where the first term accounts for the total fan-ins of all the gates at depth $d-1$. For δ suitably larger than ε, the expectation of W is much smaller than $n^{1-\delta}$, the expected number of surviving variables. Thus, if we set all the variables corresponding to the wires feeding into the balanced gates, we can set all the balanced gates to constants. In the process, we don't set too many live variables (since the expectation of W is much smaller than $n^{1-\delta}$) and further we reduce the depth of the circuit to $d-1$.

We can now apply the induction hypothesis to bound the correlation of the resulting circuit with the Parity function on (roughly) $n^{1-\delta}$ variables to get the correlation bound against depth-d circuits.

Proof. The proof is by induction on the depth d of C. The base case is $d = 1$, which is the case when C is only a single threshold gate. In this case, Corollary 2.13 tells us that $\mathrm{Corr}(C, \mathrm{Par}_n) \le O(n^{-1/2}) \le n^{-\varepsilon_1}$, since $B > 2$.

Now, we handle the inductive case when the depth $d > 1$. Our analysis proceeds in phases.

Phase 1. We first transform the circuit into a read-$2c$ circuit by setting $n/2$ variables. This may be done by Fact 4.7. This defines a restriction tree of depth $n/2$. By Fact 2.3, it suffices to show that each leaf of this restriction tree, the correlation of the restricted circuit and $\mathrm{Par}_{n/2}$ remains bounded by $O(n^{-\varepsilon_d})$.

Let n_1 now denote the new number of variables and let C_1 now be the restricted circuit at some arbitrary leaf of the restriction tree. By renaming the variables, we assume that they are indexed by the set $[n_1]$.

Phase 2. Let $\phi_1, \ldots, \phi_m$ be the threshold gates at depth $d-1$ in the circuit C_1. We call ϕ_i *large* if $\text{fan-in}(\phi_i) > n^{\delta_d}$ and *small* otherwise. Let $L \subseteq [m]$ be defined by $L = \{i \in [m] \mid \phi_i \text{ large}\}$. Assume that $|L| = \ell$. Note that $\ell \cdot n^{\delta_d} \le n^{1+\varepsilon_d}$ and hence $\ell \le n^{1+\varepsilon_d - \delta_d} \le n$.

We restrict the circuit with a random restriction $\rho = (I, y) \sim \mathcal{R}_p^{n_1}$, where $p = n^{-\delta_d/2}$. By Lemma 4.4, we know that for each $i \in [m]$ and some $t = 1/p^{\Omega(1)}$ and $q = p^{\Omega(1)}$,

$$\Pr_\rho \left[\phi_i|_\rho \text{ is } t\text{-balanced}\right] \le q. \tag{4.3}$$

Further, we also know that for each $i \in L$, the expected value of $\text{fan-in}(\phi_i|_\rho) = p \cdot \text{fan-in}(\phi_i)$, since each variable is set to a constant with probability $1-p$. Since $i \in L$, the expected fan-in of each ϕ_i ($i \in L$) is at least $n^{\delta_d/2}$. Hence, by a Chernoff bound (Theorem 2.22), we see that for any $i \in L$,

$$\Pr_\rho \left[\text{fan-in}(\phi_i|_\rho) > 2p \cdot \text{fan-in}(\phi_i)\right] \le \exp(-\Omega(n^{\delta_d/2})). \tag{4.4}$$

Finally another Chernoff bound (Theorem 2.22) tells us that

$$\Pr_{\rho=(I,y)} \left[|I| < \frac{n_1 p}{2}\right] \le \exp(-\Omega(n_1 p)) = \exp(-\Omega(np)). \tag{4.5}$$

We call a set I *generic* if $|I| \ge n_1 p/2$ and $\text{fan-in}(\phi_i|_\rho) \le 2p \cdot \text{fan-in}(\phi_i)$ for each $i \in L$. Let $\mathcal{G}$ denote the event that I is generic. By (4.4) and (4.5), we know that

$$\Pr_I [\neg\mathcal{G}] \le \ell \exp(-\Omega(n^{\delta_d/2})) + \exp(-\Omega(np)) \le \exp(-n^{\delta_d/4}).$$

In particular, conditioning on $\mathcal{G}$ doesn't change (4.3) by much.

$$\Pr_{\rho=(I,y)}\left[\phi_i|_\rho \text{ is } t\text{-balanced} \;\middle|\; \mathcal{G}\right] \leq q + \exp(-n^{\delta_d/4}) \leq 2q. \tag{4.6}$$

Our aim is to further restrict the circuit by setting *all* the input variables to the gates ϕ_i that are t-balanced. In order to analyze this procedure, we define random variables Y_i $(i \in L)$ so that $Y_i = 0$ if $\phi_i|_\rho$ is t-imbalanced and fan-in$(\phi_i|_\rho)$ otherwise. Let $Y = \sum_{i\in L} Y_i$. Note that

$$\mathop{\mathbf{E}}_{\rho}[Y_i \mid \mathcal{G}] \leq (2p\cdot \text{fan-in}(\phi_i))\cdot \Pr_{\rho}\left[\phi_i|_\rho \text{ is } t\text{-balanced} \;\middle|\; \mathcal{G}\right] \leq 4pq\cdot \text{fan-in}(\phi_i)$$

where the first inequality follows from the fact that since we have conditioned on I being generic, we have

$$\text{fan-in}(\phi_i|_\rho) \leq 2p\cdot \text{fan-in}(\phi_i)$$

with probability 1. Hence, we have

$$\mathop{\mathbf{E}}_{\rho}[Y \mid \mathcal{G}] \leq 4pq\cdot \sum_i \text{fan-in}(\phi_i) \leq 4pq\cdot n^{1+\varepsilon_d}. \tag{4.7}$$

We let $\mu := 4pq\cdot n^{1+\varepsilon_d}$. By Markov's inequality,

$$\Pr_{\rho}\left[Y \geq \frac{\mu}{\sqrt{q}} \;\middle|\; \mathcal{G}\right] \leq \sqrt{q}. \tag{4.8}$$

In particular, we can condition on a *fixed* generic $I \subseteq [n]$ such that for random $y \sim \{-1,1\}^{n_1-|I|}$, we have

$$\Pr_{y}\left[Y \geq \frac{\mu}{\sqrt{q}}\right] \leq \sqrt{q}.$$

The above gives us a restriction tree T (that simply sets all the variables in $[n_1]\setminus I$) such that at all but $1-2\sqrt{q}$ fraction of leaves λ of T, the total fan-in of the large gates at depth 1 in C_1 that are t-balanced is at most $\mu/\sqrt{q}$; call such λ *good* leaves. Let n_2 denote $|I|$, which is the number of surviving variables.

Phase 3. We will show that for any good leaf λ, we have

$$\text{Corr}(C_\lambda, \text{Par}_{n_2}) \leq n^{-\varepsilon_d} \tag{4.9}$$

where C_λ denotes $C_1|_{\rho_\lambda}$. This will prove the theorem, since we have by Fact 2.3,

$$\begin{aligned}\text{Corr}(C_1, \text{Par}_{n_1}) &\leq \mathop{\mathbf{E}}_{\lambda\sim T}[\text{Corr}(C_\lambda, \text{Par}_{n_2})]\\ &\leq \Pr_{\lambda}[\lambda \text{ not good}] + \max_{\lambda \text{ good}} \text{Corr}(C_\lambda, \text{Par}_{n_2})\\ &\leq 2\sqrt{q} + n^{-\varepsilon_d} \leq 2n^{-\varepsilon_d}\end{aligned}$$

where we have used the fact that $\text{Par}_{n_1}|_{\rho_\lambda} = \pm\text{Par}_{n_2}$ for each leaf λ, and also that $2\sqrt{q} \leq n^{-\varepsilon_d}$ for a large enough choice of the constant B.

It remains to prove (4.9). We do this in two steps.

In the first step, we set all large t-imbalanced gates to their most probable constant values. Formally, for a t-imbalanced threshold gate ϕ labeled by (w,θ), we have $|\theta| \geq t \cdot \|w\|_2$. We replace ϕ by a constant b_ϕ which is 1 if $\theta \geq t \cdot \|w\|_2$ and by -1 if $-\theta \geq t \cdot \|w\|_2$. This turns the circuit C_λ into a circuit C'_λ of at most the wire complexity of C_λ. Further, note that for any $x \in \{-1,1\}^{n_1}$, $C_\lambda(x) = C'_\lambda(x)$ unless there is a t-imbalanced threshold gate ϕ such that $\phi(x) \neq b_\phi(x)$. By the Chernoff bound (Theorem 2.20) the probability that this happens for any fixed imbalanced threshold gate is at most

$$\exp(-\Omega(t^2)) \leq \exp(-n^{\Omega(\delta_d)}).$$

By a union bound over the $\ell \leq n$ large threshold gates, we see that

$$\Pr_x\left[C_\lambda(x) \neq C'_\lambda(x)\right] \leq n\exp(-n^{\Omega(\delta_d)}).$$

In particular, we get by Fact 2.3

$$\mathrm{Corr}(C_\lambda, \mathrm{Par}_{n_2}) \leq \mathrm{Corr}(C'_\lambda, \mathrm{Par}_{n_2}) + n\exp(-n^{\Omega(\delta_d)}) \leq \mathrm{Corr}(C'_\lambda, \mathrm{Par}_{n_2}) + \exp(-n^{\varepsilon_d}) \tag{4.10}$$

where the last inequality is true for a large enough constant B.

In the second step, we further define a restriction tree T_λ such that C'_λ becomes a depth-$(d-1)$ circuit with at most cn wires at *all the leaves* of T_λ. We first restrict by setting all variables that feed into *any* of the t-balanced gates. The number of variables set in this way is at most

$$\frac{\mu}{\sqrt{q}} \leq 4p\sqrt{q}\cdot n^{1+\varepsilon_d} \leq (pn)\cdot(4\sqrt{q}n^{\varepsilon_d}) \leq \frac{pn}{8} \leq \frac{n_2}{2}$$

for a large enough choice of the constant B. This leaves $n_3 \geq n_2/2$ variables still alive. Further, all the large t-balanced gates are set to constants *with probability* 1. Finally, by Proposition 4.8, we may set all but a set S of $n_4 = n_3/2cn^{\delta_d}$ variables to ensure that with probability 1, all the small gates depend on at most one input variable each. At this point, the circuit C'_λ may be transformed to a depth-$(d-1)$ circuit C''_λ with at most as many wires as C'_λ, which is at most cn.

Note that the number of unset variables is $n_4 \geq pn/8cn^{\delta_d} \geq n^{1-2\delta_d}$, for large enough B. Hence, the number of wires is at most

$$cn \leq n_4^{\frac{1+\varepsilon_d}{1-2\delta_d}} \leq n_4^{(1+\varepsilon_d)(1+3\delta_d)} \leq n_4^{1+\varepsilon_{d-1}}$$

for suitably large B. Thus, by the inductive hypothesis, we have

$$\mathrm{Corr}(C''_\lambda, \mathrm{Par}_{n_4}) \leq O(n_4^{-\varepsilon_{d-1}}) \leq n^{-\varepsilon_d}/2$$

with probability 1 over the choice of the variables restricted in the second step. Along with (4.10) and Fact 2.3, this implies (4.9) and hence the theorem. □

4.3 Exponential correlation bounds for the Generalized Andreev function

We now prove an exponentially strong correlation bound for the Generalized Andreev function defined in Section 2.4 with any $\gamma < 1/6$. For any $d \geq 1$, we define the constants ε_d and δ_d as in (4.2), where $B > 2$ is a large constant that will be chosen below.

Proof overview. As in the case of Theorem 4.2, the proof proceeds by an iterative depth reduction. The proof of the depth reduction here is technically more challenging than the proof of Theorem 4.2 since we are trying to prove exponentially small correlation bounds and hence, we can no longer afford to ignore "bad" events that occur with polynomially small probability. This forces us to prove a more involved depth-reduction statement, which we prove as a separate lemma (Lemma 4.11 below). We show how to use this depth-reduction statement to prove our correlation bound in Theorem 4.13 and Corollary 4.3.

We begin with a definition that will be required to state our depth-reduction lemma.

Definition 4.9 (Simplicity)**.** We call a threshold circuit C (t,d,w)-*simple* if there is a set R of $r \le t$ threshold functions $g_1,\ldots,g_r$ such that for every setting of these threshold functions to bits $b_1,\ldots,b_r$, the circuit C can be represented on the corresponding inputs (i. e., inputs x satisfying $g_i(x) = b_i$ for each $i \in [r]$) by a depth-d threshold gate of wire complexity at most w.

In particular, note that a (t,d,w)-simple circuit C may be expressed as

$$C(x) = \bigvee_{b_1,\ldots,b_r \in \{-1,1\}} \left(C_{b_1,\ldots,b_r} \wedge \bigwedge_{i:b_i=-1} g_i \wedge \bigwedge_{i:b_i=1} (\neg g_i) \right). \tag{4.11}$$

where each $C_{b_1,\ldots,b_r}$ is a depth-d circuit of wire complexity at most w. Further, note that the OR appearing in the above expression is disjoint (i. e., no two terms in the OR can be simultaneously true).

We need the following elementary property of disjoint ORs.

Proposition 4.10. *Let $g_1,\ldots,g_N$ be Boolean functions defined on $\{-1,1\}^n$ such that no two of them are simultaneously true and let h denote $\bigvee_{i\in[N]} g_i$. Then, for any $f : \{-1,1\}^n \to \{-1,1\}$, we have*

$$\mathrm{Corr}(f,h) \le (N-1)\mathrm{Corr}(f,1) + \sum_{i\in[N]} \mathrm{Corr}(f,g_i),$$

where 1 *denotes the constant* 1 *function.*

Proof. Note that h being the disjoint OR of $g_1,\ldots,g_N$ translates into the following identity:

$$\frac{1-h}{2} = \sum_{i\in[N]} \frac{1-g_i}{2}.$$

Hence, we have $h = (\sum_i g_i) - (N-1)$. Thus, using the bilinearity of the inner product and the triangle inequality

$$\mathrm{Corr}(f,h) = |\langle f,h\rangle| \le \sum_{i\in[N]} \mathrm{Corr}(f,g_i) + (N-1)\mathrm{Corr}(f,1). \qquad \square$$

We are now ready to prove our main depth-reduction lemma for threshold circuits with small wire complexity.

Lemma 4.11 (Depth-reduction lemma with exponentially small failure probability)**.** *Let $d \ge 1$ be any constant and assume that ε_d, δ_d are defined as in (4.2) with B large enough. Say we are given any depth-d threshold circuit C on n variables with at most $n^{1+\varepsilon_d}$ wires.*

There is a restriction tree T of depth $n - n^{1-2\delta_d}$ with the following property: for a random leaf $\lambda \sim T$, let $\mathcal{E}(\lambda)$ denote the event that the circuit $C|_{\rho_\lambda}$ is $\exp(-n^{\varepsilon_d})$-approximated by an $(n^{\delta_d}, d-1, n^{1+\varepsilon_d})$-simple circuit. Then, $\Pr_\lambda[\neg\mathcal{E}(\lambda)] \le \exp(-n^{\varepsilon_d})$.

Comparison with Theorem 4.2. The proof of this lemma is similar to the proof of the depth-reduction in Theorem 4.2, with a few key differences. The main change is that we require the simplification of depth-d to depth-$(d-1)$ to fail only with *exponentially* small probability as opposed to the bound of $n^{-\Omega_d(1)}$ in the proof of Theorem 4.2. To ensure this, we will use the read-k Chernoff bound stated in Theorem 2.24.

However, this lemma yields strong bounds only when the number of random variables (the parameter m in the statement of Theorem 2.24) is quite large. In particular, for example, if the number of gates at depth $d-1$ is small (a constant, say), then the bound obtained from Theorem 2.24 is not very useful. In this case, instead of simplifying these gates, we simply say that for each setting of the outputs of these gates, we get a circuit of smaller depth. This is why we obtain $(t,d-1,w)$-simple circuits in general and not simply a depth-$(d-1)$ threshold circuit.

