Electronic Colloquium on Computational Complexity, Revision 1 of Report No. 151 (2013)

Hardness Amplification and the Approximate Degree of Constant-Depth Circuits

Mark Bun * Justin Thaler †

Abstract

We establish a generic form of hardness amplification for the approximability of constant-depth Boolean circuits by polynomials. Specifically, we show that if a Boolean circuit cannot be pointwise approximated by low-degree polynomials to within constant error in a certain one-sided sense, then an OR of disjoint copies of that circuit cannot be pointwise approximated even with very high error. As our main application, we show that for every sequence of degrees $d(n)$, there is an explicit depth-three circuit $F : \{-1,1\}^n \to \{-1,1\}$ of polynomial-size such that any degree-d polynomial cannot pointwise approximate F to error better than $1-\exp\left(-\tilde{\Omega}(nd^{-3/2})\right)$. As a consequence of our main result, we obtain an $\exp\left(-\tilde{\Omega}(n^{2/5})\right)$ upper bound on the the discrepancy of a function in AC^0, and an $\exp\left(\tilde{\Omega}(n^{2/5})\right)$ lower bound on the threshold weight of AC^0, improving over the previous best results of $\exp\left(-\Omega(n^{1/3})\right)$ and $\exp\left(\Omega(n^{1/3})\right)$ respectively.

Our techniques also yield a new lower bound of $\Omega\left(n^{1/2}/\log^{(d-2)/2}(n)\right)$ on the approximate degree of the AND-OR tree of depth d, which is tight up to polylogarithmic factors for any constant d, as well as new bounds for read-once DNF formulas. In turn, these results imply new lower bounds on the communication and circuit complexity of these classes, and demonstrate strong limitations on existing PAC learning algorithms.

1 Introduction

The ε-approximate degree of a Boolean function $f : \{-1,1\}^n \to \{-1,1\}$, denoted $\widetilde{\deg}_\varepsilon(f)$, is the minimum degree of a real polynomial that approximates f to error ε in the ℓ_∞ norm. Approximate degree has pervasive applications in theoretical computer science. For example, lower bounds on approximate degree underly many tight lower bounds on quantum query complexity (e.g., [2,4,6,22, 44]), and have been used to resolve long-standing open questions in communication complexity (see for example the survey paper by Sherstov [39]). Meanwhile, upper bounds on approximate degree underly many of the best known agnostic learning and PAC learning algorithms (e.g. [19,23,24,37]).

Despite the range and importance of these applications, large gaps remain in our understanding of approximate degree. The approximate degree of any *symmetric* Boolean function has been understood since Paturi's 1992 paper [35], but once we move beyond symmetric functions, few general results are known.

*Harvard University, School of Engineering and Applied Sciences. Supported by an NDSEG Fellowship and NSF grant CNS-1237235.

†Simons Institute for the Theory of Computing at UC Berkeley. Parts of this work were performed while the author as a graduate student at Harvard University, School of Engineering and Applied Sciences. This work was supported by an NSF Graduate Research Fellowship, NSF grants CNS-1011840 and CCF-0915922, and a Research Fellowship from the Simons Institute for the Theory of Computing.

ISSN 1433-8092

In this paper, we perform a careful study of the approximate degree of constant-depth Boolean circuits. In particular, we establish a generic form of hardness amplification for approximate degree: we show that if a Boolean circuit f cannot be pointwise approximated to within constant error in a certain one-sided sense by low-degree polynomials, then the circuit F obtained by taking an OR of disjoint copies of f cannot be pointwise approximated even with error exponentially close to 1. Notice that if f is computed by a circuit of polynomial size and constant depth, than so is F.

Our proof extends a recent line of work [12, 28, 38, 46] that seeks to prove approximate degree lower bounds by constructing explicit *dual polynomials*, which are dual solutions to a linear program that captures the approximate degree of any function. Specifically, we show that given a dual polynomial demonstrating that f cannot be approximated to within constant error, we can construct a dual polynomial demonstrating that F cannot be approximated even with error exponentially close to 1.

As the main application of our hardness amplification technique, we exhibit an explicit function $F : \{-1,1\}^n \to \{-1,1\}$ computed by a polynomial size circuit of depth three for which any degree-d polynomial cannot pointwise approximate F to error $1 - \exp\left(-\tilde{\Omega}(nd^{-3/2})\right)$. We then use this result to obtain new bounds on two quantities that play central roles in learning theory, communication complexity, and circuit complexity: *discrepancy* and *threshold weight*. Specifically, we prove a a new upper bound of $\exp\left(-\tilde{\Omega}(n^{2/5})\right)$ for the discrepancy of a function in AC^0, and a new lower bound of $\exp\left(\tilde{\Omega}(n^{2/5})\right)$ for the threshold weight of AC^0. Our techniques also yield new lower bounds for read-once DNF formulas and constant-depth AND-OR trees.

In Section 2, we provide a detailed summary of our results and their relationship to prior work, as well as their applications to learning theory, communication complexity, and circuit lower bounds.

2 Summary of Results

2.1 Hardness Amplification

Central to our work is a measure of the complexity of a Boolean function that we call *one-sided approximate degree*. We denote this quantity by $\widetilde{\deg}(f)$. This measure captures the least degree of a real polynomial that pointwise approximates f to within constant error in a certain one-sided sense (made precise in Section 3). This is the complexity measure that we amplify for constant-depth circuits: given a depth k circuit f on m variables that has one-sided approximate degree greater than d, we show how to generically transform f into a depth $k+1$ circuit F on $t \cdot m$ variables such that F cannot be pointwise approximated by degree d polynomials even to error $1 - 2^{-t}$.

Theorem 1. *Suppose $f : \{-1,1\}^m \to \{-1,1\}$ has one-sided approximate degree $\widetilde{\deg}_{1/2}(f) > d$. Denote by $F : \{-1,1\}^{m \cdot t} \to \{-1,1\}$ the block-wise composition $\mathrm{OR}_t(f, \ldots, f)$, where OR_t denotes the OR function on t variables. Then F cannot be pointwise approximated by degree-d polynomials even to within error $1 - 2^{-t}$ by degree-d polynomials. That is, the $(1 - 2^{-t})$-approximate degree of F is greater than d.*

A *dual formulation* of one-sided approximate degree was previously exploited by Gavinsky and Sherstov to separate the communication versions of NP and co-NP [15], as well as by the current authors [12] and independently by Sherstov [38] to resolve the approximate degree of the two-level

AND-OR tree. In this paper, we introduce a primal formulation of one-sided approximate degree. This allows us to express Theorem 1 as a form of hardness amplification from one-sided approximate degree to approximate degree.

Prior Work on Hardness Amplification for Approximate Degree. For the purposes of this discussion, we informally consider a hardness amplification result for approximate degree to be any statement of the following form: Fix two functions $f : \{-1,1\}^m \to \{-1,1\}$ and $g : \{-1,1\}^t \to \{-1,1\}$. Then the composed function $g(f,\ldots,f) : \{-1,1\}^{m\cdot t} \to \{-1,1\}$ is strictly harder to approximate in the ℓ_∞ norm by low-degree polynomials than is the original function g.

We think of such a result as establishing that application of the outer function g to t disjoint copies of f amplifies the hardness of f. Here we consider polynomial degree to be a resource, and "harder to approximate" can refer either to the amount of resources required for the approximation, to the error of the approximation, or to a combination of the two.

Two particular kinds of hardness amplification results for approximate degree have received particular attention. *Direct-sum* theorems focus on amplifying the degree required to obtain an approximation, but do not focus on amplifying the error. For example, a typical direct-sum theorem identifies conditions on f and g that guarantee that $\widetilde{\deg}_\varepsilon(g(f,\ldots,f)) \geq \widetilde{\deg}_\varepsilon(g) \cdot \widetilde{\deg}_\varepsilon(f)$. In contrast, a *direct-product* theorem focuses on amplifying both the error and the minimum degree required to achieve this error. An *XOR lemma* is a special case of either type of theorem where the combining function g is the XOR function. Ideally, an XOR lemma of the direct-product form establishes that there exists a sufficiently small constant $\delta > 0$ such that $\widetilde{\deg}_{1-2^{-\delta t}}(\mathrm{XOR}_t(f,\ldots,f)) \geq t \cdot \widetilde{\deg}_{1/3}(f)$. That is, an XOR lemma establishes that approximating the XOR of t disjoint copies of f requires a t-fold blowup in degree relative to f, even if one allows error exponentially close to 1.

O'Donnell and Servedio [34] proved an XOR lemma for *threshold degree*, establishing that $\mathrm{XOR}_t(f,\ldots,f)$ has threshold degree t times the threshold degree of f. In later work, Sherstov [46] proved a direct sum result for approximate degree that holds whenever the combining function g has low block-sensitivity. His techniques also capture O'Donnell and Servedio's XOR lemma for threshold degree as a special case. In [44], Sherstov proved a number of hardness amplification results for approximate degree. Most notably, he proved an optimal XOR lemma, as well as a direct-sum theorem that holds whenever the combining function has close to maximal approximate degree (i.e., approximate degree $\Omega(t)$). Sherstov used his XOR lemma to prove direct product theorems for quantum query complexity, and in subsequent work [45], to show direct product theorems for the multiparty communication of set disjointness.

Comparison to Prior Work. In this paper, we are interested in establishing approximate degree lower bounds for constant-depth circuits over the basis $\{\mathrm{AND}, \mathrm{OR}, \mathrm{NOT}\}$. For this purpose, it is essential to consider combining functions (such as OR, see Theorem 1) that are themselves in AC^0, ruling out the use of XOR as a combining function. Our hardness amplification result (Theorem 1) is orthogonal to direct-sum theorems: direct-sum theorems focus on amplifying degree but not error, while Theorem 1 focuses on amplifying error but not degree. Curiously, Theorem 1 is nonetheless a critical ingredient in our proof of a direct-sum type theorem for AND-OR trees of constant depth (Theorem 9 below).