Proof. Let $\phi_1,\ldots,\phi_m$ be the threshold gates appearing at depth $d-1$ in the circuit C. We say that ϕ_i is large if fan-in$(\phi_i) \geq n^{\delta_d}$ and small otherwise. Let $L = \{i \mid \phi_i \text{ large}\}$ and $S = [m]\setminus L$. Let $\ell = |L|$. Note that $\ell \leq n^{1+\varepsilon_d-\delta_d} \leq n$. Let $c = n^{\varepsilon_d}$.

As in the inductive case of Theorem 4.2, our construction proceeds in phases.

Phase 1. This is identical to Phase 1 in Theorem 4.2. We thus get a restriction tree of depth $n/2$ such that at *all* leaves of this tree, the resulting circuit is a read-$2c$ circuit with at most cn wires. Let C_1 denote the circuit obtained at some arbitrary leaf of the restriction tree and let n_1 denote the number of variables.

Phase 2. This basic idea here is similar to Phase 2 from Theorem 4.2. However, there are technical differences from Theorem 4.2 since we apply a concentration bound to ensure that the circuit simplifies with very high probability.

We restrict the circuit with a random restriction $\rho = (I,y) \sim \mathcal{R}_p^{n_1}$, where $p = n^{-\delta_d/2}$. As in Theorem 4.2, we have for some $t = 1/p^{\Omega(1)}$, $q = p^{\Omega(1)}$, and for each $i \in L$,

$$\Pr_{\rho}\left[\phi_i|_\rho \text{ is } t\text{-balanced}\right] \leq q, \tag{4.12}$$

$$\Pr_{\rho}\left[\text{fan-in}(\phi_i|_\rho) > 2p\cdot\text{fan-in}(\phi_i)\right] \leq \exp(-\Omega(n^{\delta_d/2})), \qquad \text{and} \tag{4.13}$$

$$\Pr_{\rho=(I,y)}\left[|I| < \frac{n_1 p}{2}\right] \leq \exp(-\Omega(np)). \tag{4.14}$$

Now, we partition L as $L = L_1 \cup \cdots \cup L_a$, where $a \leq 1/\varepsilon_d$, as follows. The set L_j indexes all threshold gates at depth $d-1$ of fan-in at least $n^{\delta_d+(j-1)\varepsilon_d}$ and less than $n^{\delta_d+j\varepsilon_d}$. We let ℓ_j denote $|L_j|$. For each $i \in L$, let Y_i be a random variable that is 1 if $\phi_i|_\rho$ is t-balanced and 0 otherwise. Note that this defines a collection of read-$2c$ Boolean random variables (the underlying independent random variables are $\rho(k)$ for each $k \in [n_1]$).

Let $Z_j = \sum_{i\in L_j} Y_i$, the number of t-balanced gates in L_j. We have

$$\mathbf{E}[Z_j] = \sum_{i\in L_j}\mathbf{E}[Y_i] \leq q\ell_j$$

by (4.12). Thus, by an application of the read-$2c$ Chernoff bound in Theorem 2.24, we have

$$\Pr[Z_j \geq 2q\ell_j] \leq \exp\{-\Omega(q\ell_j/c)\}.$$

Assuming that $\ell_j \geq n^{3\delta_d/4}$ and $B = \delta_d/\varepsilon_d$ is a large enough constant, the right hand side of the above inequality is upper bounded bounded by $\exp\{-2n^{\varepsilon_d}\}$. On the other hand if $\ell_j < n^{3\delta_d/4}$, then $Z_j < n^{3\delta_d/4}$ with probability 1. Hence, we have

$$\Pr_{\rho=(I,y)}\left[Z_j \geq \max\{2q\ell_j, n^{3\delta_d/4}\}\right] \leq \exp\{-2n^{\varepsilon_d}\}$$

and by a union bound

$$\Pr_{\rho=(I,y)}\left[\exists j \in [a], Z_j \geq \max\{2q\ell_j, n^{3\delta_d/4}\}\right] \leq a\exp\{-2n^{\varepsilon_d}\}. \tag{4.15}$$

We call a set I *generic* if $|I| \geq n_1 p/2$ and $\text{fan-in}(\phi_i|_\rho) \leq 2p \cdot \text{fan-in}(\phi_i)$ for each $i \in L$. Let $\mathcal{G}$ denote the event that I is generic. By (4.13) and (4.14), we know that

$$\Pr_I[\neg\mathcal{G}] \leq \ell\exp(-\Omega(n^{\delta_d/2})) + \exp(-\Omega(np)) \leq \exp(-n^{\delta_d/4}).$$

In particular, similar to Theorem 4.2, we get,

$$\Pr_{\rho=(I,y)}\left[\exists j \in [a], Z_j \geq \max\{2q\ell_j, n^{3\delta_d/4}\} \;\middle|\; \mathcal{G}\right] \leq a\exp\{-2n^{\varepsilon_d}\} + \exp(-n^{\delta_d/4}) \leq 2a\exp\{-2n^{\varepsilon_d}\}. \tag{4.16}$$

We fix any generic I such that

$$\Pr_y\left[\exists j \in [a], Z_j \geq \max\{2q\ell_j, n^{3\delta_d/4}\}\right] \leq 2a\exp\{-2n^{\varepsilon_d}\}. \tag{4.17}$$

Consider the restriction tree T that sets all the variables not in I. The tree leaves $n_2 \geq pn_1/2 = pn/4$ variables unfixed. We call a leaf λ of the tree *good* if for each $j \in [a]$ we have $Z_j < \max\{2q\ell_j, n^{3\delta_d/4}\}$ and *bad* otherwise. We have

$$\Pr_{\lambda\sim T}[\lambda \text{ a bad leaf}] \leq 2a\exp(-2n^{\varepsilon_d}). \tag{4.18}$$

For good leaves λ, we show how to approximate $C_\lambda := C_1|_{\rho_\lambda}$ as claimed in the lemma statement.

For the remainder of the argument, fix any good leaf λ. We partition $[a] = J_1 \cup J_2$ where

$$J_1 = \{j \in [a] \mid Z_j < 2q\ell_j\}.$$

Note that for any $j \in J_1$, we have

$$\begin{aligned}
\sum_{i\in L_j} Y_i \cdot \text{fan-in}(\phi_i|_{\rho_\lambda}) &\leq \sum_{i\in L_j} Y_i \cdot 2p \cdot \text{fan-in}(\phi_i) \\
&\leq 2p \cdot n^{\delta_d + j\cdot\varepsilon_d} \cdot Z_j \leq n^{\delta_d + j\cdot\varepsilon_d} \cdot 4pq\ell_j \\
&= 4pqn^{\varepsilon_d} \cdot \ell_j \cdot n^{\delta_d + (j-1)\cdot\varepsilon_d} \leq 4pqn^{\varepsilon_d} n^{1+\varepsilon_d} \\
&= 4pqn^{1+2\varepsilon_d}
\end{aligned}$$

where for the last inequality, we have used the fact that since we have ℓ_j gates of fan-in at least $n^{\delta_d+(j-1)\varepsilon_d}$ each, we must have $\ell_j \cdot n^{\delta_d+(j-1)\varepsilon_d} \leq n^{1+\varepsilon_d}$, the total wire complexity of the circuit.

In particular, we can bound the total fan-in of all the t-balanced gates indexed by $\bigcup_{j\in J_1} L_j$ by

$$\sum_{j\in J_1}\sum_{i\in L_j} Y_i \cdot \text{fan-in}(\phi_i|_{\rho_\lambda}) \leq \frac{4pqn^{1+2\varepsilon_d}}{\varepsilon_d}. \tag{4.19}$$

Phase 3. We proceed in two steps as in Theorem 4.2. Since the steps are very similar, we just sketch the arguments. In the first step, we replace all large t-imbalanced gates by their most probable values. This yields a circuit C'_λ of at most the wire complexity of C_λ and such that

$$\Pr_x\left[C_\lambda(x) \neq C'_\lambda(x)\right] \leq \ell \exp(-n^{\Omega(\delta_d)}) \leq n\exp(-n^{\Omega(\delta_d)}) \leq \exp(-n^{\varepsilon_d}). \tag{4.20}$$

In the second step, we construct another restriction tree rooted at λ that simplifies the circuit to the required form. We first restrict by setting all variables that feed into the t-balanced gates that are indexed by $\bigcup_{j\in J_1} L_j$. By (4.19), the number of variables set is bounded by

$$\frac{4pqn^{1+2\varepsilon_d}}{\varepsilon_d} \leq \frac{4pn\cdot n^{\varepsilon_d-\Omega(\delta_d)}}{\varepsilon_d} \leq \frac{pn}{8} \leq \frac{n_2}{2}$$

for a large enough choice of the constant B. This sets all the t-balanced gates indexed by $\bigcup_{j\in J_1} L_j$ to constants while leaving $n_3 \geq n_2/2$ variables still alive. Finally, by Proposition 4.8, we may set all but a set of $n_4 = n_3/2cn^{\delta_d}$ variables to ensure that with probability 1, all the small gates depend on at most one input variable each. We may replace the small gates by the unique variable they depend on or a constant (if they do not depend on any variable) without increasing the wire complexity of the circuit. Call the circuit thus obtained C''_λ.

At this point, the only threshold gates at depth $d-1$ in the circuit C''_λ are the gates indexed by the t-balanced gates in $\bigcup_{j\in J_2} L_j$. But by the definition of J_2, there can be at most $(1/\varepsilon_d)\cdot n^{3\delta_d/4} \leq n^{\delta_d}$ of them. For every setting of these threshold gates to constants, the circuit becomes a depth-$(d-1)$ circuit of size at most $n^{1+\varepsilon_d}$. Hence, we have a $(n^{\delta_d}, d-1, n^{1+\varepsilon_d})$-simple circuit, as claimed.

Note that the number of variables still surviving is given by $n_4 \geq pn/16cn^{\delta_d} \geq n^{1-2\delta_d}$, for a large enough choice of the parameter B. Hence, the restriction tree constructed satisfies the required depth constraints.

For a random leaf $\nu \sim T$, the probability $\mathcal{E}(\nu)$ does not occur is at most the probability that in Phase 2, the leaf sampled is bad. By (4.18), this is bounded by $2a\exp(-2n^{\varepsilon_d}) \leq \exp(-n^{\varepsilon_d})$ as claimed. □

We now prove the correlation bound for the Generalized Andreev function against threshold circuits with small wire complexity. For the sake of induction, it helps to prove a statement that is stronger in two ways: firstly, we consider any function $F_a = F(a,\cdot)$ where $a \in \{-1,1\}^{4n}$ has high Kolmogorov complexity and the input to F_a is further restricted by an arbitrary restriction ρ that leaves a certain number of variables alive; secondly, we prove a correlation bound against circuits which are the AND of a small threshold circuit with a small number of threshold gates.

Definition 4.12 (Intractability)**.** We say that $f : \{-1,1\}^n \to \{-1,1\}$ is (N,d,t,α)*-intractable* if for any restriction ρ on n variables that leaves $m \geq N$ variables unset, any depth-d threshold circuit C on m variables of wire complexity at most $m^{1+\varepsilon_d}$, and any set S of at most t threshold functions, we have

$$\mathrm{Corr}\Big(f, C \wedge \bigwedge_{g \in S} g\Big) \leq \alpha.$$

The main theorem of this section is the following generalized correlation bound.

Theorem 4.13 (Generalized version of strong correlation)**.** *Fix any constant $d \geq 1$. Let F be the Generalized Andreev function on $5n$ variables as defined in Section 2.4 with γ being any constant smaller than $1/6$. Let $a \in \{-1,1\}^{4n}$ be any string with $K(a) \geq 3n$. Then, the function F_a is $(n^{1-\varepsilon_d}, d, n^{\varepsilon_d}, \exp(-n^{\varepsilon_d}/2))$-intractable.*

The proof is by induction on d. The properties of F_a are only used to prove the base case of the theorem, which can then be used to prove the induction case using Lemma 4.11. We prove the base case separately below (we assume that the constant $B > 0$ is large enough so that this implies the base case of the theorem stated above).

Lemma 4.14 (Base case of induction)**.** *Let $a \in \{-1,1\}^{4n}$ be any string with $K(a) \geq 3n$. Then, the function F_a is $(\sqrt{n}, 1, \sqrt{n}, \exp(-n^{\Omega(1)}))$-intractable.*

Proof. Let $\gamma < 1/6$ in the definition of the Generalized Andreev function in Section 2.4. Let τ be any restriction of n variables leaving $m \geq \sqrt{n}$ variables unfixed. Define $f := F_a|_\tau$. Let C be a conjunction of $\sqrt{n}+1$ threshold gates each on m variables. We wish to prove that

$$\mathrm{Corr}(f, C) \leq \exp(-n^{\Omega(\gamma)}).$$

We build a restriction tree T for C of depth $m - n^\gamma$, by restricting all but n^γ arbitrarily chosen variables. For any leaf ℓ of T, the restricted circuit $C_\ell := C|_{\rho_\ell}$ is a conjunction of $\sqrt{n}+1$ threshold gates each on n^γ variables. By Corollary 2.10, each threshold function can be described using $n^{2\gamma}$ bits. Hence, the entire circuit can be described in a standard way using $(\sqrt{n}+1) \cdot O(n^{2\gamma}) < n$ bits. Then, by Lemma 2.19, we have

$$\mathrm{Corr}\big(f|_{\rho_\ell}, C_\ell\big) \leq \exp(-n^{\Omega(\gamma)}).$$

By Fact 2.3, we then obtain $\mathrm{Corr}(f, C) \leq \exp(n^{-\Omega(\gamma)})$. □

Proof of Theorem 4.13. We only need to prove the inductive case. Assume that $d \geq 2$ is given. Fix any restriction ρ that sets all but $m \geq n^{1-\varepsilon_d}$ variables and let $f = F_a|_\rho$. Let C be a depth-d threshold circuit on the surviving variables of wire complexity at most $m^{1+\varepsilon_d}$. Let S be any set of at most n^{ε_d} threshold functions on the m variables. We need to show that

$$\mathrm{Corr}\Big(f, C \wedge \bigwedge_{g \in S} g\Big) \leq \exp(-n^{\varepsilon_d/2}).$$

Apply Lemma 4.11 to circuit C to find a restriction tree T as guaranteed by the statement of the lemma. By Fact 2.3, we have

$$\begin{aligned}\mathrm{Corr}\Big(f, C\wedge\bigwedge_{g\in S} g\Big) &\le \mathop{\mathbf{E}}_{\ell\sim T}\Big[\mathrm{Corr}\Big(f_\ell, C_\ell\wedge\bigwedge_{g\in S} g_\ell\Big)\Big]\\ &\le \Pr_{\ell}[\neg\mathcal{E}(\ell)] + \max_{\ell:\mathcal{E}(\ell)\text{ holds}} \mathrm{Corr}\Big(f_\ell, C_\ell\wedge\bigwedge_{g\in S} g_\ell\Big) \end{aligned}\tag{4.21}$$

where f_ℓ denotes $f|_{\rho_\ell}$ and similarly for C_ℓ and g_ℓ, and $\mathcal{E}(\ell)$ is the event defined in the statement of Lemma 4.11.