2.2 Lower Bounds For AC^0

In our primary applications of Theorem 1, we let $f : \{-1,1\}^m \to \{-1,1\}$ be the Element Distinctness function (defined in Section 3). Aaronson and Shi showed that the approximate degree

Reference	Discrepancy Bound	Circuit Depth
Sherstov [43]	$\exp(-\Omega(n^{1/5}))$	3
Buhrman et al. [11]	$\exp(-\Omega(n^{1/3}))$	3
Sherstov [42]	$\exp(-\Omega(n^{1/3}))$	3
This work	$\exp\left(-\tilde{\Omega}(n^{2/5})\right)$	4

Table 1: Comparison of our new discrepancy bound for AC^0 to prior work. The circuit depth column lists the depth of the circuit used to exhibit the bound.

of Element Distinctness is $\Omega(m^{2/3}/\log m)$ [2]. This is the best-known lower bound for the approximate degree of a function in AC^0. We show in Appendix A that even the *one-sided* approximate degree of Element Distinctness is $\Omega(m^{2/3}/\log m)$.

Applying Theorem 1 to Element Distinctness, we obtain a depth-three Boolean circuit F with $m \cdot t$ inputs such that $\widetilde{\deg}_{\varepsilon}(F) = \tilde{\Omega}(m^{2/3})$, for $\varepsilon = 1 - 2^{-t}$. By choosing t and m appropriately, we obtain a depth-three circuit on $n = t \cdot m$ variables of size $\mathrm{poly}(n)$ such that any degree-d polynomial cannot pointwise approximate F to error better than $1 - \exp\left(-\tilde{\Omega}(nd^{-3/2})\right)$.

Corollary 2. *For every $d > 0$, there is a depth-3 Boolean circuit $F : \{-1,1\}^n \to \{-1,1\}$ of size $\mathrm{poly}(n)$ such that any degree-d polynomial cannot pointwise approximate F to error better than $1 - \exp\left(-\tilde{\Omega}(nd^{-3/2})\right)$. In particular, any polynomial of degree at most $n^{2/5}$ cannot pointwise approximate F to error better than $1 - \exp\left(-\tilde{\Omega}(n^{2/5})\right)$.*

Discrepancy Upper Bound. Discrepancy, defined formally in Section 5, is a central quantity in communication complexity and circuit complexity. For instance, upper bounds on the discrepancy of a function f immediately yield lower bounds on the cost of small-bias communication protocols for computing f (see Section 2.5 for details). The first exponentially small discrepancy upper bounds for AC^0 were proved by Burhman et al. [11] and Sherstov [42,43], who exhibited constant-depth circuits with discrepancy $\exp(-\Omega(n^{1/3}))$. Our results improve the best-known upper bound to $\exp\left(-\tilde{\Omega}(n^{2/5})\right)$.

In particular, Sherstov [42] developed a powerful technique, known as the pattern-matrix method, that allows one to automatically translate lower bounds on the ε-approximate degree of a Boolean function F into lower bounds on the *discrepancy* of a related function F' as long as ε is exponentially close to one. By combining the pattern-matrix method with Corollary 2, we obtain the following result.

Corollary 3. *There is a depth-4 Boolean circuit $F' : \{-1,1\}^n \to \{-1,1\}$ with discrepancy $\exp\left(-\tilde{\Omega}(n^{2/5})\right)$.*

Threshold Weight Lower Bounds. A *polynomial threshold function* (PTF) for a Boolean function f is a multilinear polynomial p with integer coefficients that agrees in sign with f on all Boolean inputs. The *weight* of an n-variate polynomial p is the sum of the absolute value of its coefficients. The *degree-d threshold weight* of a Boolean function $f : \{-1,1\}^n \to \{-1,1\}$, denoted $W(f,d)$, refers to the least weight of a degree-d PTF for f. We let $W(f)$ denote the quantity

$W(f,n)$, i.e., the least weight of any threshold function for f regardless of its degree. As discussed below in Section 2.5, threshold weight has important applications in learning theory.

Threshold weight is closely related to ε-approximate degree when ε is very close to 1 (see Section 3.3). We can thus translate Corollary 2 into lower bounds on the degree-d threshold weight of AC^0.

Corollary 4. *For every $d > 0$, there is a depth-3 Boolean circuit $F : \{-1,1\}^n \to \{-1,1\}$ of size $\mathrm{poly}(n)$ such that $W(F,d) \geq \exp\left(\tilde{\Omega}(nd^{-3/2})\right)$. In particular, $W(F,n^{2/5}) = \exp\left(\tilde{\Omega}(n^{2/5})\right)$.*

A result of Krause [26] allows us to extend our new degree-d threshold weight lower bound for F into a *degree independent* threshold weight lower bound for a related function F'. The previous best lower bound on the threshold weight of AC^0 was $\exp\left(\Omega(n^{1/3})\right)$, due to Krause and Pudlák [27].

Corollary 5. *There is a depth-4 Boolean circuit $F' : \{-1,1\}^n \to \{-1,1\}$ satisfying $W(F') = \exp\left(\tilde{\Omega}(n^{2/5})\right)$.*

Moreover, while the threshold weight bound of Corollary 5 is stated for polynomial threshold functions over $\{-1,1\}^n$, we show that the same threshold weight lower bound also holds for polynomials over $\{0,1\}^n$.

2.3 Lower Bounds for Read-Once DNFs and CNFs

Our techniques also yield new lower bounds on the approximate degree and degree-d threshold weight of read-once DNF and CNF formulas. Before stating our results, we discuss relevant prior work.

In their seminal work on perceptrons, Minsky and Papert exhibited a read-once DNF $f : \{-1,1\}^n \to \{-1,1\}$ with *threshold degree* $\Omega(n^{1/3})$ [31]. That is, *no* polynomial p of degree $o(n^{1/3})$ can sign-represent f, regardless of the weight of p. However, to our knowledge no non-trivial lower bound on the degree-d threshold weight of read-once DNFs was known for any $d = \omega(n^{1/3})$.

In an influential result, Beigel [9] exhibited a polynomial-size (read-many) DNF called ODD-MAX-BIT satisfying the following: there is some constant $\delta > 0$ such that $\widetilde{\deg}_{1-2^{-\delta n/d^2}}(\text{ODD-MAX-BIT}) > d$, and hence also $W(\text{ODD-MAX-BIT}, d) = \exp\left(\Omega(n/d^2)\right)$ (see Section 3.3). Klivans and Servedio showed that Beigel's lower bound is essentially tight for $d < n^{1/3}$ [24]. Very recently, Servedio, Tan, and Thaler showed an alternative lower bound on the degree-d threshold weight of ODD-MAX-BIT. Specifically, they showed that $W(\text{ODD-MAX-BIT}, d) = \exp\left(\Omega\left(\sqrt{n/d}\right)\right)$ [37]. The lower bound of Servedio et al. improves over Beigel's for any $d > n^{1/3}$, and is essentially tight in this regime (i.e., when $d > n^{1/3}$).

While ODD-MAX-BIT is a relatively simple DNF (in fact, it is a *decision list*), it is not a read-once DNF. Our results extend the lower bounds of Servedio et al. and Beigel from decision lists to read-once DNFs and CNFs. In the statement of the results below, we restrict ourselves to DNFs, as the case of CNFs is entirely analogous.

2.3.1 Extending the Lower Bound of Servedio et al. to Read-Once DNFs

In order to extend the lower bound of Servedio et al. to read-once DNFs and CNFs, we extend our hardness amplification techniques from one-sided approximate degree to a quantity we call

degree-d one-sided non-constant approximate weight. This quantity captures the least *weight* of a polynomial of degree at most d that pointwise approximates f in a certain one-sided sense (again, made precise in Section 3). We denote the degree-d one-sided approximate weight of a Boolean function f by $W^*_\varepsilon(f,d)$, where ε is an error parameter.

We prove the following analog of Theorem 1.

Theorem 6. *Fix $d > 0$. Let $f : \{-1,1\}^m \to \{-1,1\}$, and suppose that $W^*_{3/4}(f,d) > w$. Let $F : \{-1,1\}^{m\cdot t} \to \{-1,1\}$ denote the function $\mathrm{OR}_t(f,\ldots,f)$. Then any degree-$d$ polynomial that approximates F to within error $1-2^{-t}$ requires weight $2^{-5t}w$.*

Adapting a proof of Servedio et al., we can show that $W^*_{3/4}(\mathrm{AND}_m, d) \geq 2^{\Omega(m/d)}$. By applying Theorem 6 with $f = \mathrm{AND}_m$, along with standard manipulations, we are able to extend the lower bound of Servedio et al. to read-once CNFs and DNFs.

Corollary 7. *For each $d = o(n/\log^4 n)$, there is a read-once DNF F satisfying $W(F,d) = \exp\left(\Omega(\sqrt{n/d})\right)$.*

In particular, there is a read-once DNF that cannot be computed by any PTF of $\mathrm{poly}(n)$ weight, unless the degree is $\tilde{\Omega}(n)$.

2.3.2 Extending Beigel's Lower Bound to Read-Once DNFs

It is known that $\widetilde{\mathrm{odeg}}(\mathrm{AND}_m) = \Omega(m^{1/2})$. By applying Theorem 1 with $f = \mathrm{AND}_m$, we obtain the following result.

Corollary 8. *There is an (explicit) read-once DNF $F : \{-1,1\}^n \to \{-1,1\}$ with $\widetilde{\deg}_{1-2^{-n/d^2}}(F) = \Omega(d)$.*

We remark that for $d < n^{1/3}$, Corollary 8 is subsumed by Minsky and Papert's seminal result that exhibited a read-once DNF F with threshold degree $\Omega(n^{1/3})$ [31]. However, for $d > n^{1/3}$, it is not subsumed by Minsky and Papert's result, nor by Corollary 7. Indeed, Corollary 7 yields a lower bound on the degree-d threshold weight of read-once DNFs, but not a lower bound on the *approximate-degree* of read-once DNFs (see Section 3.3 for further discussion on the separation between these quantities).