Fix any leaf ℓ so that $\mathcal{E}(\ell)$ holds. We want to bound

$$\mathrm{Corr}\Big(f_\ell, C_\ell\wedge\bigwedge_{g\in S} g_\ell\Big).$$

By definition of $\mathcal{E}(\ell)$, we know that C_ℓ is $\exp(-m^{\varepsilon_d})$-approximated by a $(m^{\delta_d}, d-1, m^{1+\varepsilon_d})$-simple circuit C'_ℓ. This implies that $C_\ell\wedge\bigwedge_{g\in S} g_\ell$ is $\exp(-m^{\varepsilon_d})$-approximated by $C'_\ell\wedge\bigwedge_{g\in S} g_\ell$. Hence, we have

$$\mathrm{Corr}\Big(f_\ell, C_\ell\wedge\bigwedge_{g\in S} g_\ell\Big) \le \mathrm{Corr}\Big(f_\ell, C'_\ell\wedge\bigwedge_{g\in S} g_\ell\Big) + \exp(-m^{\varepsilon_d}).\tag{4.22}$$

Further, by the definition of simplicity and its consequence (4.11), we know that there exist $r\le m^{\delta_d}$ threshold functions $h_1^\ell,\dots,h_r^\ell$ such that

$$C'_\ell = \bigvee_{b_1,\dots,b_r\in\{-1,1\}} C_{b_1,\dots,b_r}\wedge\bigwedge_{i:b_i=-1} h_i^\ell\wedge\bigwedge_{i:b_i=1}\neg h_i^\ell$$

where each $C_{b_1,\dots,b_r}$ is a depth-$(d-1)$ threshold circuit of size at most $m^{1+\varepsilon_d}$ and the OR above is disjoint. This further implies that

$$C'_\ell\wedge\bigwedge_{g\in S} g_\ell = \bigvee_{b_1,\dots,b_r\in\{-1,1\}}\Big(C_{b_1,\dots,b_r}\wedge\bigwedge_{i:b_i=-1} h_i^\ell\wedge\bigwedge_{i:b_i=1}\neg h_i^\ell\wedge\bigwedge_{g\in S} g_\ell\Big)\tag{4.23}$$

and the OR remains disjoint.

Note that we may apply the induction hypothesis to obtain a bound on the correlation with each term in the OR at this point, since the number of surviving variables is at least $m_1 = m^{1-2\delta_d}\ge n^{1-\varepsilon_d-2\delta_d}\ge n^{1-\varepsilon_{d-1}}$ (throughout, we assume that B is a large enough constant for many of the inequalities to hold); and the wire complexity of each depth-$(d-1)$ circuit $C_{b_1,\dots,b_r}$ is at most

$$m^{1+\varepsilon_d}\le m_1^{(1+\varepsilon_d)/(1-2\delta_d)}\le m_1^{1+\varepsilon_d+3\delta_d}\le m_1^{1+\varepsilon_{d-1}}$$

and further, the number of threshold functions in each term is at most $n^{\varepsilon_d}+n^{\delta_d}<m^{\varepsilon_{d-1}}$. Thus, by the inductive hypothesis, we obtain for any $b_1,\dots,b_r$,

$$\mathrm{Corr}\Big(f, C_{b_1,\dots,b_r}\wedge\bigwedge_{i:b_i=-1} h_i^\ell\wedge\bigwedge_{i:b_i=1}\neg h_i^\ell\wedge\bigwedge_{g\in S} g_\ell\Big)\le\exp(-n^{\varepsilon_{d-1}/2}).$$

Using the fact that the OR in (4.23) is disjoint, from Proposition 4.10, we obtain

$$\mathrm{Corr}\Big(f, C'_\ell \wedge \bigwedge_{g\in S} g_\ell\Big) \le O(2^r \cdot \exp(-n^{\varepsilon_{d-1}/2})) \le O\big(2^{n^{\delta_d}} \cdot \exp(-n^{\varepsilon_{d-1}/2})\big) \le \exp(-n^{\varepsilon_d}).$$

Putting the above together with (4.21) and (4.22), we obtain

$$\mathrm{Corr}\Big(f, C \wedge \bigwedge_{g\in S} g\Big) \le \exp(-m^{\varepsilon_d}) + \exp\big(-n^{\varepsilon_d}\big) \le \exp(-n^{\varepsilon_d/2}).$$

which proves the induction case and hence the theorem. □

Corollary 4.3 (Correlation bounds for Andreev's function (restated))**.** Fix any $d \ge 1$ and let F be as in the statement of Theorem 4.13. Any depth-d threshold circuit C of wire complexity at most $n^{1+\varepsilon_d}$ satisfies $\mathrm{Corr}(C,F) \le 2\exp(-n^{\varepsilon_d/2})$.

Proof. For a random $a \in \{-1,1\}^{4n}$, we know by Fact 2.15 that $K(a) \ge 3n$ with probability $1 - \exp(-\Omega(n))$. For each such a, by Theorem 4.13, we have $\mathrm{Corr}(C_a, F_a) \le \exp(-n^{\varepsilon_d/2})$, where C_a is the circuit obtained by substituting $x_1 = a$ in C. Hence, we have

$$\mathrm{Corr}(C,F) \le \mathop{\mathbf{E}}_a[\mathrm{Corr}(C_a,F_a)] \le \exp(-\Omega(n)) + \exp(-n^{\varepsilon_d/2}) \le 2\exp(-n^{\varepsilon_d/2})$$

as claimed. □

5 Proof of Main Structural Lemma (Lemma 4.4)

We prove the following statement that was key in proving the results of Section 4.

Lemma 4.4 (Main Structural lemma for threshold gates (restated))**.** The following holds for some absolute constant $p_0 \in [0,1]$. For any threshold gate ϕ over n variables with label (w,θ) and any $p \in [0,p_0]$, we have[4]

$$\Pr_{\rho\sim\mathcal{R}_p^n}\left[\phi|_\rho \text{ is } \frac{1}{p^{\Omega(1)}}\text{-balanced}\right] \le p^{\Omega(1)}.$$

Proof overview. The proof follows a template that has been used in many results on threshold functions (see, e. g., [50, 38, 11, 20, 12, 32]). The basic idea is to divide threshold gates into one of two kinds: threshold gates such that none of their variables have too much weight (called *regular* threshold functions below) and those where some variable has large weight. For regular gates, the lemma can be easily proved by appealing to standard results on anticoncentration of sums of independent random variables. For gates that are not regular, we first try to make them regular by setting a few variables. If we succeed, then we can appeal to the regular case and we will be done. Otherwise, it must be the case that there are many variables that have large weight in the threshold gate. In this case, we show that the random restriction, with high probability, sets a large number of these variables. Since these variables have large weight, this

[4]Recall from Definition 2.11 that $\phi|_\rho$ is the threshold gate obtained by setting variables outside $\rho^{-1}(*)$ according to ρ.

results in a potentially large shift in the threshold θ of the threshold gate, causing it to become imbalanced with high probability.

We now proceed with the formal details. We need the following definitions and facts that have appeared many times before in the literature on threshold functions (see, e. g., [11]).

Definition 5.1 (Regularity and Critical index)**.** Let $\varepsilon \in [0,1]$ be a real parameter. We say that $w \in \mathbb{R}^n$ is ε-regular if for each $i \in [n]$, $|w_i| \leq \varepsilon \cdot \|w\|_2$.

Assume that the co-ordinates of the vector w are sorted so that $|w_1| \geq |w_2| \geq \cdots \geq |w_n|$. Let $w_{\geq i} \in \mathbb{R}^{n-i+1}$ denote the vector obtained by removing the first $i-1$ co-ordinates of w. We define the ε-critical index of w be the least $K = K(\varepsilon)$ so that the vector $w_{\geq K+1}$ is ε-regular. Note that $K = 0$ if w is already ε-regular and we define $K = n$ if the ε-critical index is not defined.

We say that an n-variable threshold gate ϕ labeled by (w,θ) is ε-regular if w is. Similarly, the ε-critical index of ϕ is defined to be the ε-critical index of w.

Define the parameter $L = L(\varepsilon) = (100\log^2(1/\varepsilon))/\varepsilon^2$.

The Berry–Esseen theorem (see, e. g., [14]) yields the following standard anticoncentration lemma for linear functions. (See [11, Corollary 2.2] for this particular statement.)

Lemma 5.2 (Anticoncentration for regular linear functions)**.** *Let $w \in \mathbb{R}^n$ be ε-regular and let $J \subseteq \mathbb{R}$ be any interval. Then,*

$$\left| \Pr_{x \in \{-1,1\}^n} [\langle w,x \rangle \in J] - \Phi(J) \right| \leq O(\varepsilon)$$

where $\Phi(\cdot)$ denotes the cdf of the standard Gaussian with mean 0 and variance $\|w\|_2^2$. In particular, if $|J|$ denotes the length of J, then

$$\Pr_x [\langle w,x \rangle \in J] \leq \frac{|J|}{\|w\|_2} + O(\varepsilon).$$

We now proceed with the proof of Lemma 4.4. Throughout, we work with random restrictions sampled from $\mathcal{R}_p^n$ where $p \in [0,1]$ is the probability from the statement of Lemma 4.4: equivalently, we pick a pair (I,y) where $I \subseteq [n]$ and $y \in \{-1,1\}^{n-|I|}$.

Let the threshold gate ϕ be labeled by pair (w,θ), where $w \in \mathbb{R}^n$. We may assume that the variables of the threshold gate have been sorted so that $|w_1| \geq |w_2| \geq \cdots \geq |w_n|$. Note that after applying a restriction ρ, the threshold gate $\phi|_\rho$ is labeled by pair (w',θ'), where w' is the restriction of w to the coordinates in I and

$$\theta' = \theta'(\rho) = \theta - \langle w'', y \rangle \tag{5.1}$$

where we use w'' to denote the vector w restricted to the indices in $[n] \setminus I$.

For a random restriction $\rho \sim \mathcal{R}_p^n$, define the following "bad" events:

1. $\mathcal{B}^t(\rho)$ (t a parameter): $\phi|_\rho$ is t-balanced: i. e., $\theta' \leq t \cdot \|w'\|_2$. This is the event whose probability we want to upper bound.

2. $\mathcal{B}_1(\rho)$: $\sum_{i \in I} w_i^2 \geq 2p\|w\|_2^2$.

3. $\mathcal{B}_2^{k,\ell}(\rho)$ (k,ℓ parameters): I contains at least k variables among the first ℓ variables $x_1,\dots,x_\ell$.[5]

We have the following upper bounds on the probabilities of some of the above bad events.

Claim 5.3. *Say w is ε'-regular for any $\varepsilon' \le \frac{1}{\sqrt{16\log(1/p)}}$. Then, $\Pr_\rho[\mathcal{B}_1(\rho)] \le p^2$.*

Proof. The proof follows by applying the variant of the Chernoff bound in Theorem 2.23. Since each variable is added to I independently with probability p, we have

$$\mathop{\mathbf{E}}_{\rho}\left[\sum_{i\in I} w_i^2\right] = p\cdot \|w\|_2^2.$$

Applying Theorem 2.23 to the Boolean random variables X_i which take value 1 iff $x_i \in I$ and for $X = \sum_i w_i^2 X_i$, we obtain the following bound.

$$\Pr\left[\sum_{i\in I} w_i^2 \ge 2\|w\|_2^2\right] \le \Pr\left[X \ge \mathbf{E}[X] + \|w\|_2^2\right] \le \exp(-\|w\|_2^2/8\max_i\{w_i^2\}) \le p^2$$

where for the last inequality we have used the fact that w is ε'-regular and hence

$$\max_i w_i^2 \le (\varepsilon')^2\cdot\|w\|_2^2 = \|w\|_2^2/(16\log(1/p)).$$

□

Claim 5.4. $\Pr_\rho\left[\mathcal{B}_2^{k,\ell}(\rho)\right] \le (ep\ell/k)^k$.

Proof. The probability that any specific set of k variables is a subset of I is p^k. By a union bound, for any ℓ, we have

$$\Pr_\rho\left[\mathcal{B}_2^{k,\ell}(\rho)\right] \le \binom{\ell}{k} p^k \le (e\ell/k)^k\cdot p^k.$$

□

We start with a simpler subcase of the lemma that follows almost directly from Lemma 5.2.

Lemma 5.5 (The regular case)**.** *Say that w is ε'-regular for some $\varepsilon' \le \frac{1}{\sqrt{16\log(1/p)}}$. Then*

$$\Pr_\rho\left[\mathcal{B}^t(\rho)\right] \le O(t\sqrt{p}+\varepsilon').$$

Proof. We bound $\Pr_\rho[\mathcal{B}^t(\rho)]$ as follows.

$$\begin{aligned}\Pr_\rho\left[\mathcal{B}^t(\rho)\right] &\le \Pr_\rho[\mathcal{B}_1(\rho)] + \Pr_\rho\left[\mathcal{B}^t(\rho)\mid \neg(\mathcal{B}_1(\rho))\right]\\ &\le p^2 + \Pr_\rho\left[\mathcal{B}^t(\rho)\mid\neg(\mathcal{B}_1(\rho))\right] \qquad (5.2)\end{aligned}$$

where the bound on $\Pr[\mathcal{B}_1(\rho)]$ follows from Claim 5.3.

Now, note that the event $\neg\mathcal{B}_1(\rho)$ only depends on the choice of $\rho^{-1}(*) = I$. Hence we can condition on an I so that this event occurs; choosing ρ is now equivalent to choosing a random assignment y to the variables in $[n]\setminus I$.

We have $\theta' = \theta - \langle w'', y\rangle$. Using the fact that $\mathcal{B}_1(\rho)$ doesn't occur, we have

[5] We actually only need to analyze the events $\mathcal{B}_2^{1,\ell}$ for the proof of Lemma 4.4. However, we will need a slightly more general bound for the proof of Lemma 9.1 later.

- $\|w''\|_2 \geq \|w\|_2\sqrt{1-2p} \geq \|w\|_2/2$. Using the ε'-regularity of w, for each $i \notin I$, we have

$$|w_i| \leq \varepsilon'\|w\|_2 \leq 2\varepsilon'\|w''\|_2.$$

  Thus, w'' is $2\varepsilon'$-regular.

- $\|w'\|_2 \leq 2p^{1/2}\|w\|_2 \leq 4p^{1/2}\|w''\|_2$,

Using the above, we can see that the probability that

$$\begin{aligned}\Pr_\rho\left[\mathcal{B}^t(\rho)\mid\neg\mathcal{B}_1(\rho)\right] &\leq \Pr_y\left[|\theta'|\leq t\cdot\|w'\|_2\right] \leq \Pr_y\left[|\theta'|\leq 4tp^{1/2}\cdot\|w''\|_2\right]\\ &\leq \Pr_y\left[\langle w'',y\rangle\in[\theta-4tp^{1/2}\cdot\|w''\|_2,\theta+4tp^{1/2}\cdot\|w''\|_2]\right]\\ &\leq O(tp^{1/2})+O(\varepsilon')\end{aligned}$$

where the final inequality uses the anti-concentration bound in Lemma 5.2. Putting the above together with (5.2), we are done. □

Proof of Lemma 4.4. The proof of the lemma is a standard case analysis based on the ε-critical index of the threshold gate ϕ (see [50, 38, 11, 32]). Let $\varepsilon = p^{1/8}$ and $t = p^{-1/16}$. The parameter p_0 will be chosen in the proof below.

The first case is when the critical index $K \leq L$. In this case, we bound the probability of $\mathcal{B}^t(\rho)$ by

$$\begin{aligned}\Pr_\rho\left[\mathcal{B}^t(\rho)\right] &\leq \Pr_\rho\left[\mathcal{B}_2^{1,K}(\rho)\right]+\Pr_\rho\left[\mathcal{B}^t(\rho)\mid\neg\mathcal{B}_2^K(\rho)\right]\\ &\leq epK+\Pr_\rho\left[\mathcal{B}^t(\rho)\mid\neg\mathcal{B}_2^K(\rho)\right]\leq\sqrt{p}+\Pr_\rho\left[\mathcal{B}^t(\rho)\mid\neg\mathcal{B}_2^K(\rho)\right]\end{aligned} \tag{5.3}$$

where the second inequality follows from Claim 5.4 and the final inequality follows from the fact that $epK \leq epL \leq \sqrt{p}$ by our choice of parameters. The event $\neg\mathcal{B}_2(\rho)$ only depends on the choice of the sub-restriction $\rho|_{[K]}$ and we can condition on $\rho|_{[K]}$ so that this event occurs. From now on, the random choice will be a restriction $\rho' \sim \mathcal{R}_p^{n-K}$ on the remaining variables.

Since the restricted linear function is now ε-regular by the definition of the ε-critical index, we can apply Lemma 5.5 to conclude that

$$\Pr_{\rho'}\left[\mathcal{B}^t(\rho)\,\middle|\,\neg\mathcal{B}_2^{1,K}(\rho)\right]\leq p^{\Omega(1)}$$

(note that $\varepsilon = p^{1/8} \leq 1/\sqrt{16\log(1/p)}$ as long as p is smaller than some absolute constant p_0, so that Lemma 5.5 is applicable). Along with (5.1), this implies the lemma in the case that $K \leq L$.

The second case is when $K > L$. As in previous cases, we first condition on some bad event not occurring. We have

$$\begin{aligned}\Pr_\rho\left[\mathcal{B}^t(\rho)\right] &\leq \Pr_\rho\left[\mathcal{B}_2^{1,L}(\rho)\right]+\Pr_\rho\left[\mathcal{B}^t(\rho)\,\middle|\,\neg\mathcal{B}_2^{1,L}(\rho)\right]\\ &\leq epL+\Pr_\rho\left[\mathcal{B}^t(\rho)\,\middle|\,\neg\mathcal{B}_2^{1,L}(\rho)\right]\leq\sqrt{p}+\Pr_\rho\left[\mathcal{B}^t(\rho)\,\middle|\,\neg\mathcal{B}_2^{1,L}(\rho)\right].\end{aligned} \tag{5.4}$$

As above, we can condition on a fixed I so that $\mathcal{B}_2^{1,L}(\rho)$ does not occur (i. e., none of the first L variables belong to I). We then use the following claim that is implicit in [11].

Proposition 5.6. *Assume that $L' = (10r\log(1/\varepsilon))/\varepsilon^2$ and that the ε-critical index $K > L'$. Let y be a random assignment to any set of variables including the first L' variables. Then, the probability over y that the restricted threshold gate is not $(1/\varepsilon)$-imbalanced is at most 2^{-r}.*

Applying the above proposition with $L' = L$ and $r = 10\log(1/\varepsilon)$, we have

$$\Pr_{\rho}\left[\mathcal{B}^t(\rho) \mid \neg\mathcal{B}_2^{1,L}(\rho)\right] \le \varepsilon^{10} < p.$$

Putting this together with (5.4), we have the claimed upper bound on $\Pr_\rho\left[\mathcal{B}^t(\rho)\right]$ in the case that $K > L$. □

For completeness, we give below a proof sketch of Proposition 5.6.

Proof of Proposition 5.6. Let J be the set of variables being set and let $y \in \{-1,1\}^{|J|}$ denote the random assignment chosen. Let $L_0 = (1/\varepsilon^2)\cdot 3\log(1/\varepsilon)$. It can be checked that for any $i < L' - L_0$, we have

$$\|w_{\ge(i+L_0)}\|_2^2 \le \frac{\varepsilon^2}{9}\cdot\|w_{\ge i}\|^2 \le \frac{w_i^2}{9}.$$

Hence, we can choose indices $i_1 = 1, i_2 = 1+L_0, \cdots, i_{r+2} = 1+(r+1)L_0 \le L'$ such that

$$|w_{i_{j+1}}| \le \frac{|w_{i_j}|}{3} \qquad \text{and} \qquad \|w_{\ge i_{j+1}}\|_2^2 \le \frac{\varepsilon^2}{9}\cdot\|w_{\ge i_j}\|_2^2.$$

Further, we have

$$\sum_{i\notin J} w_i^2 \le \|w_{\ge L'}\|^2 \le \|w_{\ge i_{r+2}}\|^2 \le \frac{\varepsilon^2}{9}\cdot\|w_{\ge i_{r+1}}\|_2^2 \le \frac{\varepsilon^2}{81}\cdot w_{i_r}^2.$$

We condition on any setting of variables other than $y_{i_1},\dots,y_{i_r}$. This means that the constant term of the restricted threshold gate θ' is given by

$$\theta' = \theta'' - \sum_{j\in[r]} w_{i_j} y_{i_j}$$

for some $\theta'' \in \mathbb{R}$. The probability that the threshold gate is not $1/\varepsilon$-imbalanced is at most

$$\begin{aligned}\Pr_{y_{i_1},\dots,y_{i_r}}\left[|\theta'| \le \frac{1}{\varepsilon}\cdot\sqrt{\sum_{i\notin J} w_i^2}\right] &\le \Pr_{y_{i_1},\dots,y_{i_r}}\left[|\theta'| \le \frac{1}{9}\cdot|w_{i_r}|\right] \\ &= \Pr_{y_{i_1},\dots,y_{i_r}}\left[\sum_{j=1}^{r} w_{i_j}y_{i_j} \in [\theta'' - \frac{1}{9}\cdot|w_{i_r}|, \theta'' + \frac{1}{9}\cdot|w_{i_r}|]\right].\end{aligned}$$

Now, as a result of the exponentially decreasing nature of the $|w_{i_j}|$, it follows that for any interval of length at most $|w_{i_r}|/2$, there can be at most one choice of $y_{i_1},\dots,y_{i_r}$ such that the $\sum_j w_{i_j}y_{i_j}$ lies in that interval. Thus, we have the given bound.

□

6 Satisfiability algorithms beating brute-force search

In this section, we give satisfiability algorithms beating brute force search for bounded-depth threshold circuits with few wires. Until now, such algorithms were only known for threshold circuits of depth 2. We will assume that each threshold gate on m input bits is given as a pair (w,θ), where $w \in \mathbb{Z}^m$ and $\theta \in \mathbb{Z}$, and θ as well as each component of w has bit complexity $\mathrm{poly}(n)$. Note that this assumption is without loss of generality for a threshold function, and that some assumption on representability of threshold functions is necessary in an algorithmic context.

The satisfiability algorithm relies on an algorithmic version of Lemma 4.11, along with a couple of additional ideas. Essentially, we use the algorithmized version of the lemma to reduce the satisfiability of bounded-depth circuits to satisfiability of ANDs of threshold functions, which we can then solve using a recent result of Williams, stated below.

Theorem 6.1. *[54] There is a deterministic algorithm, which given a bounded-depth circuit C on n variables of size $2^{n^{o(1)}}$ with ANDs, ORs and threshold gates, and with the threshold gates appearing only at the bottom layer, decides if C is satisfiable in time $2^{n-n^{\varepsilon'}}\mathrm{poly}(n)$, where $\varepsilon' > 0$ is a constant that depends only on the depth of the circuit.*

We also need the following fact about threshold gates on n input bits: the set of inputs evaluating to 1 (and dually, the set of inputs evaluating to -1) of a linear threshold gate can be enumerated in time proportional to the number of such inputs, modulo a $\mathrm{poly}(n)$ factor.

Proposition 6.2. *Let (w,θ) represent a threshold function ϕ on m input bits, where $w \in \mathbb{Z}^m$ and $\theta \in \mathbb{Z}$ are integers of bit complexity $\mathrm{poly}(m)$. Let S be the set of inputs on which ϕ evaluates to 1. Then S can be enumerated in time $|S|\mathrm{poly}(m)$.*

Proof. We will show how to construct a decision tree for ϕ in time $|S|\mathrm{poly}(m)$, where S is the set of inputs on which ϕ takes value 1. Given a decision tree of size at most $|S|\mathrm{poly}(m)$, it is easy to enumerate the set of inputs on which ϕ takes value 1 in time $|S|\mathrm{poly}(m)$ by scanning through leaves labeled 1 and outputting all assignments corresponding to any such leaf.

The decision tree is constructed recursively as follows. Check if ϕ restricted according to the current partial assignment is satisfiable (in the sense that there is a total assignment consistent with the partial assignment for which ϕ evaluates to 1). Note that satisfiability of a linear threshold gate with polynomial bit complexity of the weights can be done trivially in polynomial time. If the satisfiability check fails, make the current node a leaf and label it with -1. If it succeeds, check if the current partial assignment is falsifiable. If this check fails, make the current node a leaf and label it with 1. Otherwise, branch on an arbitrary unassigned variable and recurse.

Clearly, this decision tree can be constructed with polynomial work at each node, and hence in time $N\mathrm{poly}(m)$, where N is the number of leaves of the tree. We show that $N \le |S|m$. Indeed, we prove inductively that for any internal node v of the tree of height $h \ge 1$, the number of -1 leaves of the tree rooted at v is at most h times the number of 1 leaves, from which the claim follows as the height of the tree $\le m$.

For the inductive claim, the base case $h = 1$ is clear as any node at height 1 must have one leaf labeled 1 and the other labeled -1. Assume the claim for height h. Consider a node v at height $h+1$. Either one of its children is a leaf, or not. If one of the children is a leaf, then the other one v' is not and by the induction hypothesis, since it is of height h, has at most h times as many -1 leaves as 1 leaves. The number of -1 leaves of v is at most one plus the number of -1 leaves of v', and hence at most $h+1$ times the number of 1 leaves. In case both children of v are internal nodes, then they are both of height at most h, and by the induction hypothesis, both have at most h times as many -1 leaves as 1 leaves, which implies that the same holds for v. □

Definition 6.3. We use THR to refer to the class of linear threshold functions. We use AND $\circ$ THR to refer to the class of polynomial-size circuits with an AND gate at the top and threshold gates at the bottom layer.

Theorem 6.4. *For each integer $d > 0$, there is a constant $\varepsilon_d > 0$ such that satisfiability of a depth-d threshold circuit with at most $n^{1+\varepsilon_d}$ wires on n variables can be solved by a randomized algorithm in time $2^{n-\Omega(n^{\varepsilon_d})}\text{poly}(n)$.*

Proof. As the proof follows the proof of Lemma 4.11 closely, we just give a sketch. Call a circuit depth-d AND $\circ$ THR-skew if the top gate is an AND and all but one child of the top gate is a bottom-level threshold gate, with the possibly exceptional child being a depth-$(d-1)$ threshold circuit with few wires. We follow the depth reduction argument in the lemma to give a recursive algorithm which reduces satisfiability of polynomial-size depth-d AND $\circ$ THR-skew circuits to the satisfiability of polynomial-size depth-$(d-1)$ AND $\circ$ THR-skew circuits by appropriately restricting variables.

For the base case $d = 1$, we simply appeal to the algorithm given by Theorem 6.1, which solves satisfiability of AND $\circ$ THR circuits of polynomial size in time $2^{n-n^{\varepsilon'}}\text{poly}(n)$ for some constant $\varepsilon' > 0$.

For the inductive case, we simulate the proof of Lemma 4.11, which performs and analyzes a certain kind of adaptive random restriction. Various bad events might happen at Phases 2 and 3 of this random restriction process, however each step of the restriction process as well as the check that a bad event happens can be implemented in polynomial time. Moreover, the probability that a bad event happens is at most $2^{-n^{\varepsilon_d}}$. Whenever a bad event happens, we simply do brute force search on the remaining variables of the circuit, but thanks to the exponentially small probability that a bad event happens, with high probability, we only spend time $2^{n-n^{\varepsilon_d}}$ on such brute force searches.

In Phase 3 of the restriction process, we replace imbalanced gates by their most probable values. This changes the functionality of the circuit and might lose us satisfying assignments or give us new invalid satisfying assignments. To get around this, for each such imbalanced gate, we use Proposition 6.2 to efficiently enumerate the inputs evaluating to the minority value for each imbalanced gate, and for each such input check whether it satisfies the original circuit. If it does, we just output "yes." We also append to the top gate of the skew circuit a child representing the assignment of the imbalanced gate to its majority value—this needs to be done so that we don't end up with "false positives" in the base case of the recursive algorithm. Although each such false positive can be tested, there might be too many of them, and this could destroy all the savings we accrue through the course of the algorithm. The total time spent in enumerating minority values of imbalanced gates is again at most $2^{n-n^{\varepsilon_d}}\text{poly}(n)$, with high probability, using the efficient enumeration and the imbalance property.

Finally, there are a few balanced gates—with high probability at most $O(n^{\delta_d})$ of them—for which we need to try all possible values. This could be expensive, but is compensated for by an increased savings for depth $d-1$, just by setting the constant B large enough in the proof of Lemma 4.11. We also need to set B large enough so that the savings given by the application of Williams' algorithm in the base case overwhelms the loss due to branching on balanced threshold gates at depth $d=2$.

Thus the total running time, once B is chosen appropriately, is $2^{n-\Omega(n^{\varepsilon_d})}\mathrm{poly}(n)$, using the fact that $\varepsilon_d < \varepsilon_{d-1} < \ldots < \varepsilon_2$. □

7 Threshold formulas

A threshold formula is a threshold circuit such that the fan-out of each gate is at most 1. A formula can be viewed as a rooted tree. Note that a depth-2 threshold circuit can always be converted to a threshold formula without increasing either the wire complexity or the gate complexity (recall that the gate complexity only measures the number of *non-input* gates).

Theorem 7.1. *There are $2^{O(sn)}$ distinct Boolean functions computed by threshold formulas with s wires on n inputs.*

Proof. A formula with s wires is a rooted tree with $s+1$ nodes. The number of different rooted trees with $s+1$ nodes is $2^{O(s)}$ [39]. For a fixed tree structure, we label the leaves by variables $x_1,\ldots,x_n$, and label the internal nodes by LTFs. Since the number of leaves is at most s, there are $O(n^s)$ different ways of labeling the leaves.

We next label the internal nodes. Without loss of generality, we assume leaves feeding into the same node are labeled by distinct variables, and each internal nodes has fan-in at least two. Consider a topological order of the internal nodes, say, $h_1,\ldots,h_k$, where $k\le s$. For $i=1,\ldots,k$, let s_i be the fan-in of h_i; then $\sum_i s_i = s$. We label each h_i by a LTF on s_i inputs. For fixed labels of h_j for $1\le j<i$, by Theorem 2.9, there are $2^{O(ns_i)}$ possible choices for h_i.

Therefore, the number of different functions computed by threshold formulas of s wires is

$$2^{O(s)}\cdot O(n^s)\cdot 2^{O(ns_1)}\cdots 2^{O(ns_k)} \le 2^{O(ns)}.$$ □

The main result of this section is the following.

Theorem 7.2. *Let $\gamma>0$ be any constant parameter. Any threshold formula on $5n$ variables with at most $n^{2-3\gamma}$ wires has correlation at most $\exp(-n^{\Omega(\gamma)})$ with the Generalized Andreev function as defined in Section 2.4 with parameter γ.*

Proof. Let C be a threshold formula with n inputs and $s=n^{2-3\gamma}$ wires. Let L be the number of leaves in the formula tree; then $L\le s\le 2L$. We build a restriction tree T for C up to depth $n-pn$, for $p=n^\gamma/n$, by greedily restricting the most frequent variables appearing in the leaves. Since the most frequent variable appears at least L/n times, after restricting one variable, the formula tree has at most $L(1-1/n)$ leaves left. Continue this until pn variables left unrestricted; then the number of remaining leaves is at most

$$L\cdot\frac{n-1}{n}\cdot\frac{n-2}{n-1}\cdot\cdots\cdot\frac{pn}{pn+1}=pL.$$

Thus, for any leaf l of T, the restricted formula $C|_{\rho_l}$, on $n' := pn = n^\gamma$ variables, has

$$s' := s(C|_{\rho_l}) \le 2pL \le 2ps$$

wires. By Theorem 7.1, there are at most $2^{O(n's')}$ different threshold formulas of s' wires on n' variables, and thus each such formula has a description of length $O(n's') = O(p^2sn) = O(n^{1-\gamma}) \ll n$.

Recall that $F\colon \{-1,1\}^{4n} \times \{-1,1\}^n \to \{-1,1\}$ is the Generalized Andreev function defined in Section 2.4. Let $a \in \{-1,1\}^{4n}$ be a string with Kolmogorov complexity $K(a) \geq 3n$, and let $F_a(x) := F(a,x)$. Then, by Lemma 2.19,

$$\mathrm{Corr}(F_a|_{\rho_l}, C|_{\rho_l}) \leq \exp(-n^{\Omega(\gamma)}).$$

Note that this holds for every leaf l of T. By Fact 2.3, $\mathrm{Corr}(F_a, C) \leq \exp(-n^{\Omega(\gamma)})$.