2.4 Approximate Degree Lower Bounds for AND-OR Trees

The d-level AND-OR tree on n variables is a function described by a read-once circuit of depth d consisting of alternating layers of AND gates and OR gates. We assume for simplicity that all gates have fan-in $n^{1/d}$. For example, the two-level AND-OR tree is a read-once CNF in which all gates have fan-in $n^{1/2}$.

Until recently, the approximate degree of AND-OR trees of depth two or greater had resisted characterization, despite 19 years of attention [4,12,18,33,38,46,48]. The case of of depth two was reposed as challenge problem by Aaronson in 2008 [1], as it captured the limitations of existing lower bound techniques. This case was resolved earlier this year by the current authors [12], and independently by Sherstov [38], who proved a lower bound of $\Omega(\sqrt{n})$, matching an upper bound of Høyer, Mosca, and de Wolf [18]. However, the case of depth three or greater remained open. To our knowledge, the best known lower bound for $d \geq 3$ was $\Omega(n^{1/4+1/2d})$, which follows by combining

the depth-two lower bound [12, 38] with an earlier direct-sum theorem of Sherstov [46, Theorem 3.1].

By combining our hardness amplification result (Theorem 1) with techniques of our earlier work [12], we improve this lower bound to $\Omega\left(n^{1/2}/\log^{(d-2)/2}(n)\right)$ for any constant $d \geq 2$. A result of Sherstov [40] yields an upper bound of $O(n^{1/2})$ for constant d, demonstrating that our result is optimal up to polylogarithmic factors.

Theorem 9. *Let* AND-OR$_{d,n}$ *denote the d-level AND-OR tree on n variables. Then* $\widetilde{\deg}(\text{AND-OR}_{d,n}) = \Omega\left(n^{1/2}/\log^{(d-2)/2} n\right)$ *for any constant* $d > 0$.

2.5 Applications

In this section, we detail applications of the results described above to communication complexity, circuit complexity, and computational learning theory.

2.5.1 Communication Complexity

Let $f : X \times Y \to \{-1, 1\}$, where X and Y are finite sets. Consider a two-party communication problem in which Alice is given an input $x \in X$, Bob is given an input $y \in Y$, and their goal is to compute $f(x, y)$ with probability $1/2 + \beta$ for some bias $\beta > 0$. Alice and Bob each have access to an arbitrarily long sequence of private random bits, and the cost $C(P)$ of a protocol P is the worst-case number of bits they must exchange over all inputs $(x, y) \in X \times Y$. Babai et al. [5] defined the *PP communication* model to capture the complexity of computing f with small bias. The PP communication complexity of f, denoted by $\text{PP}(f)$, is the minimum value of $C(P) + \log(1/\beta(P))$ over all protocols P that compute f with positive bias.

It is well known [21] that PP communication is essentially characterized by discrepancy: if $f : \{-1, 1\}^n \times \{-1, 1\}^n \to \{-1, 1\}$, then $\text{PP}(f) = \Theta\left(\log\left(1/\operatorname{disc}(f)\right) + \log n\right)$. It follows immediately that our $\exp\left(-\tilde{\Omega}(n^{2/5})\right)$ upper bound on the discrepancy of an AC^0 function f implies an $\tilde{\Omega}(n^{2/5})$ lower bound on $\text{PP}(f)$. The previous best lower bound on $\text{PP}(f)$ for an AC^0 function f was $\Omega(n^{1/3})$ [11, 42].

2.5.2 Circuit Complexity

Constant-depth circuits of majority gates are known to be surprisingly powerful. Most strikingly, Allender [3] showed that any function in AC^0 can be computed by a depth three circuit of majority gates of quasipolynomial size. This prompted Krause and Pudlák [27] to ask whether every AC^0 function could be computed by depth *two* majority gates of polynomial size. This question was resolved in the negative by Sherstov [43], who exhibited an AC^0 function that cannot be computed even by majority-of-threshold circuits of size $\exp(n^{1/5})$ (later sharpened to $\exp(n^{1/3})$ [42]), and independently by Buhrman, Vereshchagin, and de Wolf [11], who obtained an $\exp(n^{1/3})$ lower bound on the size of majority-of-threshold circuits computing a different AC^0 function.

It is well-known that a discrepancy upper bound for F yields a lower bound on the size of majority-of-threshold circuits computing F [16, 17, 32, 43], and indeed, the circuit lower bounds of [11, 42, 43] are all proved using discrepancy. Through this connection, our discrepancy upper

bound of Corollary 3 sharpens the previous lower bounds by yielding a depth-four Boolean circuit F of polynomial size such that any majority-of-threshold circuit computing F requires size $\exp\left(\tilde{\Omega}(n^{2/5})\right)$.

Corollary 10. *There is a depth-four Boolean circuit $F : \{-1,1\}^n \to \{-1,1\}$ of size $\mathrm{poly}(n)$ such that every majority-of-threshold circuit computing F has size $\exp\left(\tilde{\Omega}(n^{2/5})\right)$.*

2.5.3 Learning Theory

Our results have a number of consequences in computational learning theory. We discuss them below.

Technical Background: The Generalized Winnow Algorithm. The Generalized Winnow algorithm is one of the most powerful known algorithms for online learning [24,30,37]. Suppose we are given a concept class $\mathcal{C}$ of functions mapping n-bit inputs to $\{-1,1\}$, as well as a collection of polynomial-time computable "feature" functions $\mathcal{F}$. The Generalized Winnow algorithm learns a concept in $\mathcal{C}$ by maintaining as a hypothesis a low-weight linear threshold function of features in $\mathcal{F}$.

Suppose that each $f \in \mathcal{C}$ has a low-weight linear threshold representation

$$f(x) = \mathrm{sgn}\left(\sum_{h_i \in \mathcal{F}} w_i h_i(x)\right),$$

where each w_i is an integer, and $\sum_i |w_i| \leq W$. A remarkable property of the Generalized Winnow algorithm is that its mistake bound depends only *logarithmically* on the size of the feature set $\mathcal{F}$, and polynomially on the weight bound W (here the mistake bound refers to the worst-case number of mistakes an online learning algorithm makes over any sequence of examples). Meanwhile, its running time per example is polynomial in the size of the feature set. Standard techniques can be used to transform any online learning algorithm into a PAC learning algorithm whose sample complexity is proportional to the mistake bound.

PAC Learning $\mathbf{AC^0}$ via Generalized Winnow. Valiant famously posed the problem of PAC learning DNF formulas in his original paper [51] introducing the PAC model. The fastest known algorithm for this problem is due to Klivans and Servedio. It is based on linear programming, and takes time $\exp\left(\tilde{O}(n^{1/3})\right)$ [23]. At the core of this algorithm is a fundamental structural result for DNFs: Klivans and Servedio showed that every DNF of size s can be computed by a polynomial threshold function of degree $O(n^{1/3}\log s)$. However, the weight of the PTF arising in this construction can grow doubly-exponentially with n. Klivans and Servedio asked whether it is possible that every polynomial-size DNF has a PTF of degree $\tilde{O}(n^{1/3})$, and weight $\exp\left(\tilde{O}(n^{1/3})\right)$ – an affirmative answer to this question would imply that the Generalized Winnow Algorithm (run over the feature set of all low-degree parities) can also PAC learn DNFs in time $\exp\left(\tilde{O}(n^{1/3})\right)$. Such a result would be attractive, as the Generalized Winnow algorithm is substantially simpler than the linear programming algorithm of Klivans and Servedio.

While we do not resolve the question of Klivans and Servedio for DNFs, we do resolve it in the negative for depth-three circuits. In fact, we rule out the possibility of the Generalized Winnow algorithm PAC learning depth-three Boolean circuits in time $\exp\left(\tilde{O}(n^{2/5})\right)$ *regardless* of the underlying feature set. That is, our lower bound holds even on feature sets that are not low-degree parities.

Specifically, Corollary 4 implies the following result. The proof is identical to [43, Theorem 8.1] and is omitted for brevity.

Corollary 11. *Let $\mathcal{C}$ denote the concept class of polynomial-size depth-three Boolean circuits. Let $\mathcal{F} = \{h_1, \ldots, h_m : \{-1,1\}^n \to \{-1,1\}\}$ be arbitrary Boolean functions such that every $f \in \mathcal{C}$ can be expressed as $f(x) = \operatorname{sgn}\left(\sum_{i=1}^{m} w_i h_i(x)\right)$ for some integers $w_1, \ldots, w_m$ with $|w_1| + \cdots + |w_m| \leq W$. Then $m \cdot W > \exp\left(\tilde{\Omega}(n^{2/5})\right)$.*

PAC Learning $\mathbf{AC^0}$ via Boosting. While an $\exp\left(\tilde{\Omega}(n^{1/3})\right)$-time algorithm is known for PAC learning polynomial-size DNF formulas, no $\exp(o(n))$-time algorithm is known even for learning polynomial-size depth-three Boolean circuits. A natural approach to this problem is as follows. Suppose that every function f in a concept class $\mathcal{C}$ can be computed by a PTF (of arbitrary degree) over $\{0,1\}^n$ with weight at most W. The well-known *discriminator lemma* of Hajnal et al. [17] implies that under *any* distribution, there is some conjunction (possibly of width $\Omega(n)$) that has correlation at least $1/W$ with f. One can then apply an agnostic learning algorithm for conjunctions (such as the $\exp\left(\tilde{O}(n^{1/2})\right)$-time polynomial regression algorithm of Kalai et al. [19]), combined with standard boosting techniques, to PAC-learn $\mathcal{C}$ in time polynomial in $\max\left(\exp\left(\tilde{O}(n^{1/2})\right), W\right)$.

Thus, if one could prove an $\exp(\tilde{O}(n^{1/2}))$ upper bound (for PTFs over $\{0,1\}^n$) on the threshold weight of AC^0, one would obtain an $\exp(\tilde{O}(n^{1/2}))$-time algorithm for PAC learning AC^0. While our $\exp(\tilde{\Omega}(n^{2/5}))$ threshold weight lower bound for AC^0 does not rule out this possibility, it does establish new limitations for this technique. In particular, our threshold weight lower bound implies that even if faster algorithms for agnostically learning conjunctions are discovered, this boosting-based approach to learning AC^0 cannot run in time better than $\exp\left(\tilde{\Omega}(n^{2/5})\right)$.