Let D be a threshold formula with $5n$ inputs and $n^{2-3\gamma}$ wires, and let $D_a(x) := D(a,x)$. Then D_a is a formula on n inputs with at most $n^{2-3\gamma}$ wires, and thus

$$\Pr_x[F_a(x) = D_a(x)] \leq 1/2 + \exp(-n^{\Omega(\gamma)}).$$

Since a random $a \in \{-1,1\}^{4n}$ has $K(a) \geq 3n$ with probability $1 - 2^{-\Omega(n)}$, the correlation of D and F is at most $2^{-\Omega(n)} + \exp(-n^{\Omega(\gamma)}) = \exp(-n^{\Omega(\gamma)})$. □

8 AC^0 circuits with a few threshold gates

In this section, we extend the noise sensitivity and correlation bounds from Section 3 to the more general setting of small AC^0 circuits (i. e., a constant-depth circuit made up of AND and OR gates) augmented with a small number of threshold gates.

We prove noise sensitivity bounds for Boolean functions computed by such circuits. As consequences of this, we are able to prove correlation bounds against such circuits (as in Section 3) and also devise learning algorithms for such circuits under the uniform distribution (as in [28, 17]).

Following Gopalan and Servedio [17], we define $\mathsf{TAC}^0[k]$ to be the class of constant-depth circuits made up of AND and OR gates and at most k arbitrary threshold gates. The inputs to the circuit are allowed to be arbitrary literals over the underlying variables.

We prove upper bounds on the noise sensitivity of small depth-d $\mathsf{TAC}^0[k]$ circuits for k much smaller than $n^{1/2(d-1)}$. The main theorem of the section is the following.

Theorem 8.1. *Fix any constant $d \geq 1$. Let C be a depth-d $\mathsf{TAC}^0[k]$ circuit with at most M gates overall. Then, for any $p, q \in [0,1]$ and any $D \geq 1$, we have*

$$\mathrm{NS}_{p^{d-1}q}(C) \leq O(k\alpha(p,D) + \alpha(q,D) + M(10pD)^D)$$

where $\alpha(p,D) := \sqrt{p} \cdot (\log(1/p))^{O(D\log D)} \cdot 2^{O(D^2\log D)}$ and $O(\cdot)$ hides an absolute constant (independent of p and D).

This implies a correlation bound for the parity function for such circuits as in Section 3 (see Corollary 8.4 below). Using Theorem 8.1 along with a general idea due to Klivans et al. [28], we also get the following subexponential-time (i. e., $2^{o(n)}$-time) learning algorithms for $\mathsf{TAC}^0[k]$ circuits of small size.

Theorem 8.2. *Let d be any fixed constant. The class of $\mathsf{TAC}^0[k]$ circuits of depth d and size M where*

$$M = n^{o(\sqrt{\log n/\log\log n})} \qquad \text{and} \qquad k = \delta n^{1/2(d-1)}$$

for some $\delta > 0$ can be learned to within error $\varepsilon > 0$ in time $n^{O(m)}$ where

$$m = \max\{k^{2(d-1)}/\varepsilon^{2d}, n^{1/4+o(1)}/\varepsilon^2\}.$$

In particular, if $\delta = n^{-\Omega(1)}$ and $\varepsilon = \Omega(1)$, then the running time of the algorithm is $2^{o(n)}$.

8.1 Noise sensitivity bounds for $\mathsf{TAC}^0[k]$ circuits (Theorem 8.1)

We now prove Theorem 8.1. The basic idea is the same as in Theorem 3.2, but we also need to use the following powerful result of Kane [26, Corollary 3].

Lemma 8.3 (Kane [26]). *Let f be a degree-D Polynomial Threshold function (PTF).*[6] *Then, for any $p > 0$,*

$$\mathrm{NS}_p(f) \leq \sqrt{p} \cdot (\log(1/p))^{O(D\log D)} \cdot 2^{O(D^2\log D)}.$$

Proof of Theorem 8.1. This is a standard switching argument (see, e. g., [21]) augmented with the ideas of Theorem 3.2. We assume throughout that $q \leq 1/2$ without loss of generality since otherwise $\alpha(q,D) \geq q \geq 1/2$ and the claim is trivial.

We say that a threshold gate is a *true* threshold gate if it is not an AND or OR gate.

For any parameters $k_1, d_1, t_1, s_1 \in \mathbb{N}$ with $d_1 \geq 2$, we define $\mathsf{TAC}^0[k_1, d_1, t_1, s_1]$ to be the class of constant-depth circuits made up of AND, OR and threshold gates such that:

- the overall depth is at most d_1,
- the total number of gates at *depth at most* $d_1 - 2$ in the circuit is at most s_1,
- all the true threshold gates are at depth at most $d_1 - 2$ and there are at most k_1 of them, and
- the bottom fan-in of the circuit (i. e., the maximum fan-in of a gate at depth $d_1 - 1$) is at most t_1.

Note that the circuit C in the statement of the theorem is in the class $\mathsf{TAC}^0[k, d+1, 1, M]$, since we may replace the input literals with (say) AND gates of fan-in 1 at the expense of increasing the depth by 1 but in the process satisfying all the criteria of the above definition. We prove the following stronger statement: for any p, q, D as in the statement of the theorem, and any C from the class $\mathsf{TAC}^0[k, d, D, M]$ with $d \geq 2$, we have

$$\mathrm{NS}_{p^{d-2}q}(C) = \mathop{\mathbf{E}}_{\rho_d}\left[\mathrm{Var}(C|_{\rho_d})\right] \leq O(k\alpha(p,D) + \alpha(q,D) + M(10pD)^D) \tag{8.1}$$

where $\rho_d \sim \mathcal{R}^n_{p_d}$ and $p_d := 2p^{d-2}q \in [0,1]$. Proving (8.1) will clearly prove the theorem.

The proof is by induction on d. The base case is $d = 2$. In this case, since there are no true threshold gates at depth $d - 1$ by assumption, a true threshold gate can only occur as the output gate of the circuit

[6] We refer the reader to Section 2.2 for the definitions of Polynomial Threshold functions.

C. Since AND and OR gates are also threshold gates, we can assume that the output gate is a threshold gate. The bottom fan-in being at most D implies that each gate at depth 1 can be represented exactly as a polynomial of degree at most D and therefore that the function computed by C is a degree-D PTF. Hence, Lemma 8.3 trivially implies the result.

Now assume $d > 2$. Let $\psi_1, \ldots, \psi_s$ denote the AND and OR gates at depth exactly $d-2$ in the circuit and let $\phi_1, \ldots, \phi_m$ denote the true threshold gates. By assumption $m \le k$ and $s \le M$. We sample a random restriction $\rho \sim \mathcal{R}_p^n$ and consider the restricted circuit $C|_\rho$.

Håstad's switching lemma [21] tells us that for each $i \in [s]$, we have

$$\Pr_\rho\left[\text{DT-depth}(\psi_i|_\rho) \ge D\right] \le (10pD)^D, \tag{8.2}$$

and hence by a union bound,

$$\Pr_\rho\left[\exists i \in [s] : \text{DT-depth}(\psi_i|_\rho) \ge D\right] \le s(10pD)^D. \tag{8.3}$$

Also, as in the base case, we see that each ϕ_j computes a degree-D PTF. Hence, Lemma 8.3 gives us

$$\mathop{\mathbf{E}}_\rho\left[\sum_{j\in[m]} \mathrm{Var}(\phi_j|_\rho)\right] \le m\alpha(p,D). \tag{8.4}$$

Consider the circuit C'_ρ obtained from $C|_\rho$ as follows: if there is an $i \in [s]$ such that $\text{DT-depth}(\psi_i|_\rho) \ge D$, then C'_ρ is defined to be a trivial circuit that always outputs 1; otherwise, C'_ρ is the depth-$(d-1)$ circuit obtained from $C|_\rho$ as follows:

- We replace each $\phi_j|_\rho$ by a bit $b_{j,\rho} \in \{-1,1\}$ so that by Fact 2.5, we have

$$\Pr_{x\in\{-1,1\}^{|\rho^{-1}(*)|}}\left[\phi_j(x) \ne b_{j,\rho}\right] \le O(\mathrm{Var}(\phi_j)),$$

- Since each $\psi_i|_\rho$ is a depth-D decision tree, we can write it as a D-DNF or D-CNF or as a *disjoint sum of terms* of size at most D each. For each gate χ at depth at most $d-3$ that takes ψ_i as an input, we do the following:
  - If χ is an OR gate, then we take the D-DNF representing $\psi_i|_\rho$ and feed the terms of the DNF directly into χ, eliminating the output OR gate of the D-DNF.
  - If χ is an AND gate, we do the same as above, except that we use the D-CNF representation of $\psi_i|_\rho$ and eliminate the output AND gate.
  - If χ is a threshold gate, then we write $\psi_i|_\rho$ as a disjoint sum of terms of size at most D each and feed each of the terms directly to χ. The gate χ now has many inputs in the place of $\psi_i|_\rho$, and the weight given to each of these inputs is the same as the weight given to $\psi_i|_\rho$.

  Note that the above operations do not increase the number of gates at depth at most $d-3$ in the circuit.

Note that C'_ρ has depth $d-1$ and bottom fan-in at most D. Further, the number of gates at depth at most $d-3$ in C'_ρ is at most $M-s$. Hence, C'_ρ is a circuit from the class $\mathsf{TAC}^0[k-m,d-1,D,M]$. We can thus apply the induction hypothesis and obtain

$$\mathop{\mathbf{E}}_{\rho_{d-1}}\left[\mathrm{Var}(C'_\rho|_{\rho_{d-1}})\right] \le O((k-m)\alpha(p,D)+\alpha(q,D)+(M-s)(10pD)^D). \tag{8.5}$$

To obtain (8.1), we use

$$\begin{aligned}\mathop{\mathbf{E}}_{\rho_d}[\mathrm{Var}(C)] = \mathop{\mathbf{E}}_{\rho_{d-1}}\left[\mathop{\mathbf{E}}_{\rho}\left[\mathrm{Var}(C|_\rho)|_{\rho_{d-1}}\right]\right] &\le \mathop{\mathbf{E}}_{\rho_{d-1}}\left[\mathop{\mathbf{E}}_{\rho}\left[\mathrm{Var}(C'_\rho)|_{\rho_{d-1}}\right]\right] + O\left(\mathop{\mathbf{E}}_{\rho}\left[\delta(C|_\rho, C'_\rho)\right]\right)\\ &= \mathop{\mathbf{E}}_{\rho}\left[\mathop{\mathbf{E}}_{\rho_{d-1}}\left[\mathrm{Var}(C'_\rho)|_{\rho_{d-1}}\right]\right] + O\left(\mathop{\mathbf{E}}_{\rho}\left[\delta(C|_\rho, C'_\rho)\right]\right)\end{aligned} \tag{8.6}$$

where the inequality follows from Proposition 3.3. Inequality (8.5) allows us to bound the first term on the right hand size.

It remains to analyze the last term on the right hand side of (8.6). Define a Boolean random variable $Z = Z(\rho)$ which is 1 iff there is an $i \in [s]$ such that ϕ_i is not a depth-D decision tree. Let $\Delta = \Delta(\rho)$ be the random variable defined by

$$\Delta := Z + \sum_{j\in[m]} \mathrm{Var}(\phi_j|_\rho).$$

It easily follows from the definition of C'_ρ that for any choice of ρ, either $Z = 1$—in which case we can trivially bound $\delta(C'_\rho, C|_\rho)$ by 1—or

$$\delta(C'_\rho, C|_\rho) \le \sum_j \delta(\phi_j|_\rho, b_{j,\rho}) = \sum_j \Pr_x\left[\phi_j|_\rho(x) \ne b_{j,\rho}\right].$$

Hence, for any choice of ρ, we get

$$\delta(C'_\rho, C|_\rho) \le Z + \sum_{j\in[m]} \Pr_{x\in\{-1,1\}^{|\rho^{-1}(*)|}}\left[\phi_j|_\rho(x) \ne b_{j,\rho}\right] \le O(\Delta).$$

Further, by (8.3) and (8.4), we have

$$\mathop{\mathbf{E}}_{\rho}[\Delta] \le O(m\alpha(p,D) + s(10pD)^D).$$

Putting this together with (8.5) and (8.6)[7] gives the claimed bound. This completes the induction. □

This yields the following correlation bound as in Corollary 3.1.

Corollary 8.4. *The following is true for any constant $d \ge 2$. Say C is a depth-d $\mathsf{TAC}^0[k]$ circuit with at most M gates where*

$$k \le \delta \cdot n^{1/2(d-1)} \qquad \text{and} \qquad M = n^{o(\sqrt{\log n/\log\log n})}.$$

Then $\mathrm{Corr}(C, \mathrm{Par}_n) \le n^{o(1)} \cdot \delta^{1-(1/d)}$. *In particular, if $\delta = n^{-\Omega(1)}$, then* $\mathrm{Corr}(C, \mathrm{Par}_n) = n^{-\Omega(1)}$.

[7] Of course, we need to be judicious in our choice of constants in the $O(\cdot)$. We leave this matter to the interested reader.

Proof. We choose a D such that $\omega(1) \leq D \leq o(\sqrt{\log n / \log\log n})$ so that $M \leq n^{o(D)}$ and p, q as in Corollary 3.1. We can then use Theorem 8.1 to obtain

$$\mathrm{NS}_{1/n}(C) \leq n^{o(1)} \cdot \delta^{1-(1/d)} + M \cdot (10pD)^D \leq 1/n^{\omega(1)}.$$

Thus, we get $\mathrm{NS}_{1/n}(C) \leq n^{o(1)} \delta^{1-(1/d)}$. By Proposition 2.6, we have $\mathrm{Corr}(C, \mathrm{Par}_n) \leq O(\mathrm{NS}_{1/n}(C))$, which proves the claim. □

Remark 8.5. The above corollary can be strengthened considerably if a widely believed strengthening of Lemma 8.3—named the Gotsman-Linial conjecture [18]—is known to hold. The Gotsman-Linial conjecture is a conjecture about the *average sensitivity* of low-degree PTFs. We do not recall the exact statement of the conjecture here, and refer the reader to the article by Gopalan and Servedio [17] instead. As noted by [17, Corollary 13], the Gotsman-Linial conjecture implies that for any p and any degree D PTF, we have $\mathrm{NS}_p(f) \leq O(D\sqrt{p})$. Plugging in this bound in place of Lemma 45, it is not hard to see that we can obtain $\mathrm{Corr}(C, \mathrm{Par}_n) = o(1)$ for any circuit C of size $2^{n^{o(1)}}$ from the class $\mathsf{TAC}^0[k]$ where $k = n^{1/2(d-1)-\Omega(1)}$. This is almost a complete generalization of the result of Beigel [6] who proved such a result in the setting where all the threshold gates are of polynomial weight. In contrast, the results of Podolskii [43] and Gopalan and Servedio [17] can prove such correlation bounds only if $k < \log n$.

8.2 Learning algorithms for $\mathsf{TAC}^0[k]$ circuits (Theorem 8.2)

To prove Theorem 8.2, we use Theorem 8.1 along with an observation of Klivans, O'Donnell, and Servedio [28]. We have the following lemma that can be obtained by putting together Fact 9 and Corollary 15 in [28].

Lemma 8.6. *Let $\mathcal{F}$ be a class of Boolean functions defined on $\{-1,1\}^n$. Assume that we know that for some $\varepsilon > 0$ and $f \in \mathcal{F}$, there is a $\gamma > 0$ such that $\mathrm{NS}_\gamma(f) \leq \varepsilon/3$. Then, there is an algorithm that learns $\mathcal{F}$ with error ε in time $n^{O(1/\gamma)}$.*

We now prove Theorem 8.2.

Proof of Theorem 8.2. We can assume that $\varepsilon \geq 1/n^{1/2d}$ since otherwise, we can just run a brute force algorithm that takes time $2^{O(n)}$. We choose a D such that $\omega(1) \leq D \leq o(\sqrt{\log n / \log\log n})$ so that $M = n^{o(D)}$. Theorem 8.1 tells us that for any $p, q \geq 1/n$ any C from the class of circuits described in the theorem statement, we have

$$\mathrm{NS}_{p^{d-1}q}(C) \leq Ak\sqrt{p} + B\sqrt{q} + O(M(10pD)^D)$$

where A and B are $n^{o(1)}$.

We choose p, q so that the first two terms above are each bounded by $\varepsilon/10$. This requires $p \leq \varepsilon^2/O(k^2A^2)$ and $q \leq \varepsilon^2/O(B^2)$. Further, to ensure that the last term is at most $\varepsilon/10$, it suffices to choose $p \leq n^{-\Omega(1)}$ (in fact, this ensures that the third term is $n^{-\omega(1)}$ whereas $\varepsilon \geq n^{-1/2d}$ by assumption). Thus, we fix $p = \min\{\varepsilon^2/O(k^2A^2), n^{-1/4d}\}$ and $q = \varepsilon^2/O(B^2)$ so that all the above conditions are satisfied. This gives

$$\mathrm{NS}_\gamma(C) \leq \varepsilon/3$$

where $\gamma = p^{d-1}q$. Hence, by Lemma 8.6, we obtain the statement of the theorem. □

Remark 8.7. Assuming the Gotsman Linial conjecture, the above technique yields subexponential time constant-error learning algorithms as long as $M \leq 2^{n^{o(1)}}$ and $\delta = n^{-\Omega(1)}$. To contrast again with the work of Gopalan and Servedio [17], the results of [17]—even assuming the Gotsman Linial conjecture—only yield subexponential-time learning algorithms in the setting when $k < \log n$. However, the dependence on the error parameter in [17] is better than the dependence we obtain here. (The running time there has a ε^3 in place of the ε^{2d} that we obtain here.)

9 Depth-3 threshold circuit lower bounds

In this section, we compare our lower bound techniques for threshold functions with the simultaneous independent work of Kane and Williams [27], which proves improved lower bounds for depth-3 threshold circuits. While both results analyze the effects of random restrictions on threshold gates, the random restriction lemma obtained in [27] is different from ours. It thus makes sense to ask if our techniques can also be used to prove the lower bounds of [27]. We show that a modified version of our structural lemma (Lemma 4.4) can be used to recover the lower bound of [27] up to $n^{o(1)}$ factors.

The modified structural lemma is as follows. We say that a threshold gate ϕ with label (w,θ) is (t,k)-*imbalanced* if there is a set S of at most k input variables to ϕ such that for each setting to these variables, the resulting threshold gate is t-imbalanced. We prove the following modification of Lemma 4.4.

Lemma 9.1. *Let ϕ be a threshold gate on n variables with label (w,θ). For a parameter $p \in (0,1)$, and any $t,k \in \mathbb{N}$ such that $t = (1/p)^{o(1)}$ and $k = \omega(1)$ (i. e., $\log_{1/p}(t) \to 0$ and $k \to \infty$ as $p \to 0$), we have*

$$\Pr_{\rho \in \mathcal{R}_p^n}\left[\phi|_\rho \text{ is not } (t,k)\text{-imbalanced}\right] \leq p^{1/2-o(1)}.$$

(Here, upper bounding the probability of an event by $p^{1/2-o(1)}$ means that the probability is at most $\sqrt{p} \cdot g(p)$ where $\lim_{p\to 0} g(p)p^{\delta} = 0$ for any fixed $\delta > 0$.)

Remark 9.2. The statement of Lemma 9.1 is somewhat incomparable to that of Lemma 4.4. On the one hand, Lemma 4.4 yields a stronger property: namely that the threshold gate is $(t,0)$-imbalanced (or equivalently, not t-balanced) with high probability; Lemma 9.1 only yields that the gate is (t,k)-imbalanced with high probability (and also holds only for a smaller value of t). However, the probability estimates in Lemma 9.1 are stronger than those in Lemma 4.4: it can be checked that the proof of Lemma 4.4 cannot yield an upper bound of p^c for some $c < 1/2$.

Replacing the random restriction lemma of [27] with Lemma 9.1 in the lower bound proof of [27] yields the same lower bound for depth-3 circuits up to $n^{o(1)}$ factors. We state the result below after a few basic definitions.

An *unweighted majority* gate is a threshold gate whose label (w,θ) satisfies $|w_i| = 1$ for all i. We let $\mathrm{MAJ} \circ \mathrm{THR} \circ \mathrm{THR}$ denote the class of depth-3 circuits that have an unweighted majority gate as the output gate above two layers of general threshold gates.

Theorem 9.3 (Kane-Williams [27]). *For infinitely many $n \in \mathbb{N}$, there is an explicit (polynomial-time computable) function defined on $2n$ variables that does not have $\mathrm{MAJ} \circ \mathrm{THR} \circ \mathrm{THR}$ circuits with fewer than $n^{1.5-o(1)}$ gates or $n^{2.5-o(1)}$ wires.*

9.1 A modified structural lemma

In this section, we prove Lemma 9.1. We will need the following easy consequence of Theorem 2.20.

Fact 9.4. *Let ϕ be a (t,k)-imbalanced threshold gate such that for some set S of at most k input variables to ϕ and for each setting to these variables, the resulting threshold gate is t-imbalanced. The threshold gate ϕ' obtained from ϕ by keeping only the variables in S (with their weights) and removing the others satisfies*

$$\Pr_x\left[\phi(x) \neq \phi'(x)\right] \leq \exp(-\Omega(t^2)).$$

We now prove Lemma 9.1.

Proof of Lemma 9.1. The proof of the above lemma will follow along the lines of the proof of Lemma 4.4. We also use the definitions and notation from Section 5.

Throughout, we assume that p is a small enough constant since the statement of the lemma only needs to be proved for $p \to 0$. We can also assume that $k = o(\lg(1/p))$ since assuming an upper bound on k only makes the statement stronger. We sample $\rho = (I,y)$ from $\mathcal{R}_p^n$.

We will need a number of parameters in the proof. Let $\varepsilon = m\sqrt{p}$ for an integer parameter $m \leq (1/p)^{o(1)}$ to be chosen later. Define $q = 16m^2 p \lg(1/p)$. Note that $q = p^{1-o(1)}$ since $m \leq (1/p)^{o(1)}$. Let $L = 100\log^2(1/\varepsilon)/\varepsilon^2$. At some places in the proof, we will assume that p is smaller than some constants, since the statement of the lemma is only non-trivial for $p \to 0$.

We assume that the variables of the threshold gate have been sorted so that $|w_1| \geq |w_2| \geq \cdots \geq |w_n|$. After applying a restriction ρ, the threshold gate $\phi|_\rho$ is labeled by pair (w',θ'), where w' is the restriction of w to the coordinates in I and

$$\theta' = \theta'(\rho) = \theta - \langle w'', y\rangle. \tag{9.1}$$

where w'' denotes the vector w restricted to the indices in $[n]\setminus I$.

For a random restriction $\rho \sim \mathcal{R}_p^n$, we define the bad events $\mathcal{B}^t(\rho), \mathcal{B}_1(\rho)$ and $\mathcal{B}_2^{k,\ell}(\rho)$ exactly as in the proof of Lemma 4.4 (here ℓ is a varying parameter we will choose below and k is as fixed above). Additionally, we define the bad event $\tilde{\mathcal{B}}^t(\rho)$ to be the event that ϕ is not (t,k)-imbalanced: note that this is exactly the event whose probability we need to upper bound. Also note that $\mathcal{B}^t(\rho)$ holds whenever $\tilde{\mathcal{B}}^t(\rho)$ does and hence any upper bound on the probability of $\mathcal{B}^t(\rho)$ also applies to $\tilde{\mathcal{B}}^t(\rho)$.

Exactly as in Lemma 4.4, we first the case that w is ε-regular and bound $\Pr_\rho\left[\tilde{\mathcal{B}}^t(\rho)\right]$ in this case. This easily follows since $\Pr\left[\tilde{\mathcal{B}}^t(\rho)\right] \leq \Pr\left[\mathcal{B}^t(\rho)\right]$ and hence by Lemma 5.5 we have

$$\Pr_\rho\left[\tilde{\mathcal{B}}^t(\rho)\right] \leq \Pr_\rho\left[\mathcal{B}^t(\rho)\right] \leq O(t\sqrt{p}+\varepsilon) = p^{1/2-o(1)} \tag{9.2}$$

where we have used the fact that $t = (1/p)^{o(1)}$ and $\varepsilon = m\sqrt{p} = p^{1/2-o(1)}$ since $m = (1/p)^{o(1)}$.

In the non-regular case, we proceed by case analysis based on the ε-critical index K of w. Specifically, we proceed based on whether $K \leq L$ or not.

We first show how to handle the former case, i. e., $K \leq L$. In this case, we bound the probability of $\tilde{\mathcal{B}}^t(\rho)$ as follows

$$\begin{aligned}\Pr_\rho\left[\tilde{\mathcal{B}}^t(\rho)\right] &\leq \Pr_\rho\left[\mathcal{B}_2^{k,K}(\rho)\right] + \Pr_\rho\left[\mathcal{B}^t(\rho) \,\middle|\, \neg\mathcal{B}_2^{k,K}(\rho)\right] \\ &\leq (epK/k)^k + \Pr_\rho\left[\mathcal{B}^t(\rho) \,\middle|\, \neg\mathcal{B}_2^{k,K}(\rho)\right] \leq \sqrt{p} + \Pr_\rho\left[\mathcal{B}^t(\rho) \,\middle|\, \neg\mathcal{B}_2^{k,K}(\rho)\right] \qquad (9.3)\end{aligned}$$

where the first inequality follows from Claim 5.4 and we ensure the final inequality by choosing m so that $(epK/k)^k \leq (epL/k)^k \leq \sqrt{p}$. This can be done by choosing $m = O(\log(1/p) \cdot (1/p)^{1/2k})$ since we have

$$\left(\frac{epL}{k}\right)^k \leq (3pL)^k = \left(\frac{300p\log^2(1/\varepsilon)}{\varepsilon^2}\right)^k \leq \left(\frac{300p\log^2(1/p)}{m^2 p}\right)^k \leq \left(\frac{300\log^2(1/p)}{m^2}\right)^k \leq \sqrt{p}$$

for m as chosen above. Note that $m = (1/p)^{o(1)}$ since we have $k = \omega(1)$ by assumption.

The event $\neg\mathcal{B}_2^{k,K}(\rho)$ only depends on the choice of the sub-restriction $\rho|_{[K]}$ and we can condition on $\rho|_{[K]}$ so that this event occurs. From now on, the random choice will be a restriction $\rho' \sim \mathcal{R}_p^{n-K}$ on the remaining variables.

Let S denote the set of variables in I among the first K variables. We show that with high probability over the choice of ρ', we obtain a t-imbalanced threshold gate for each setting of the variables in S.

There are 2^k settings to the variables in S. For each such setting, the restricted linear function is now ε-regular by the definition of the ε-critical index. Hence we can appeal to the regular case (i. e., inequality (9.1)) and use a union bound over the settings to the variables in S to conclude that

$$\Pr_{\rho'}\left[\mathcal{B}^t(\rho) \,\middle|\, \neg\mathcal{B}_2^{k,K}(\rho)\right] \leq 2^k p^{1/2-o(1)} \leq p^{1/2-o(1)}$$

where we have used our assumption that $k = o(\log(1/p))$ for the final inequality. Along with (9.3), this implies the lemma in the case that $K \leq L$.

In the case that $K > L$, we use a slightly different strategy.

$$\begin{aligned}\Pr_\rho\left[\tilde{\mathcal{B}}^t(\rho)\right] &\leq \Pr_\rho\left[\mathcal{B}_2^{k,L}(\rho)\right] + \Pr_\rho\left[\tilde{\mathcal{B}}^t(\rho) \,\middle|\, \neg\mathcal{B}_2^{k,L}(\rho)\right] \\ &\leq (epL/k)^k + \Pr_\rho\left[\tilde{\mathcal{B}}^t(\rho) \,\middle|\, \neg\mathcal{B}_2^{k,L}(\rho)\right] \leq \sqrt{p} + \Pr_\rho\left[\tilde{\mathcal{B}}^t(\rho) \,\middle|\, \neg\mathcal{B}_2^{k,L}(\rho)\right]. \qquad (9.4)\end{aligned}$$

As we did above, we condition on a fixed I so that $\neg\mathcal{B}_2^{k,L}(\rho)$ occurs (i. e., at most k among the first L variables belong to I). Let S denote the set of variables in I among the first L variables. Note that $|S| \leq k$.

Now, note that if $\tilde{\mathcal{B}}^t(\rho)$ does occur, then on a uniformly random setting to the variables in S, the probability that the restricted threshold gate is not t-imbalanced is at least $1/2^k$.

On the other hand, we analyze what happens when we set all variables not in I and *also the variables in S* uniformly at random. Note that Proposition 5.6 is now applicable with $L' = L$ and $r = 10\log(1/\varepsilon)$ since we are setting all of the first L variables (and also others). Applying Proposition 5.6, the probability that the restricted threshold gate is not t-imbalanced is at most $\varepsilon^{10} < p$. Along with the previous paragraph, this implies that

$$\Pr\left[\tilde{\mathcal{B}}^t(\rho) \,\middle|\, \neg\mathcal{B}_2^{k,L}(\rho)\right] \leq p \cdot 2^k \leq \sqrt{p}$$

since $k = o(\lg(1/p))$.

Putting this together with (9.4), we have the claimed upper bound on $\tilde{\mathcal{B}}^t(\rho)$ in the case that $K > L$. $\square$

9.2 Superlinear gate lower bounds and superquadratic wire lower bounds for threshold circuits

We now prove Theorem 9.3. The proof closely follows [27] except that we use the structural lemma proved above in lieu of the random restriction lemma from [27].

The following proposition (a weaker version of which is implicit in [27]) will help us prove lower bounds on circuits with an unweighted majority gate on top.

Proposition 9.5. *Let h be a Boolean function computed by an unweighted majority gate of fan-in $s \geq 1$. Let $f, g_1, \ldots, g_s$ be Boolean functions of n variables such that* $\max_{i\in[s]} \mathrm{Corr}(f, g_i) \leq (1/16s)$ *and* $\mathrm{Corr}(f, 1) \leq (1/16s)$ *where* 1 *denotes the constant function that takes value* 1 *at all inputs. Then,* $\mathrm{Corr}(f, h(g_1, \ldots, g_s)) \leq 1 - (1/8s)$.

Proof. We can write $h(x_1, \ldots, x_s) = \mathrm{sgn}(\sum_{i\leq s} w_i x_i - \theta)$ where $|w_i| = 1$ for each i. We can also assume that $|\theta| \leq s+1$ since otherwise we may replace θ by $\theta' := \mathrm{sgn}(\theta)\cdot(s+1)$ without changing the function. Similarly, by changing θ if necessary, we can assume that θ is an integer multiple of $1/2$. Let

$$L'(y) := \sum_{i\leq s} w_i g_i(y) - \theta$$

and $h'(y) = h(g_1(y), \ldots, g_s(y))$.

From the fact that $\mathrm{Corr}(f, g_i) \leq 1/16s$ for all $i \in [s]$ and $\mathrm{Corr}(f, 1) \leq 1/16s$, we obtain

$$|\langle f, L'\rangle| \leq \sum_i |w_i|\cdot|\langle f, g_i\rangle| + |\theta|\cdot|\langle f, 1\rangle| \leq s\cdot\frac{1}{16s} + \frac{(s+1)}{16s} < \frac{1}{4}. \tag{9.5}$$

Assume for the sake of contradiction that $\mathrm{Corr}(f, h') > 1 - (1/8s)$. Then, we have

$$\max\left\{ \Pr_{y\in\{-1,1\}^n}\left[f(y) = h'(y)\right], \Pr_{y\in\{-1,1\}^n}\left[f(y) \neq h'(y)\right] \right\} \geq 1 - (1/16s).$$

We assume that $\Pr_{y\in\{-1,1\}^n}[f(y) = h'(y)] \geq 1 - (1/16s)$ and obtain a contradiction to (9.5) (the other case is similar).