Attribute-Efficient Learning. Attribute-efficient learning is a clean framework that captures the challenging and important problem of learning in the presence of irrelevant information [10]. A class $\mathcal{C}$ of Boolean functions over $\{-1,1\}^n$ is said to be attribute-efficiently learnable if there is a $\mathrm{poly}(n)$-time online algorithm that learns any $f \in \mathcal{C}$ with mistake bound polynomial in the representation size of f. For example, the concept class of read-once DNFs that depend on $k \ll n$ of their input variables is attribute-efficiently learnable if there is an online learning algorithm for this class that runs in time $\mathrm{poly}(n)$ per example and achieves mistake bound $\mathrm{poly}(k, \log n)$.

Attribute-efficient learning is a challenging problem, and many simple concept classes are not known to be attribute-efficiently learnable, including decision lists and read-once DNFs. The Generalized Winnow algorithm, run over the feature-space of low-degree parities, marks the best progress toward attribute-efficient learning of these concept classes (see e.g. [24, 37]). Prior to our work, it was unknown whether this approach could learn read-once DNFs depending on k variables in time $\exp\left(\tilde{O}(n^{1/3})\right)$ per example and with mistake bound $\mathrm{poly}(k, \log n)$, as such a guarantee would hold if every read-once DNF on n variables were computed by a polynomial threshold function of degree

$\tilde{O}(n^{1/3})$ and weight poly(n). Corollary 7 rules out this possibility in a very strong sense, as it implies the existence of a read-once DNF that cannot be computed by any PTF of poly(n) weight, unless the degree is $\tilde{\Omega}(n)$. Similarly, Corollary 2 establishes new limitations on the efficiency of the Generalized Winnow algorithm in the context of attribute-efficient learning of depth-three Boolean circuits.

2.6 Organization

Section 3 establishes terminology, introduces our main technique based on LP-duality, and proves essential technical lemmas. Section 4 establishes our central hardness amplification result for approximate degree (Theorem 1). It then applies this result to the Element Distinctness function to obtain our new lower bounds on "accuracy vs. degree" tradeoffs for pointwise approximating AC^0 by polynomials (Corollary 2). Section 5 proves our new discrepancy upper bound for an AC^0 function (Corollary 3). Section 6 proves our new threshold weight lower bound for AC^0 (Corollaries 4 and 5). Section 7 proves our new lower bounds for read-once DNFs (Theorem 6, Corollary 7, and Corollary 8). Section 8 proves our new approximate degree lower bound for AND-OR trees (Theorem 9). Section 9 concludes with suggestions for further research directions.

3 Preliminaries

We work with Boolean functions $f : \{-1,1\}^n \to \{-1,1\}$ under the standard convention that 1 corresponds to logical false, and -1 corresponds to logical true. For a real-valued function $r : \{-1,1\}^n \to \mathbb{R}$, we let $\|r\|_\infty = \max_{x\in\{-1,1\}^n} |r(x)|$ denote the ℓ_∞ norm of r. We let OR_n and AND_n denote the OR function and AND function on n variables respectively. Define $\widetilde{\mathrm{sgn}}(t) = -1$ if $t < 0$ and 1 otherwise. For a set $S \subseteq [n] = \{1,\ldots,n\}$, let $\chi_S(x) := \prod_{i\in S} x_i$ denote the parity function over variables indexed by S.

We now define the notions of approximate degree, approximate weight, threshold degree, threshold weight, and their one-sided variants.

3.1 Polynomial Approximations and their Dual Characterizations

3.1.1 Approximate Degree

The ε-approximate degree of a function $f : \{-1,1\}^n \to \{-1,1\}$, denoted $\widetilde{\deg}_\varepsilon(f)$, is the minimum (total) degree of any real polynomial p such that $\|p - f\|_\infty \leq \varepsilon$, i.e., $|p(x) - f(x)| \leq \varepsilon$ for all $x \in \{-1,1\}^n$. We use $\widetilde{\deg}(f)$ to denote $\widetilde{\deg}_{1/3}(f)$, and use this to refer to the *approximate degree* of a function without qualification. The choice of $1/3$ is arbitrary, as $\widetilde{\deg}(f)$ is related to $\widetilde{\deg}_\varepsilon(f)$ by a constant factor for any constant $\varepsilon \in (0,1)$.

Given a Boolean function f, let p be a real polynomial that minimizes $\|p - f\|_\infty$ among all polynomials of degree at most d. Since we work over $x \in \{-1,1\}^n$, we may assume without loss of generality that p is multilinear with the representation $p(x) = \sum_{|S|\leq d} c_S \chi_S(x)$ where the coefficients c_S are real numbers. Then p is an optimum of the following linear program.

min	ε	
such that	$\left\lvert f(x) - \sum_{\lvert S\rvert \le d} c_S \chi_S(x)\right\rvert \le \varepsilon$	for each $x \in \{-1,1\}^n$
	$c_S \in \mathbb{R}$	for each $\lvert S\rvert \le d$
	$\varepsilon \ge 0$	

The dual LP is as follows.

max	$\sum_{x\in\{-1,1\}^n} \phi(x) f(x)$	
such that	$\sum_{x\in\{-1,1\}^n} \lvert\phi(x)\rvert = 1$	
	$\sum_{x\in\{-1,1\}^n} \phi(x)\chi_S(x) = 0$	for each $\lvert S\rvert \le d$
	$\phi(x) \in \mathbb{R}$	for each $x \in \{-1,1\}^n$

Strong LP-duality thus yields the following well-known dual characterization of approximate degree (cf. [42]).

Theorem 12. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be a Boolean function. Then $\widetilde{\deg}_\varepsilon(f) > d$ if and only if there is a polynomial $\phi : \{-1,1\}^n \to \mathbb{R}$ such that*

$$\sum_{x\in\{-1,1\}^n} f(x)\phi(x) > \varepsilon, \tag{1}$$

$$\sum_{x\in\{-1,1\}^n} |\phi(x)| = 1, \tag{2}$$

and

$$\sum_{x\in\{-1,1\}^n} \phi(x)\chi_S(x) = 0 \text{ for each } |S| \le d. \tag{3}$$

If ϕ satisfies Eq. (3), we say ϕ has *pure high degree d*. We refer to any feasible solution ϕ to the dual LP as a *dual polynomial* for f.

3.1.2 One-Sided Approximate Degree

We introduce a relaxed notion of the approximate degree of f which we call the one-sided ε-approximate degree, denoted by $\widetilde{\mathrm{odeg}}_\varepsilon(f)$. This is the least degree of a real polynomial p with one-sided distance at most ε from f, where the *one-sided distance* between p and f is defined to be the smallest ε such that

1. $|p(x) - 1| \le \varepsilon$ for all $x \in f^{-1}(1)$.
2. $p(x) \le -1 + \varepsilon$ for all $x \in f^{-1}(-1)$.

That is, we require p to be very accurate on inputs in $f^{-1}(1)$, but only require "one-sided accuracy" on inputs in $f^{-1}(-1)$. We use $\widetilde{\mathrm{odeg}}(f)$ to denote $\widetilde{\mathrm{odeg}}_{1/3}(f)$, and refer to this quantity without qualification as the *one-sided approximate degree* of f.

The primal and dual LPs change in a simple but crucial way if we look at one-sided approximate degree rather than approximate degree. Let $p(x) = \sum_{|S|\le d} c_S \chi_S(x)$ be a polynomial of degree d that minimizes the one-sided distance from f. Then p is an optimum of the following linear program.

min	ε	
such that	$\left\lvert f(x) - \sum_{\lvert S\rvert \le d} c_S \chi_S(x)\right\rvert \le \varepsilon$	for each $x \in f^{-1}(1)$
	$\sum_{\lvert S\rvert \le d} c_S \chi_S(x) \le -1 + \varepsilon$	for each $x \in f^{-1}(-1)$
	$c_S \in \mathbb{R}$	for each $\lvert S\rvert \le d$
	$\varepsilon \ge 0$	

The dual LP is as follows.

max	$\sum_{x\in\{-1,1\}^n} \phi(x) f(x)$	
such that	$\sum_{x\in\{-1,1\}^n} \lvert\phi(x)\rvert = 1$	
	$\sum_{x\in\{-1,1\}^n} \phi(x)\chi_S(x) = 0$	for each $\lvert S\rvert \le d$
	$\phi(x) \le 0$ for each $x \in f^{-1}(-1)$	
	$\phi(x) \in \mathbb{R}$	for each $x \in \{-1,1\}^n$

We again appeal to strong LP-duality for the following dual characterization of one-sided approximate degree.

Theorem 13. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be a Boolean function. Then $\widetilde{\mathrm{odeg}}_\varepsilon(f) > d$ if and only if there is a polynomial $\phi : \{-1,1\}^n \to \mathbb{R}$ such that*

$$\sum_{x\in\{-1,1\}^n} f(x)\phi(x) > \varepsilon, \tag{4}$$

$$\sum_{x\in\{-1,1\}^n} |\phi(x)| = 1, \tag{5}$$

$$\sum_{x\in\{-1,1\}^n} \phi(x)\chi_S(x) = 0 \text{ for each } |S| \le d, \tag{6}$$

and

$$\phi(x) \le 0 \text{ for each } x \in f^{-1}(-1). \tag{7}$$

Observe that a feasible solution ϕ to this dual LP is a feasible solution to the dual LP for approximate degree, with the additional constraint that $\phi(x)$ agrees in sign with $f(x)$ whenever $x \in f^{-1}(-1)$. We refer to any such feasible solution ϕ as a dual polynomial for f with *one-sided error*. Dual polynomials with one-sided error have recently played an important role in resolving open problems in communication complexity [15] and resolving the approximate degree of the two-level AND-OR tree [12, 38]. They will play a critical role in our proof of Theorem 1 as well.

3.1.3 Approximate Weight

We define the *degree-d ε-approximate weight* of f, $W_\varepsilon(f,d)$, to be the minimum weight of a degree-d polynomial that approximates f pointwise to error ε. Recall that the weight of a polynomial p is the L_1 norm of its coefficients. If $\widetilde{\deg}_\varepsilon(f) > d$, we define $W_\varepsilon(f,d) = \infty$.

For a fixed error parameter ε and degree d, the degree-d ε-approximate weight of a function f is captured by the following linear program.

min	$\sum_{\lvert S\rvert \le d} \lvert c_S\rvert$	
such that	$\left\lvert f(x) - \sum_{\lvert S\rvert \le d} c_S \chi_S(x)\right\rvert \le \varepsilon$	for each $x \in \{-1,1\}^n$
	$c_S \in \mathbb{R}$	for each $\lvert S\rvert \le d$

The dual LP is as follows.

max	$\sum_{x\in\{-1,1\}^n} \phi(x) f(x) - \varepsilon \sum_{x\in\{-1,1\}^n} \lvert\phi(x)\rvert$	
such that	$\left\lvert \sum_{x\in\{-1,1\}^n} \phi(x)\chi_S(x)\right\rvert \le 1$	for each $\lvert S\rvert \le d$
	$\phi(x) \in \mathbb{R}$	for each $x \in \{-1,1\}^n$

We thus obtain the following duality theorem.

Theorem 14. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be a Boolean function. Then $W_\varepsilon(f,d) > w$ if and only if there is a polynomial $\phi : \{-1,1\}^n \to \mathbb{R}$ such that*

$$\sum_{x\in\{-1,1\}^n} f(x)\phi(x) - \varepsilon \sum_{x\in\{-1,1\}^n} |\phi(x)| > w, \tag{8}$$

$$\left|\sum_{x\in\{-1,1\}^n} \phi(x)\chi_S(x)\right| \le 1 \text{ for each } |S| \le d. \tag{9}$$

3.1.4 One-Sided Non-Constant Approximate Weight

To derive our new lower bound on the degree-d threshold weight of read-once DNFs (Corollary 7), we need the following technical variation on approximate weight. Given a polynomial $p(x) = \sum_S c_S\chi_S(x)$, define the *non-constant weight* of p to be the L_1 norm of its coefficients excluding the constant term, i.e., $\sum_{S\neq\emptyset} |c_S|$. We then define the *degree-d one-sided non-constant ε-approximate weight* of f, denoted by $W^*_\varepsilon(f,d)$ to be the minimum non-constant weight of a polynomial that approximates f to one-sided distance ε. Linear programming duality yields the following characterization of $W^*_\varepsilon(f,d)$.

Theorem 15. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be a Boolean function. Then $W^*_\varepsilon(f,d) > w$ if and only if there is a polynomial $\phi : \{-1,1\}^n \to \mathbb{R}$ such that*

$$\sum_{x\in\{-1,1\}^n} f(x)\phi(x) - \varepsilon \sum_{x\in\{-1,1\}^n} |\phi(x)| > w, \tag{10}$$

$$\left|\sum_{x\in\{-1,1\}^n} \phi(x)\chi_S(x)\right| \le 1 \text{ for each } 0 < |S| \le d, \tag{11}$$

$$\sum_{x\in\{-1,1\}^n} \phi(x) = 0, \tag{12}$$

$$\phi(x) \le 0 \text{ for each } x \in f^{-1}(-1). \tag{13}$$

3.1.5 Threshold Degree and Threshold Weight

We say a polynomial $p(x) = \sum_S c_S \chi_S(x)$ with *integer* coefficients is a polynomial threshold function (PTF) for a Boolean function f if p sign-represents f at all Boolean inputs, i.e., if $f(x)p(x) > 0$ for all $x \in \{-1,1\}^n$. The *threshold degree* of f, $\deg_\pm(f)$, is the minimum degree of a PTF for f.

The *threshold weight* $W(f)$ is the minimum weight of any PTF for f. Observe that this definition is only meaningful because the coefficients of any PTF for f are required to be integers, as any positive constant multiple of a PTF for f also sign-represents f. More generally, it is of interest to study the tradeoff between the weight and degree necessary for PTF representations. To this end, we define the *degree-d threshold weight* $W(f,d)$ to be the minimum weight of a degree-d PTF for f. If $\deg_\pm(f) > d$, define $W(f,d) = \infty$.

While threshold weight is naturally captured by an *integer* program rather than a linear program, it still admits an important dual characterization, obtained by combining results of Freund [14] and Hajnal et al. [17] (see also [16, 42]).

Theorem 16. *Let $f : \{-1,1\}^n \to \{-1,1\}$ and fix an integer $d \geq \deg_\pm(f)$. Then for every probability distribution μ on $\{-1,1\}^n$,*

$$|\mathbb{E}_{x\sim\mu}[f(x)\chi_S(x)]| \geq \frac{1}{W(f,d)} \text{ for each } |S| \leq d. \tag{14}$$

Moreover, there exists a distribution μ for which

$$|\mathbb{E}_{x\sim\mu}[f(x)\chi_S(x)]| \leq \left(\frac{2n}{W(f,d)}\right)^{1/2} \text{ for each } |S| \leq d. \tag{15}$$

3.2 The One-Sided Approximate Degree of AC^0

We now exhibit a depth-two circuit having one-sided approximate degree $\tilde{\Omega}(n^{2/3})$. Let N and R be positive integers such that $N \geq R$ and R is a power of 2. We define the Element Distinctness function with range R as follows. The function takes $n = N \log R$ bits as input, and interprets its input as N blocks $(x_1, \ldots, x_N)$ with each block consisting of $\log R$ bits. Each block is interpreted as a number in the range $[R]$, and the function evaluates to TRUE if and only if all N numbers are distinct.

It is straightforward to check that for $R = \text{poly}(N)$, the Element Distinctness function with range R is computed by a CNF formula of polynomial size. Indeed, the function evaluates to TRUE if and only if there is no number $K \in [R]$ for which there is a pair of distinct indices $i, j \in [N]$ such that $x_i = x_j = K$. Thus, the following natural CNF computes Element Distinctness (noting that for any fixed K, the inner formula is computed by a bitwise OR):

$$f(x_1, \ldots, x_N) = \bigwedge_{K=1}^{R} \bigvee_{i \neq j} (x_i \neq K) \vee (x_j \neq K).$$

Aaronson and Shi [2] showed that when $R > 3N/2$, the approximate degree of Element Distinctness is $\Omega(N^{2/3})$. Ambainis [4] extended the lower bound to the "small-range" case where $R = N$. For the remainder of the paper, we will use the term Element Distinctness without qualification to refer to the small-range case.

By manipulating a dual witness for the high approximate degree of Element Distinctness, we can in fact show that $\widetilde{\mathrm{odeg}}_{1/3}(\text{Element Distinctness}) = \Omega(N^{2/3})$. We provide the argument in Appendix A.

3.3 Relating Degree-d Threshold Weight to High-Error Approximations

In this paper, we will often need to translate lower bounds on $\widetilde{\deg}_{\varepsilon}(f)$ for some function f with ε very close to 1 into lower bounds on the degree-d threshold weight of f. This is possible because degree-d PTFs of weight w are closely related to degree-d pointwise approximations with error $1 - 1/w$. In fact, these notions are essentially equivalent when $w \geq \binom{n}{d}$ [42]. The relationships we will need are formalized in the following lemma.

Lemma 17. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be a Boolean function, and let $w > 0$. Then (1) $\Rightarrow$ (2) $\Rightarrow$ (3).*

(1) $\widetilde{\deg}_{1-\frac{1}{w}}(f) > d$.

(2) $W_{1-\frac{1}{w}}(f,d) > 1$.

(3) $W(f,d) > w$.

Lemma 17 implies that a PTF of degree d and weight w can be transformed into $(1-1/w)$-approximation of degree d. Indeed, the proof will go by way of such a transformation.

Proof. Clearly (1) implies (2), since $W_{1-\frac{1}{w}}(f,d) = \infty$ when $\widetilde{\deg}_{1-\frac{1}{w}}(f) > d$. To show that (2) implies (3), suppose there is a PTF p for f having weight w and degree d. Since p has integer coefficients and is nonzero on Boolean inputs, $|p(x)| \geq 1$ on $\{-1,1\}^n$. Moreover, $|p(x)| \leq w$ by the weight bound, so the polynomial $\frac{1}{w}p(x)$ is a $(1-\frac{1}{w})$-approximation to f with weight 1. □

Remark: We stress that the converse of Lemma 17 fails badly when $w \ll \binom{n}{d}$. For example, we show in Corollary 7 that for any $d > 0$ there exists a read-once DNF F satisfying $W(F,d) \geq \exp\left(\sqrt{n/d}\right)$. In particular, this yields an exponential lower bound on the degree-d threshold weight of F for any $d = n^{1-\delta}$, with $\delta > 0$ a constant. Yet it follows from a result of Sherstov [40] that $\widetilde{\deg}_{1/3}(F) = O(n^{1/2})$ for any read-once DNF F.

4 Lower Bounds for $\mathbf{AC^0}$

4.1 Hardness Amplification for Approximate Degree

In this section, we show how to generically transform a circuit f with one-sided approximate degree d into a circuit F with ε-approximate degree d for $\varepsilon = 1 - 2^{-t}$. That is, while f cannot be approximated to error $1/2$ by degree d polynomials, F cannot even be approximated to error $1 - 2^{-t}$ by polynomials of the same degree.

Intuitively, our proof proceeds by taking a dual witness ψ to the high one-sided approximate degree of f, and a certain dual witness Ψ for the function OR_t, and combines them to obtain a dual witness for the fact that $\widetilde{\deg}_{1-2^{-t}}(\mathrm{OR}_t(f,\ldots,f)) > d$. Our analysis of the combined dual witness crucially exploits two properties: first, that ψ has one-sided error and second, that the vector whose

entries are all equal to -1 has very large (in fact, maximal) Hamming distance from the unique input in $\mathrm{OR}_t^{-1}(1)$.