It is easy to check using our assumptions on the w_i and θ that whenever $f(y) = h'(y)$, then $f(y)L'(y) \geq 1/2$. Also, for any y such that $f(y) \neq h'(y)$, we have $f(y)L'(y) \geq -2s-1$. Since

$$\Pr_y\left[f(y) = h'(y)\right] \geq 1 - (1/16s),$$

we get

$$\begin{aligned} |\langle f, L'\rangle| \geq \langle f, L'\rangle &\geq \Pr_y\left[f(y) = L'(y)\right]\cdot\frac{1}{2} - \Pr_y\left[f(y) \neq L'(y)\right]\cdot(2s+1) \\ &\geq \left(1 - \frac{1}{16s}\right)\cdot\frac{1}{2} - \frac{(2s+1)}{16s} > \frac{1}{4}. \end{aligned}$$

This contradicts (9.5). $\square$

Recall the statement of Theorem 9.3.

Theorem 9.3 (Restated)**.** For infinitely many $n \in \mathbb{N}$, there is an explicit (polynomial-time computable) function defined on $2n$ variables that does not have $\mathrm{MAJ} \circ \mathrm{THR} \circ \mathrm{THR}$ circuits with fewer than $n^{1.5-o(1)}$ gates or $n^{2.5-o(1)}$ wires.

The explicit function we will use is the same as that defined by Kane and Williams [27, Section 6], which is also a generalization of the Andreev function. The Generalized Andreev function from Section 2.4 can also be suitably modified (by changing the parameters) to yield the same lower bounds.

To prove the above theorem, we need some definitions and some lemmas which are closely related to statements proved in [27].

For positive integers m,k,ℓ, let $\mathrm{THR} \circ \mathrm{THR}[m,k,\ell]$ denote the class of $\mathrm{THR} \circ \mathrm{THR}$ circuits on m variables with at most ℓ gates on the bottom layer that have fan-in greater than k. Similarly, let $\mathrm{MAJ} \circ \mathrm{THR} \circ \mathrm{THR}[m,k,\ell,s]$ denote the class of $\mathrm{THR} \circ \mathrm{THR}$ circuits on m variables with at most ℓ gates on the bottom layer that have fan-in greater than k and at most s gates on the second layer.

Lemma 9.6. *Let m,k,ℓ be arbitrary integers. The number of distinct Boolean functions in* $\mathrm{THR} \circ \mathrm{THR}[m,k,\ell]$ *is bounded by* 2^r*, where* $r = O(2^{O(k^2)} \cdot m^{k+1} + m^2\ell)$.

Proof. We show that each circuit from the class $\mathrm{THR} \circ \mathrm{THR}[m,k,\ell]$ can be described using r bits, where r is as above. This will prove the claim.

To see this, note that by Corollary 2.10 the number of distinct threshold functions of m bits is at most $2^{O(m^2)}$ and hence, all threshold gates of fan-in greater than k can be described using at most $O(m^2\ell)$ bits. The gates of fan-in at most k can compute at most

$$\binom{m}{k} \cdot 2^{O(k^2)} \leq m^k 2^{O(k^2)}$$

distinct functions; we assume that all these functions appear in the bottom layer.

The threshold gate on top thus computes a threshold function of at most $\ell + m^k 2^{O(k^2)}$ threshold functions at the bottom layer. By Theorem 2.9, each such function can be described using

$$O\left(\left(\ell + m^k 2^{O(k^2)}\right) \cdot m\right) \leq O\left(\ell m + 2^{O(k^2)} \cdot m^{k+1}\right)$$

many bits.

Thus, the total description length is $O(\ell m^2 + 2^{O(k^2)} \cdot m^{k+1})$ as required. □

We need the following lemma, implicit in [27].

Lemma 9.7 (Implicit in Corollary 6.2 from [27])**.** *Fix any parameters $M,n \in \mathbb{N}$ such that M is a power of 2 and $M \leq 2^{n/2}$. There is an explicit polynomial-time computable function $F : \{-1,1\}^{\log M} \times \{-1,1\}^n \to \{-1,1\}$ such that given any class of Boolean functions $\mathcal{F}$ on $\log M$ Boolean variables with $|\mathcal{F}| \leq 2^{o(n)}$, there is an $x \in \{-1,1\}^n$ such that for any $f \in \mathcal{F}$, we have* $\mathrm{Corr}(f, F(\cdot, x)) \leq O(\sqrt{n})/M^{1/4}$.

Using the function F from the above lemma, we define our hard function, as in [27], in the following way. Set $M = n^{16}$ in the above lemma. In what follows, we assume that n is a power of 2 and that $\log M \mid n$. The function $B_n : \{-1,1\}^n \times \{-1,1\}^n \to \{-1,1\}$ is defined to be $B_n(x,y) = F(z,x)$ where $z \in \{-1,1\}^{\log M}$ is obtained by partitioning the y variables into $\log M$ blocks of size $n/\log M$ and taking the parity of the bits in the ith block to obtain z_i ($i \in [\log M]$).

Lemmas 9.6 and 9.7 imply a strong correlation bound for B_n against certain kinds of circuits. We will say that a restriction on the y variables of B_n is *live* if it leaves at least one variable unset in each of the $\log M$ blocks defined above.

Lemma 9.8. *Let $n \in \mathbb{N}$ and $M = n^{16}$ be as above. Fix any $m,k,\ell \in \mathbb{N}$ such that the number of functions in $\mathrm{THR} \circ \mathrm{THR}[\log M, k, \ell]$ is $2^{o(n)}$. Then there is an $x \in \{-1,1\}^n$ such that for any live restriction ρ on the y variables that leaves m variables unset, the restricted function $B_{n,\rho} := B(x,\cdot)|_\rho$ satisfies $\mathrm{Corr}(C, B_{n,\rho}) \le 1 - \Omega\left(1/n^3\right)$ for each circuit C from the class $\mathrm{MAJ} \circ \mathrm{THR} \circ \mathrm{THR}[m,k,\ell,n^3]$.*

Proof. By Lemma 9.7, we can fix an $x \in \{-1,1\}^n$ such that the function $F(\cdot, x)$ has correlation at most $O(1/n^{3.5})$ with any function from the class $\mathrm{THR} \circ \mathrm{THR}[\log M, k, \ell]$. Proposition 9.5 then implies that any circuit from the class $\mathrm{MAJ} \circ \mathrm{THR} \circ \mathrm{THR}[m,k,\ell,n^3]$ has correlation at most $1 - \Omega\left(1/n^3\right)$ with F.

For each $i \in [\log M]$, fix a variable $y^{(i)}$ in the ith block of y variables that is left unset by ρ. Let Y denote the set of $\log M$ variables chosen in this way. Construct a restriction tree T that sets all the variables not in Y that are not set by ρ. For each leaf λ of T, we let C_λ denote the circuit $C|_{\rho_\lambda}$ and similarly $B_{n,\lambda}$ for $B_{n,\rho}|_{\rho_\lambda}$. By Fact 2.3, we have

$$\mathrm{Corr}(C, B_{n,\rho}) \le \mathop{\mathbf{E}}_{\lambda \sim T}\left[\mathrm{Corr}(C_\lambda, B_{n,\lambda}\right]$$

and in particular there is a λ such that $\mathrm{Corr}(C, B_{n,\rho}) \le \mathrm{Corr}(C_\lambda, B_{n,\lambda})$. Fix any such λ for the rest of the proof.

Note that C_λ is a circuit from the class $\mathrm{MAJ} \circ \mathrm{THR} \circ \mathrm{THR}[m,k,\ell,n^3]$. Also note that $B_{n,\lambda}$ is exactly a copy of the function F, with some subset of the variables of F being replaced by their negations. By negating some of the variables in C_λ if necessary, we can obtain a circuit C'_λ from the class $\mathrm{MAJ} \circ \mathrm{THR} \circ \mathrm{THR}[m,k,\ell,n^3]$ such that

$$\mathrm{Corr}(C_\lambda, B_{n,\lambda}) = \mathrm{Corr}(C'_\lambda, F) \le 1 - \Omega\left(\frac{1}{n^3}\right)$$

where the latter inequality follows from our reasoning above. By our choice of λ, we obtain

$$\mathrm{Corr}(C, B_{n,\rho}) \le \mathrm{Corr}(C_\lambda, B_{n,\lambda}) \le 1 - \Omega(1/n^3),$$

which concludes the proof of the lemma. □

Proof of Theorem 9.3. The proof follows closely the ideas of [27, Theorem 1.3]. The main new ingredient is the use of Lemma 9.1, which replaces the use of the random restriction lemmas from [27].

Fix $m,k \in \mathbb{N}$ such that $\omega((\log n)^2) \le m \le n^{o(1)}$, $2^{k^2} \le n^{o(1)}$ and $m^k = n^{\omega(1)}$. (For example, we could set $m = 2^{\lceil (\log n)^{5/6} \rceil}$ and $k = \lceil (\log n)^{1/4} \rceil$.) Let $\ell = n/m^3$.

By our choice of parameters and Lemma 9.6, the number of functions in $\mathrm{THR} \circ \mathrm{THR}[\log M, k, \ell]$ is $2^{o(n)}$. By Lemma 9.8, we can find an x so that the restricted function $B_n(x,\cdot)$ is as guaranteed by Lemma 9.8. We fix any such x for the rest of the proof and consider the restricted function $B_n(x,\cdot)$.

We begin with the gates case. Let C be a $\mathrm{MAJ} \circ \mathrm{THR} \circ \mathrm{THR}$ circuit computing $B_n(x,\cdot)$ with at most $s = n^{1.5}/m^{10}r$ gates where $r \leq n^{o(1)}$ is a parameter that will be chosen below.

We will apply a random restriction $\rho = (I, w) \sim \mathcal{R}_p^n$ where $p = m/n$. We bound the probability of the following bad events:

1. Event $\mathcal{E}_1(\rho)$: This is the event that the restriction ρ is not live. For this event to occur, there is some $i \in [\log M]$ such that each of the bits of y that z_i depends on must be fixed to a constant by ρ. For each i, the probability that this happens is at most $(1 - m/n)^{n/\log M} \leq 1/2^{\Omega(m/\log M)} \leq 1/n^{\omega(1)}$. By a union bound over $i \in [\log M]$, the probability of $\mathcal{E}_1$ is at most $1/n^{\omega(1)}$.

2. Event $\mathcal{E}_2(\rho)$: By Lemma 9.1, the probability that any threshold gate is not (t,k)-imbalanced for $t = \log n$ is at most $p^{1/2-\delta}$ where $\delta \to 0$ as $n \to \infty$. As $p \geq 1/n$, we may bound this probability by $\sqrt{p} \cdot n^{\delta}$. Set $r = n^{\delta}$.

   The expected number of gates that are not (t,k)-imbalanced is at most $s \cdot \sqrt{p} \cdot r \leq n/m^9$. The event $\mathcal{E}_2$ is that the number of gates that are not (t,k)-imbalanced is at least $2n/m^9$. By Markov's inequality, the probability of this event is at most $1/2$.

By a union bound, there exists a restriction ρ such that neither of the events $\mathcal{E}_1$ nor $\mathcal{E}_2$ occurs. Fix such a restriction ρ and consider the circuit $C|_\rho$. The number of gates that are not (t,k)-imbalanced in $C|_\rho$ is bounded by $2n/m^9 \leq \ell$.

By Fact 9.4, any (t,k)-imbalanced threshold gate ϕ in $C|_\rho$ can be approximated to error $\exp(-\Omega(t^2)) = 1/n^{\omega(1)}$ by a threshold gate ϕ' of fan-in at most k. By a union bound, the circuit C'_ρ obtained by replacing each (t,k)-imbalanced ϕ by a ϕ' as defined above satisfies

$$\Pr_{y' \in \{-1,1\}^{|I|}} \left[C|_\rho(y') \neq C'_\rho(y') \right] \leq s \cdot \frac{1}{n^{\omega(1)}} \leq \frac{1}{n^{\omega(1)}}.$$

In particular, since $C|_\rho$ computes the function $B' := B_n(x,\cdot)|_\rho$ we have

$$\Pr_{y' \in \{-1,1\}^{|I|}} \left[B'(y') \neq C'_\rho(y') \right] \leq \frac{1}{n^{\omega(1)}}.$$

and hence $\mathrm{Corr}(B', C'_\rho) \geq 1 - (1/n^{\omega(1)})$ which contradicts Lemma 9.8. This proves the theorem for the case of circuits with at most $n^{1.5-o(1)}$ gates.

We now consider the wires case, which is very similar. Let C now be a $\mathrm{MAJ} \circ \mathrm{THR} \circ \mathrm{THR}$ circuit computing $B_n(x,\cdot)$ with at most $s = n^{2.5}/m^{10}r$ *wires* where r is as defined above. We will say that a gate on the bottom layer is *large* if its fan-in is at least n/m^2 and *small* otherwise. We use L to denote the set of large gates and S to denote the set of small gates. Let $s_L := |L|$. Note that $s_L \leq s/(n/m^2) = n^{1.5}/m^8 r$.

As in the gates case, we apply a random restriction $\rho \sim \mathcal{R}_p^n$ to the circuit C. We now consider the following bad events.

1. Event $\mathcal{E}_1(\rho)$: This event is as defined above. As noted above, the probability of this event is bounded by $1/n^{\omega(1)}$.

2. Event $\mathcal{E}'_2(\rho)$: This is the event that some gate in S has fan-in at least k *after* the restriction. Fix any threshold gate $\phi \in S$. Since ϕ has fan-in at most n/m^2, the probability that it has fan-in at least k after the restriction is bounded by

$$\binom{n/m^2}{k} \cdot \left(\frac{m}{n}\right)^k \leq \left(\frac{enm}{knm^2}\right)^k \leq \frac{1}{m^k} = \frac{1}{n^{\omega(1)}}$$

where the last equality follows from our choice of parameters. Since the circuit C has at most s wires, it also has at most s gates and hence by a union bound over the gates in S, the probability that $\mathcal{E}'_1(\rho)$ occurs is at most $n^{2.5}/n^{\omega(1)} \leq 1/n^{\omega(1)}$.

3. Event $\mathcal{E}'_3(\rho)$: This is the event that the number of gates that are not (t,k)-imbalanced is at most $2n/m^7$. Similar to the gates case, we can argue that the expected number of gates in L that are not (t,k)-imbalanced is at most $s_L\sqrt{p} \cdot r \leq n/m^7$. Thus, by Markov's inequality, we see that the probability that $\mathcal{E}'_3(\rho)$ occurs is at most $1/2$.

By a union bound, we see that there is a ρ such that none of $\mathcal{E}_1(\rho), \mathcal{E}'_2(\rho)$, or $\mathcal{E}'_3(\rho)$ occur. From here on, we can follow the proof of the gates case verbatim to derive a contradiction to Lemma 9.8. This proves the theorem in the wires case.

□

Acknowledgements. A major portion of this work was carried out while visiting the Simons Institute for Theory of Computing at Berkeley. We thank the Institute for their support. The first and second authors conducted the work when the authors were at University of Edinburgh, and were supported by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ ERC Grant Agreement no. 615075. The third author would like to thank the School of Informatics at the University of Edinburgh for hosting him in March-April 2015, when this research project was initiated.

We would also like to thank the anonymous CCC 2016 reviewers for their feedback and suggestions and also for pointing out to us the article by Janson [25]. We also thank anonymous reviewers for *Theory of Computing* and Prahladh Harsha for their detailed feedback on the write-up and corrections.