Our method of combining the two dual witnesses was first introduced by Sherstov [46, Theorem 3.3] and independently by Lee [28]. This method has also been exploited by the authors in [12] to resolve the approximate degree of the two-level AND-OR tree, and by Sherstov [44] to prove direct sum and direct product theorems for polynomial approximation. Our principle insight in the proof of Theorem 1 lies in our choice of the appropriate dual witness for OR_t to use in the proof, and subsequent analysis of the combined dual witness.

Theorem 1. *Let $f : \{-1,1\}^m \to \{-1,1\}$ be a function with $\widetilde{\mathrm{odeg}}_{1/2}(f) > d$. Let $F : \{-1,1\}^{mt} \to \{-1,1\}$ denote the function $\mathrm{OR}_t(f,\ldots,f)$. Then $\widetilde{\deg}_{1-2^{-t}}(F) > d$.*

We remark that it is necessary that the *one-sided* approximated degree of f is large, rather than that just the approximate degree of f is large. Theorem 1 is easily seen to be false with one-sided approximate degree replaced by approximate degree. Consider for example the case where $f = \mathrm{OR}_m$. Then $F = \mathrm{OR}_t(\mathrm{OR}_m,\ldots,\mathrm{OR}_m) = \mathrm{OR}_{mt}$. It is well-known that $\widetilde{\deg}(\mathrm{OR}_m) = \Omega(\sqrt{m})$, so Theorem 1 with $\widetilde{\deg}$ in place of $\widetilde{\mathrm{odeg}}$ would say that $\widetilde{\deg}_{1-2^{-t}}(\mathrm{OR}_{mt}) = \Omega(\sqrt{mt})$. Yet the polynomial $q(y) = \frac{1}{mt}(1/2 - \sum_{i=1}^{t}\sum_{j=1}^{m} y_{ij})$ demonstrates that $\widetilde{\deg}_{1-\frac{1}{2mt}}(\mathrm{OR}_{mt}) = 1$ for all values of t. However, Theorem 1 does not apply because the one-sided approximate degree of $f = \mathrm{OR}_m$ is constant.

Proof. Let ψ be a dual polynomial for f with one-sided error whose existence is guaranteed by the assumption that $\widetilde{\mathrm{odeg}}_{1/2}(f) > d$. By Theorem 13, ψ satisfies:

$$\sum_{x\in\{-1,1\}^m} \psi(x)f(x) > 1/2, \tag{16}$$

$$\sum_{x\in\{-1,1\}^m} |\psi(x)| = 1, \tag{17}$$

$$\sum_{x\in\{-1,1\}^m} \psi(x)\chi_S(x) = 0 \text{ for each } |S| \le d \text{ and} \tag{18}$$

$$\psi(x) < 0 \text{ for each } x \in f^{-1}(-1). \tag{19}$$

We will construct a dual solution ζ that witnesses the fact that $\widetilde{\deg}_{1-2^{-t}}(F) > d$. Specifically, ζ must satisfy the three conditions of Theorem 12:

$$\sum_{(x_1,\ldots,x_t)\in(\{-1,1\}^m)^t} \zeta(x_1,\ldots,x_t)F(x_1,\ldots,x_t) > 1 - 2^{-t}. \tag{20}$$

$$\sum_{(x_1,\ldots,x_t)\in(\{-1,1\}^m)^t} |\zeta(x_1,\ldots,x_t)| = 1. \tag{21}$$

$$\sum_{(x_1,\ldots,x_t)\in(\{-1,1\}^m)^t} \zeta(x_1,\ldots,x_t)\chi_S(x_1,\ldots,x_t) = 0 \text{ for each } |S| \le d. \tag{22}$$

Let $\mathbf{1}$ denote the all-ones vector. Let $\Psi : \{-1,1\}^t \to \{-1,1\}$ be defined such that $\Psi(\mathbf{1}) = 1/2$, $\Psi(-\mathbf{1}) = -1/2$, and $\Psi(x) = 0$ for all other x. Notice that

$$\sum_{(x_1,\ldots,x_t)\in(\{-1,1\}^m)^t} \Psi(x_1,\ldots,x_t) = 0 \tag{23}$$

We define $\zeta : (\{-1,1\}^m)^t \to \mathbb{R}$ by

$$\zeta(x_1,\ldots,x_t) := 2^t \Psi(\ldots,\widetilde{\mathrm{sgn}}(\psi(x_i)),\ldots) \prod_{i=1}^{t} |\psi(x_i)|, \tag{24}$$

where $x_i = (x_{i,1},\ldots,x_{i,m})$.

Eq. (24) combines dual functions Ψ and ψ to obtain a dual witness ζ in exactly the same manner as in the works of Sherstov [46, Theorem 3.3] and Lee [28]. The analysis in these works implies without modification that ζ satisfies Equations Eq. (21) and Eq. (22). We provide this analysis in Appendix B for completeness, and here focus on arguing that (20) holds. As we remarked earlier, the properties we exploit to show this are (1) that ψ has one-sided error and (2) that the the vector $-\mathbf{1}$ has Hamming distance t from the (unique) input in $\mathrm{OR}_t^{-1}(1)$.

We now prove that (20) holds. Let μ be the distribution on $(\{-1,1\}^m)^t$ given by $\mu(x_1,\ldots,x_t) = \prod_{i=1}^t |\psi(x_i)|$. Since ψ is orthogonal to the constant polynomial, it has expected value 0, and hence the string $(\ldots,\widetilde{\mathrm{sgn}}(\psi(x_i)),\ldots)$ is distributed uniformly in $\{-1,1\}^t$ when one samples $(x_1,\ldots,x_t)$ according to μ. Observe that

$$\begin{aligned}
&\sum_{(x_1,\ldots,x_t)\in(\{-1,1\}^m)^t} \zeta(x_1,\ldots,x_t) F(x_1,\ldots,x_t) \\
&= 2^t \mathbf{E}_\mu[\Psi(\ldots,\widetilde{\mathrm{sgn}}(\psi(x_i)),\ldots)\,\mathrm{OR}_t(\ldots,f(x_i),\ldots)] \\
&= \sum_{z\in\{-1,1\}^t} \Psi(z) \left(\sum_{(x_1,\ldots,x_t)\in(\{-1,1\}^m)^t} \mathrm{OR}_t(\ldots,f(x_i),\ldots)\,\mu(x_1,\ldots,x_t|z) \right),
\end{aligned} \tag{25}$$

where $\mu(\mathbf{x}|z)$ denotes the probability of $\mathbf{x}$ under μ, conditioned on $(\ldots,\widetilde{\mathrm{sgn}}(\psi(x_i)),\ldots) = z$.

Let $A_1 = \{x \in \{-1,1\}^m : \psi(x) \geq 0, f(x) = -1\}$ and $A_{-1} = \{x \in \{-1,1\}^m : \psi(x) < 0, f(x) = 1\}$, so $A_1 \cup A_{-1}$ is the set of all inputs x where the sign of $\psi(x)$ disagrees with $f(x)$. Notice that $\sum_{x\in A_1\cup A_{-1}} |\psi(x)| < 1/4$ because ψ has correlation $1/2$ with f.

As noted in [46], for any given $z \in \{-1,1\}^t$, the following two random variables are identically distributed:

- The string $(\ldots,f(x_i),\ldots)$ when one chooses $(\ldots,x_i,\ldots)$ from the conditional distribution $\mu(\cdot|z)$.
- The string $(\ldots,y_i z_i,\ldots)$, where $y \in \{-1,1\}^t$ is a random string whose ith bit independently takes on value -1 with probability $2\sum_{x\in A_{z_i}} |\psi(x)| < 1/2$.

Thus, Expression (25) equals

$$\sum_{z\in\{-1,1\}^t} \Psi(z) \cdot \mathbf{E}[\mathrm{OR}_t(\ldots,y_i z_i,\ldots)], \tag{26}$$

where $y \in \{-1,1\}^t$ is a random string whose ith bit independently takes on value -1 with probability $2\sum_{x \in A_{z_i}} |\psi(x)| < 1/2$. We first argue that the term corresponding to $z = \mathbf{1}$ contributes $\Psi(z) = 1/2$ to Expression (26). By Eq. (19), if $f(x) = -1$, then $\widetilde{\mathrm{sgn}}(\psi(x)) = -1$. This implies that A_1 is empty; that is, if $\widetilde{\mathrm{sgn}}(\psi(x)) = 1$, then it must be the case that $f(x) = 1$. Therefore, for $z = \mathbf{1}$, the y_i's are all 1 with probability 1, and hence $\mathbf{E}_y[\mathrm{OR}_t(\dots, y_i z_i, \dots)] = \mathrm{OR}_t(\mathbf{1}) = 1$. Thus the term corresponding to $z = \mathbf{1}$ contributes $\Psi(z)\,\mathrm{OR}_t(z) = 1/2$ to Expression (26) as claimed.

All $z \notin \{\mathbf{1}, -\mathbf{1}\}$ are given zero weight by Ψ and hence contribute nothing to the sum. All that remains is to show that the contribution of the term $z = -\mathbf{1}$ to the sum is $\frac{1}{2}(1 - 2^{-t})$. Since each $y_i = 1$ independently with probability at least $1/2$, and $\mathrm{OR}_t(\dots, -y_i, \dots) = 1$ as long as there is at least one $y_i \neq -1$, we conclude that $\mathbf{E}[\mathrm{OR}_t(\dots, y_i z_i, \dots)] \geq 1 - 2^{-t+1}$. It follows that the term corresponding to $z = -\mathbf{1}$ contributes at least $\frac{1}{2}(1 - 2^{-t+1})$ to the sum. Thus,

$$\sum_{z \in \{-1,1\}^t} \Psi(z) \cdot \mathbf{E}[\mathrm{OR}_t(\dots, y_i z_i, \dots)] \geq \frac{1}{2} + \frac{1}{2}(1 - 2^{-t+1}) = 1 - 2^{-t}.$$

This completes the proof. □

Remark: Since the set A_1 within the proof of Theorem 1 is empty, the "combined" dual witness ζ constructed in the proof in fact has one-sided error. Thus, the proof establishes that $\widetilde{\mathrm{odeg}}_{1-2^{-t}}(F) > d$, which is a stronger conclusion than the $\widetilde{\deg}_{1-2^{-t}}(F) > d$ bound appearing in the theorem statement. We chose to state Theorem 1 as an approximate degree lower bound, rather than as a one-sided approximate degree lower bound, for easier comparison with prior work on approximate degree.

We are now in a position to prove our new lower bound on "accuracy vs. degree" tradeoffs for pointwise approximating AC^0 functions by polynomials.

Corollary 2. *For every $d > 0$, there is a depth-3 Boolean circuit $F : \{-1,1\}^n \to \{-1,1\}$ of size $\mathrm{poly}(n)$ such that any degree-d polynomial cannot pointwise approximate F to error better than $1 - \exp\left(-\tilde{\Omega}(nd^{-3/2})\right)$. In particular, any polynomial of degree at most $n^{2/5}$ cannot pointwise approximate F to error better than $1 - \exp\left(-\tilde{\Omega}(n^{2/5})\right)$.*

Proof. Let $t = n/d^{3/2}$, and $m = d^{3/2}$. Define $F = \mathrm{OR}_t(f, \dots, f)$ where $f : \{-1,1\}^m \to \{-1,1\}$ computes the Element Distinctness problem. The discussion in Section 3.2 implies that f is computed by a depth-2 circuit, and that f has one-sided approximate degree $\tilde{\Omega}(m^{2/3})$. The claim now follows by Theorem 1. □

4.2 On the Tightness of Theorem 1 and Corollary 2

We now argue that the approximate degree lower bound proved in Theorem 1 is essentially tight. In particular, we show that the function F for which Corollary 2 yields a $(1 - \exp(-\tilde{\Omega}(n^{2/5}))$-error lower bound for approximating polynomials of degree $n^{2/5}$ actually admits a $(1 - \exp(-\tilde{O}(n^{2/5}))$-approximating polynomial of degree $\tilde{O}(n^{2/5})$.

Our nearly-matching upper bound makes use of a well-known paradigm for constructing low-weight PTFs (and hence, by Lemma 17, low-accuracy pointwise approximations) for composed

functions by way of *rational approximations* (see e.g. [46]). Suppose $f : \{-1,1\}^m \to \{-1,1\}$ is pointwise approximated by a rational function in the sense that for every $x \in \{-1,1\}^m$,

$$\left| f(x) - \frac{p(x)}{q(x)} \right| < \frac{1}{t},$$

where p, q are polynomials of degree d and weight w and $q(x) > 0$ on $\{-1,1\}^m$. Then observe that the block composition

$$\mathrm{OR}_t(f(x_1), \ldots, f(x_t)) = \mathrm{sgn}(1 - t + f(x_1) + \cdots + f(x_t)) = \mathrm{sgn}\left(t - t + \frac{p(x_1)}{q(x_1)} + \ldots \frac{p(x_t)}{q(x_t)}\right).$$

Multiplying $\left(1 - t + \frac{p(x_1)}{q(x_1)} + \ldots \frac{p(x_t)}{q(x_t)}\right)$ by the positive quantity $q(x_1) \cdot \cdots \cdot q(x_t)$ and clearing denominators yields a PTF for the composed function of degree td and weight at most $w^t(m + tw)$.

We now construct a rational approximation for $f = $ Element Distinctness with the desired properties. Recall from Section 3.2 that Element Distinctness on m variables has a CNF representation where the top AND gate has fan-in $s := O(m^3)$ and each OR gate has fan-in $O(\log m)$. It is easy to check that $\mathrm{AND}_s : \{-1,1\}^s \to \{-1,1\}$ admits the rational approximation

$$\frac{ts - 1 + t\sum_{i=1}^s x_i}{ts + 1 + t\sum_{i=1}^s x_i}$$

with error $1/t$, degree $d = 1$, and weight $w = O(st)$. Moreover, each bottom OR gate in the CNF can be computed exactly by a degree $O(\log m)$ polynomial with weight 1. Composing these constructions yields a rational approximation for Element Distinctness with error $1/t$, degree $d = O(\log m) = O(\log t)$ and weight $O(st) = \mathrm{poly}(t)$. Therefore, F has a PTF of degree $\tilde{O}(t)$ and weight $\exp(\tilde{O}(t))$. By the construction of Lemma 17, F also has a $(1 - \exp(-\tilde{O}(t)))$-approximation of degree $\tilde{O}(t)$. Taking $t = n^{2/5}$ gives the desired result.

4.3 A Sharp Threshold in Accuracy-Degree Tradeoffs

The rational approximations developed in the previous section, combined with the lower bound of Theorem 1 and Corollary 2, reveal a "sharp threshold" in the degree required to approximate a particular function F within a given error parameter. Recall that Theorem 1 and Corollary 2 yield a lower bound of $d = \Omega(m^{2/3}/\log m)$ on the ε-approximate degree of $F = \mathrm{OR}_t(f, \ldots, f)$, where f is the Element Distinctness function on m variables and $\varepsilon = 1 - 2^{-t}$. In the following discussion, consider any $t = d^{1-\Omega(1)}$.

If our goal is to approximate F to within error $(1 - \exp(-\tilde{O}(t)))$, then the rational approximation techniques described in the preceding section yield an approximating polynomial of degree $\tilde{O}(t)$. On the other hand, if we desire even slightly better error of $1 - 2^{-t}$, then our accuracy-degree tradeoff lower bound of Theorem 1 shows that we require degree $d = \omega(t)$. That is, if we demand error that is slightly better than $1 - \exp(-\tilde{O}(t))$, there is an asymptotic jump from $\tilde{O}(t)$ to $\Omega(d)$ in the required degree.

5 Discrepancy of AC^0

In this section we prove our new exponentially small upper bound on the discrepancy of a function in AC^0. Consider a Boolean function $f : X \times Y \to \{-1,1\}$, and let $M^{(f)}$ be its communication

matrix $M^{(f)} = [f(x,y)]_{x \in X, y \in Y}$. A combinatorial rectangle of $X \times Y$ is a set of the form $A \times B$ with $A \subseteq X$ and $B \subseteq Y$. For a distribution μ over $X \times Y$, the discrepancy of f with respect to μ is defined to be the maximum over all rectangles R of the *bias* of f on R. That is:

$$\mathrm{disc}_\mu(f) = \max_R \left| \sum_{(x,y) \in R} \mu(x,y) f(x,y) \right|.$$

The discrepancy of f, $\mathrm{disc}(f)$ is defined to be $\min_\mu \mathrm{disc}_\mu(f)$.

Sherstov's pattern matrix method [42] shows how to generically transform an AC^0 function with high threshold degree or high threshold weight into another AC^0 function with low discrepancy.

Theorem 18 ([42], adapted from Corollary 1.2 and Theorem 7.3)**.** *Let $F : \{-1,1\}^n \to \{-1,1\}$ be given, and define the communication problem $F' : \{-1,1\}^{4n} \times \{-1,1\}^{4n} \to \{-1,1\}$ by*

$$F'(x,y) = F(\dots, \vee_{j=1}^4 (x_{i,j} \wedge y_{i,j}), \dots).$$

Then for every integer $d \geq 0$,

$$\mathrm{disc}(F')^2 \leq \max\left\{ \frac{2n}{W(F, d-1)}, 2^{-d} \right\}.$$

We apply this theorem to the function $F : \{-1,1\}^n \to \{-1,1\}$ of Corollary 2. This function has ε-approximate degree $n^{2/5}$ for $\varepsilon = 1 - 2^{-\tilde{\Omega}(n^{2/5})}$, and hence by by Lemma 17 it holds that $W(f, n^{2/5}) = 2^{\tilde{\Omega}(n^{2/5})}$. We thus obtain our new discrepancy upper bound for AC^0 as stated in Corollary 3, restated here for the reader's convenience.

Corollary 3. *There is a depth-4 Boolean circuit $F' : \{-1,1\}^n \to \{-1,1\}$ with discrepancy $\exp\left(-\tilde{\Omega}(n^{2/5})\right)$.*

6 Threshold Weight of AC^0

Combing Lemma 17 with Corollary 2 yields Corollary 4, restated here for the reader's convenience.

Corollary 4. *For every $d > 0$, there is a depth-3 Boolean circuit $F : \{-1,1\}^n \to \{-1,1\}$ of size $\mathrm{poly}(n)$ such that $W(F,d) \geq \exp\left(\tilde{\Omega}(nd^{-3/2})\right)$. In particular, $W(F, n^{2/5}) = \exp\left(\tilde{\Omega}(n^{2/5})\right)$.*

A result of Krause [26] allows us to extend our new degree-d threshold weight lower bound for AC^0 into an $\exp\left(\tilde{\Omega}\left(n^{2/5}\right)\right)$ *degree independent* threshold weight lower bound for a related function F'. We give a slight modification (Lemma 19) that is cleaner to apply, and asymptotically recovers Krause's result when the weights under consideration are superpolynomially large. Our restatement admits a new and simple proof based on LP duality that we present in Appendix C.

Lemma 19. *Let $F : \{-1,1\}^n \to \{-1,1\}$ be a Boolean function, and define $F' : \{-1,1\}^{3n} \to \{-1,1\}$ by*

$$F'(x_1, \dots, x_n, y_1, \dots, y_n, z_1, \dots, z_n) := F(\dots, (\bar{z}_i \wedge x_i) \vee (z_i \wedge y_i), \dots).$$

Then

$$W(F')^2 \geq \min\left\{ \frac{W(F,d)}{2n}, 2^d \right\}.$$

Combining Corollary 4 and Lemma 19 yields Corollary 5. This improves over the previous best threshold weight lower bound for AC^0, which was $\exp\left(\Omega(n^{1/3})\right)$ [27].