References

[1] SCOTT AARONSON AND AVI WIGDERSON: Algebrization: A new barrier in complexity theory. *ACM Trans. Comput. Theory*, 1(1):2:1–2:54, 2009. Conference version in STOC'08. [doi:10.1145/1490270.1490272] 2

[2] MIKLÓS AJTAI: Σ^1_1-formulae on finite structures. *Ann. Pure Appl. Logic*, 24(1):1–48, 1983. [doi:10.1016/0168-0072(83)90038-6] 2

[3] JAMES ASPNES, RICHARD BEIGEL, MERRICK L. FURST, AND STEVEN RUDICH: The expressive power of voting polynomials. *Combinatorica*, 14(2):135–148, 1994. Conference version in STOC'91. [doi:10.1007/BF01215346] 4, 6, 17, 18

[4] LÁSZLÓ BABAI, NOAM NISAN, AND MARIO SZEGEDY: Multiparty protocols, pseudorandom generators for Logspace, and time-space trade-offs. *J. Comput. System Sci.*, 45(2):204–232, 1992. Conference version in STOC'89. [doi:10.1016/0022-0000(92)90047-M] 5

[5] THEODORE BAKER, JOHN GILL, AND ROBERT SOLOVAY: Relativizations of the P =? NP question. *SIAM J. Comput.*, 4(4):431–442, 1975. [doi:10.1137/0204037] 2

[6] RICHARD BEIGEL: When do extra majority gates help? Polylog(N) majority gates are equivalent to one. *Comput. Complexity*, 4(4):314–324, 1994. Conference version in STOC'92. [doi:10.1007/BF01263420] 4, 5, 42

[7] RUIWEN CHEN, VALENTINE KABANETS, ANTONINA KOLOKOLOVA, RONEN SHALTIEL, AND DAVID ZUCKERMAN: Mining circuit lower bound proofs for meta-algorithms. *Comput. Complexity*, 24(2):333–392, 2015. Conference version in CCC'14. [doi:10.1007/s00037-015-0100-0] 2, 11, 12

[8] RUIWEN CHEN, RAHUL SANTHANAM, AND SRIKANTH SRINIVASAN: Average-case lower bounds and satisfiability algorithms for small threshold circuits. In *Proc. 31st Conf. on Computational Complexity (CCC'16)*, pp. 1:1–1:35. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2016. [doi:10.4230/LIPIcs.CCC.2016.1] 1, 8

[9] FAN R. K. CHUNG AND LINCOLN LU: Survey: Concentration inequalities and martingale inequalities: A survey. *Internet Mathematics*, 3(1):79–127, 2006. [doi:10.1080/15427951.2006.10129115] 13

[10] THOMAS M. COVER AND JOY A. THOMAS: *Elements of Information Theory*. Wiley Online Library, 2005. [doi:10.1002/047174882X] 13

[11] ILIAS DIAKONIKOLAS, PARIKSHIT GOPALAN, RAGESH JAISWAL, ROCCO A. SERVEDIO, AND EMANUELE VIOLA: Bounded independence fools halfspaces. *SIAM J. Comput.*, 39(8):3441–3462, 2010. Conference version in FOCS'09. [doi:10.1137/100783030] 5, 6, 30, 31, 33

[12] ILIAS DIAKONIKOLAS, PRASAD RAGHAVENDRA, ROCCO A. SERVEDIO, AND LI-YANG TAN: Average sensitivity and noise sensitivity of polynomial threshold functions. *SIAM J. Comput.*, 43(1):231–253, 2014. Conference version in STOC'10. [doi:10.1137/110855223, arXiv:0909.5011] 5, 30

[13] DEVDATT P. DUBHASHI AND ALESSANDRO PANCONESI: *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge Univ. Press, 2009. LINK AT ACM DL. 12

[14] WILLIAM FELLER: *An Introduction to Probability Theory and Its Applications*. John Wiley & Sons, 1968. LINK at Wiley. 31

[15] MERRICK FURST, JAMES B. SAXE, AND MICHAEL SIPSER: Parity, circuits, and the polynomial-time hierarchy. *Mathematical Systems Theory*, 17(1):13–27, 1984. Conference version in FOCS'81. [doi:10.1007/BF01744431] 2

[16] DMITRY GAVINSKY, SHACHAR LOVETT, MICHAEL E. SAKS, AND SRIKANTH SRINIVASAN: A tail bound for read-k families of functions. *Random Struct. Algorithms*, 47(1):99–108, 2015. [doi:10.1002/rsa.20532, arXiv:1205.1478] 7, 13

[17] PARIKSHIT GOPALAN AND ROCCO A. SERVEDIO: Learning and lower bounds for AC^0 with threshold gates. In *Proc. 13th Internat. Workshop on Approximation Algorithms for Combinatorial Optimization Problems (APPROX'10)*, pp. 588–601. Springer, 2010. [doi:10.1007/978-3-642-15369-3_44] 5, 14, 16, 38, 42, 43

[18] CRAIG GOTSMAN AND NATHAN LINIAL: Spectral properties of threshold functions. *Combinatorica*, 14(1):35–50, 1994. [doi:10.1007/BF01305949] 42

[19] KRISTOFFER ARNSFELT HANSEN AND PETER BRO MILTERSEN: Some meet-in-the-middle circuit lower bounds. In *Proc. 29th Internat. Symp. on Math. Foundations of Comp. Sci. (MFCS'04)*, pp. 334–345. Springer, 2004. [doi:10.1007/978-3-540-28629-5_24] 5

[20] PRAHLADH HARSHA, ADAM KLIVANS, AND RAGHU MEKA: Bounding the sensitivity of polynomial threshold functions. *Theory of Computing*, 10(1):1–26, 2014. Conference version in STOC'10. [doi:10.4086/toc.2014.v010a001, arXiv:0909.5175] 5, 30

[21] JOHAN HÅSTAD: Almost optimal lower bounds for small depth circuits. In *Proc. 18th STOC*, pp. 6–20. ACM Press, 1986. [doi:10.1145/12130.12132] 2, 6, 39, 40

[22] RUSSELL IMPAGLIAZZO, WILLIAM MATTHEWS, AND RAMAMOHAN PATURI: A satisfiability algorithm for AC^0. In *Proc. 23rd Ann. ACM-SIAM Symp. on Discrete Algorithms (SODA'12)*, pp. 961–972. ACM Press, 2012. [doi:10.1137/1.9781611973099.77, arXiv:1107.3127] 2

[23] RUSSELL IMPAGLIAZZO, MOHAN PATURI, AND STEFAN SCHNEIDER: A satisfiability algorithm for sparse depth two threshold circuits. In *Proc. 54th FOCS*, pp. 479–488. IEEE Comp. Soc. Press, 2013. [doi:10.1109/FOCS.2013.58, arXiv:1212.4548] 3

[24] RUSSELL IMPAGLIAZZO, RAMAMOHAN PATURI, AND MICHAEL E. SAKS: Size-depth tradeoffs for threshold circuits. *SIAM J. Comput.*, 26(3):693–707, 1997. Conference version in STOC'93. [doi:10.1137/S0097539792282965] 2, 4, 5, 17

[25] SVANTE JANSON: Large deviations for sums of partly dependent random variables. *Random Struct. Algorithms*, 24(3):234–248, 2004. [doi:10.1002/rsa.20008] 13, 50

[26] DANIEL M. KANE: The correct exponent for the Gotsman–Linial conjecture. *Comput. Complexity*, 23(2):151–175, 2014. Conference version in CCC'13. [doi:10.1007/s00037-014-0086-z, arXiv:1210.1283] 5, 6, 39

[27] DANIEL M. KANE AND RYAN WILLIAMS: Super-linear gate and super-quadratic wire lower bounds for depth-two and depth-three threshold circuits. In *Proc. 48th STOC*, pp. 633–643. ACM Press, 2016. [doi:10.1145/2897518.2897636, arXiv:1511.07860] 6, 7, 8, 17, 43, 46, 47, 48

[28] ADAM R. KLIVANS, RYAN O'DONNELL, AND ROCCO A. SERVEDIO: Learning intersections and thresholds of halfspaces. *J. Comput. System Sci.*, 68(4):808–840, 2004. Conference version in FOCS'02. [doi:10.1016/j.jcss.2003.11.002] 5, 14, 38, 42

[29] ILAN KOMARGODSKI AND RAN RAZ: Average-case lower bounds for formula size. In *Proc. 45th STOC*, pp. 171–180. ACM Press, 2013. [doi:10.1145/2488608.2488630] 11

[30] NATHAN LINIAL, YISHAY MANSOUR, AND NOAM NISAN: Constant depth circuits, Fourier transform, and learnability. *J. ACM*, 40(3):607–620, 1993. Conference version in FOCS'89. [doi:10.1145/174130.174138] 3

[31] SHACHAR LOVETT AND SRIKANTH SRINIVASAN: Correlation bounds for poly-size AC^0 circuits with $n^{1-o(1)}$ symmetric gates. In *Proc. 14th Internat. Workshop on Approximation Algorithms for Combinatorial Optimization Problems (APPROX'11)*, pp. 640–651. Springer, 2011. [doi:10.1007/978-3-642-22935-0_54] 5

[32] RAGHU MEKA AND DAVID ZUCKERMAN: Pseudorandom generators for polynomial threshold functions. *SIAM J. Comput.*, 42(3):1275–1301, 2013. Conference version in STOC'10. [doi:10.1137/100811623, arXiv:0910.4122] 5, 6, 30, 33

[33] NOAM NISAN: Pseudorandom bits for constant depth circuits. *Combinatorica*, 11(1):63–70, 1991. Conference version in FOCS'88. [doi:10.1007/BF01375474] 2

[34] NOAM NISAN: The communication complexity of threshold gates. In *Combinatorics: Paul Erdös is Eighty*, pp. 301–315. János Bolyai Mathematical Society, 1993. Available at author's website. 5, 14

[35] NOAM NISAN AND AVI WIGDERSON: Hardness vs. randomness. *J. Comput. System Sci.*, 49(2):149–167, 1994. Conference version in FOCS'88. [doi:10.1016/S0022-0000(05)80043-1] 2

[36] RYAN O'DONNELL: Hardness amplification within NP. *J. Comput. System Sci.*, 69(1):68–94, 2004. Conference version in CCC'02. [doi:10.1016/j.jcss.2004.01.001] 6

[37] RYAN O'DONNELL: *Analysis of Boolean Functions*. Cambridge Univ. Press, 2014. LINK. 9, 10, 11, 17

[38] RYAN O'DONNELL AND ROCCO A. SERVEDIO: The Chow parameters problem. *SIAM J. Comput.*, 40(1):165–199, 2011. Conference version in STOC'08. [doi:10.1137/090756466] 5, 6, 30, 33

[39] RICHARD OTTER: The number of trees. *Ann. of Math.*, 49(3):583–599, 1948. [doi:10.2307/1969046] 37

[40] RAMAMOHAN PATURI, PAVEL PUDLÁK, AND FRANCIS ZANE: Satisfiability coding lemma. *Chicago J. Theoret. Comput. Sci.*, (11), 1999. Conference version in FOCS'97. [doi:10.4086/cjtcs.1999.011] 3

[41] RAMAMOHAN PATURI AND MICHAEL E. SAKS: Approximating threshold circuits by rational functions. *Inform. and Comput.*, 112(2):257–272, 1994. [doi:10.1006/inco.1994.1059] 2, 4, 17

[42] YUVAL PERES: Noise stability of weighted majority, 2004. [arXiv:math/0412377] 5, 11

[43] VLADIMIR V. PODOLSKII: Exponential lower bound for bounded depth circuits with few threshold gates. *Inform. Process. Lett.*, 112(7):267–271, 2012. [doi:10.1016/j.ipl.2011.12.011] 5, 42

[44] ANUP RAO: Extractors for low-weight affine sources. In *Proc. 24th Conf. on Computational Complexity (CCC'09)*, pp. 95–101. IEEE Comp. Soc. Press, 2009. [doi:10.1109/CCC.2009.36] 12

[45] ALEXANDER A. RAZBOROV: Lower bounds on the size of bounded-depth networks over the complete basis with logical addition. *Math. Notes of the Acad. Sci. USSR*, 41(4):333–338, 1987. [doi:10.1007/BF01137685] 2

[46] ALEXANDER A. RAZBOROV AND STEVEN RUDICH: Natural proofs. *J. Comput. System Sci.*, 55(1):24–35, 1997. Conference version in STOC'94. [doi:10.1006/jcss.1997.1494] 2

[47] ALEXANDER A. RAZBOROV AND AVI WIGDERSON: $n^{\Omega(\log n)}$ lower bounds on the size of depth-3 threshold circuits with AND gates at the bottom. *Inform. Process. Lett.*, 45(6):303–307, 1993. [doi:10.1016/0020-0190(93)90041-7] 5

[48] VWANI ROYCHOWDHURY, KAI-YEUNG SIU, AND ALON ORLITSKY: Neural models and spectral methods. In *Theoretical Advances in Neural Computation and Learning*, pp. 3–36. Springer, 1994. [doi:10.1007/978-1-4615-2696-4_1] 11

[49] RAHUL SANTHANAM: Fighting perebor: New and improved algorithms for formula and QBF satisfiability. In *Proc. 51st FOCS*, pp. 183–192. IEEE Comp. Soc. Press, 2010. [doi:10.1109/FOCS.2010.25] 2

[50] ROCCO A. SERVEDIO: Every linear threshold function has a low-weight approximator. *Comput. Complexity*, 16(2):180–209, 2007. Conference version in CCC'06. [doi:10.1007/s00037-007-0228-7] 5, 6, 30, 33

[51] ALEXANDER A. SHERSTOV: Optimal bounds for sign-representing the intersection of two half-spaces by polynomials. *Combinatorica*, 33(1):73–96, 2013. Conference version in STOC'10. [doi:10.1007/s00493-013-2759-7, arXiv:0910.4224] 4

[52] KAI-YEUNG SIU, VWANI P. ROYCHOWDHURY, AND THOMAS KAILATH: Rational approximation techniques for analysis of neural networks. *IEEE Trans. Inform. Theory*, 40(2):455–466, 1994. [doi:10.1109/18.312168] 4, 5, 6, 14, 16

[53] ROMAN SMOLENSKY: Algebraic methods in the theory of lower bounds for boolean circuit complexity. In *Proc. 19th STOC*, pp. 77–82. ACM Press, 1987. [doi:10.1145/28395.28404] 2

[54] RYAN WILLIAMS: New algorithms and lower bounds for circuits with linear threshold gates. In *Proc. 46th STOC*, pp. 194–202. ACM Press, 2014. [doi:10.1145/2591796.2591858, arXiv:1401.2444] 35

[55] RYAN WILLIAMS: Nonuniform ACC circuit lower bounds. *J. ACM*, 61(1):2:1–2:32, 2014. Conference version in CCC'11. [doi:10.1145/2559903] 2

[56] ANDREW YAO: Separating the polynomial-time hierarchy by oracles. In *Proc. 26th FOCS*, pp. 1–10. IEEE Comp. Soc. Press, 1985. [doi:10.1109/SFCS.1985.49] 2

AUTHORS

Ruiwen Chen
rwchenmail@gmail.com

Rahul Santhanam
Department of Computer Science
University of Oxford
Oxford, U.K.
rahul.santhanam@cs.ox.ac.uk

Srikanth Srinivasan
Department of Mathematics
IIT Bombay, Mumbai, India
srikanth@math.iitb.ac.in

ABOUT THE AUTHORS

RUIWEN CHEN graduated with a Ph. D. in Computer Science from Simon Fraser University in 2014, advised by Valentine Kabanets, and then worked as a postdoc with Rahul Santhanam until 2017.

RAHUL SANTHANAM got his Ph. D. in Computer Science from the University of Chicago in 2005, advised by Lance Fortnow and Janos Simon. After stints at Simon Fraser University, the University of Toronto and the University of Edinburgh, he is now a Professor of Computer Science at the University of Oxford. His main research interest is in complexity lower bounds.

SRIKANTH SRINIVASAN got his undergraduate degree from the Indian Institute of Technology Madras. He received his Ph. D. from The Institute of Mathematical Sciences in Chennai, India in 2011, where his advisor was V. Arvind. His research interests include circuit complexity, derandomization, and related areas of mathematics.