Corollary 5. *There is a depth-4 Boolean circuit $F' : \{-1,1\}^n \to \{-1,1\}$ satisfying $W(F') = \exp\left(\tilde{\Omega}(n^{2/5})\right)$.*

Proof. Let F be the circuit of Corollary 2 and let F' be the depth-four circuit obtained by applying Lemma 19 to F. Let $d = n^{2/5}/\log^c n$ for a sufficiently large constant c. Then Corollary 4 implies that $W(F,d) \geq 2n2^d$, and hence $W(F') \geq 2^{d/2} = 2^{\tilde{\Omega}(n^{2/5})}$ by Lemma 19. □

Remark: While the threshold weight bound of Corollary 5 is stated for polynomial threshold functions over $\{-1,1\}^n$ (i.e., for polynomials that are integer linear combinations of parities), the same threshold weight lower bound also holds for polynomials over $\{0,1\}^n$, or equivalently, for integer linear combinations of conjunctions. This can be seen as follows.

Given a set $S \subseteq [n]$, let $\mathrm{AND}_S : \{-1,1\}^n \to \{-1,1\}$ denote the AND function restricted to variables in S. Given a sign-representation $p = \sum_S c_S \mathrm{AND}_S$ for F of weight w, let $\sum_S \hat{p}(S)\chi_S$ denote the Fourier representation of p. It is easy to check that the L_1-norm of the Fourier coefficients of each conjunction AND_S is at most 3, so the weight of the Fourier expansion of p is $w' := \sum_S |\hat{p}(S)| \leq 3w$. However, we cannot simply conclude that $w/3 \geq w' \geq W(f)$ because the coefficients $\hat{p}(S)$ are not necessarily integers.

Nonetheless, note that $|p(x)| \geq 1$ for all $x \in \{-1,1\}^n$, since p has integer coefficients. That is, p is a sign-representation for f over $\{-1,1\}^n$ of weight w' and with margin at least 1. It follows by Theorem 16 that $\exp\left(\tilde{\Omega}(n^{2/5})\right) = W(f) \leq 2n(w')^2 = \mathrm{poly}(n,w)$. We conclude that $w = \exp\left(\tilde{\Omega}(n^{2/5})\right)$ as desired.

The same argument shows that all of our lower bounds on degree-d threshold weight proved in this paper hold for PTFs over $\{0,1\}^n$, in addition to PTFs over $\{-1,1\}^n$.

7 Lower Bounds for Read-Once DNFs

In this section we derive new approximate degree and degree-d threshold weight lower bounds for read-once DNF formulas. The lower bounds we prove are essentially identical to those proved by Beigel [8] and Servedio et al. [37] for the *decision list* ODD-MAX-BIT, which is not computable by a read-once DNF. Our first construction (Corollary 7) yields a degree-d threshold weight lower bound of $2^{\Omega(\sqrt{n/d})}$, matching the lower bound proved by Servedio et al. for the decision list ODD-MAX-BIT. In Section 7.3, we show that this is essentially optimal in the "high-degree" regime where $d = \Omega(n^{1/3})$.

Our second lower bound (Corollary 8) exhibits a DNF with $(1 - 2^{-n/d^2})$-approximate degree $\Omega(d)$, matching Beigel's lower bound for ODD-MAX-BIT. As we remarked in the introduction, for $d < n^{1/3}$, Corollary 8 is subsumed by Minsky and Papert's seminal result exhibiting a read-once DNF F with threshold degree $\Omega(n^{1/3})$. However, for $d > n^{1/3}$, it is not subsumed by Minsky and Papert's result, nor by Corollary 7. While Corollary 7 yields a lower bound on the degree-d threshold weight of read-once DNFs, it does not yield a lower bound on the *approximate-degree* of read-once DNFs. As described in Section 3.3, while $\widetilde{\deg}_{1-\frac{1}{w}}(F) > d$ implies that $W(F,d) > w$, the reverse implication does *not* hold when $w \ll \binom{n}{d}$ (and in fact the read-once DNF considered in Corollary 7 is an explicit example of the reverse implication failing badly).

7.1 Extending the Lower Bound of Servedio et al. to Read-Once DNFs

7.1.1 Hardness Amplification for Approximate Weight

We now extend our hardness amplification techniques from approximate degree to approximate weight. This extension forms the technical heart of our proof that the lower bound of Servedio et al. applies to read-once DNFs.

Theorem 6. *Let $f : \{-1,1\}^m \to \{-1,1\}$ be a function with one-sided non-constant approximate weight $W^*_{3/4}(f,d) > w$. Let $F : \{-1,1\}^{mt} \to \{-1,1\}$ denote the function $\mathrm{OR}_t(f,\dots,f)$. Then F has degree-d $(1-2^{-t})$-approximate weight $W_{1-2^{-t}}(F,d) > 2^{-5t}w$.*

Proof. Let ψ be a dual polynomial for f with one-sided error whose existence is guaranteed by the assumption that $W^*_{3/4}(f,d) > w$. Then by Theorem 15, ψ satisfies:

$$\sum_{x\in\{-1,1\}^m} \psi(x)f(x) - \frac{3}{4}\sum_{x\in\{-1,1\}^m} |\psi(x)| > w, \tag{27}$$

$$\left|\sum_{x\in\{-1,1\}^m} \psi(x)\chi_S(x)\right| \leq 1 \text{ for each } 0 < |S| \leq d, \tag{28}$$

$$\sum_{x\in\{-1,1\}^m} \psi(x) = 0, \text{ and} \tag{29}$$

$$\psi(x) < 0 \text{ for each } x \in f^{-1}(-1). \tag{30}$$

We will construct a dual solution ζ that witnesses the fact that $W_{1-2^{-t}}(F,d) > 2^{-5t}w$. Specifically, by Theorem 14, ζ must satisfy the following conditions:

$$\sum_{(x_1,\dots,x_t)\in(\{-1,1\}^m)^t} \zeta(x_1,\dots,x_t)F(x_1,\dots,x_t) - (1-2^{-t})|\zeta(x_1,\dots,x_t)| > 2^{-5t}w. \tag{31}$$

$$\left|\sum_{(x_1,\dots,x_t)\in(\{-1,1\}^m)^t} \zeta(x_1,\dots,x_t)\chi_S(x_1,\dots,x_t)\right| \leq 1 \text{ for each } |S| \leq d. \tag{32}$$

As before, let $\Psi : \{-1,1\}^t \to \{-1,1\}$ be defined such that $\Psi(\mathbf{1}) = 1/2$, $\Psi(-\mathbf{1}) = -1/2$, and $\Psi(x) = 0$ for all other x, where $\mathbf{1}$ denotes the all-ones vector. We define $\zeta : (\{-1,1\}^m)^t \to \mathbb{R}$ by

$$\zeta(x_1,\dots,x_t) := M_t\Psi(\dots,\widetilde{\mathrm{sgn}}(\psi(x_i)),\dots)\prod_{i=1}^{t}|\psi(x_i)|, \tag{33}$$

where $x_i = (x_{i,1},\dots,x_{i,m})$ and M_t is a normalization term to be determined later.

We start with Eq. (32) to determine an appropriate choice of M_t. Notice that since Ψ is orthogonal on $\{-1,1\}^t$ to constant functions, its expected value is 0. Thus, we may write the Fourier representation for Ψ as

$$\Psi(z) = \sum_{\substack{T\subseteq\{1,\dots,t\}\\ T\neq\emptyset}} \hat{\Psi}(T)\chi_T(z)$$

for some real numbers $\hat{\Psi}(T)$. We can thus write

$$\zeta(x_1, \ldots, x_t) = M_t \sum_{T \neq \emptyset} \hat{\Psi}(T) \prod_{i \in T} \psi(x_i) \prod_{i \notin T} |\psi(x_i)|.$$

Given a subset $S \subseteq \{1, \ldots, t\} \times \{1, \ldots, m\}$ with $|S| \leq d$, partition $S = (\{1\} \times S_1) \cup \cdots \cup (\{t\} \times S_t)$ where each $S_i \subseteq \{1, \ldots, m\}$. Then

$$\sum_{(x_1,\ldots,x_t) \in (\{-1,1\}^m)^t} \zeta(x_1, \ldots, x_t) \chi_S(x_1, \ldots, x_t)$$

$$= M_t \sum_{T \neq \emptyset} \hat{\Psi}(T) \prod_{i \in T} \underbrace{\left(\sum_{x_i \in \{-1,1\}^m} \psi(x_i) \chi_{S_i}(x_i) \right)}_{} \prod_{i \notin T} \left(\sum_{x_i \in \{-1,1\}^m} |\psi(x_i)| \chi_{S_i}(x_i) \right).$$

Since $|S| \leq d$, we have that $|S_i| \leq d$ for every index $i \in \{1, \ldots, t\}$. For each set T, each of the underbraced factors is bounded in absolute value by 1 by (28). Writing

$$\|\psi\|_1 := \sum_{x \in \{-1,1\}^m} |\psi(x)|$$

for notational convenience, we see that

$$\left| \sum_{(x_1,\ldots,x_t) \in (\{-1,1\}^m)^t} \zeta(x_1, \ldots, x_t) \chi_S(x_1, \ldots, x_t) \right| \leq M_t \sum_{T \neq \emptyset} \hat{\Psi}(T) \|\psi\|_1^{t-|T|} \leq M_t \cdot t 2^{t-1} \|\psi\|_1^{t-1}.$$

Taking $M_t = 2^{-2t} \|\psi\|_1^{1-t}$ gives (32).

We now proceed to verify (31). Let μ be the distribution on $(\{-1,1\}^m)^t$ given by $\mu(x_1, \ldots, x_t) = \|\psi\|_1^{-t} \prod_{i=1}^t |\psi(x_i)|$. Since ψ is orthogonal to the constant polynomial, it has expected value 0, and hence the string $(\ldots, \widetilde{\mathrm{sgn}}(\psi(x_i)), \ldots)$ is distributed uniformly in $\{-1,1\}^t$ when one samples $(x_1, \ldots, x_t)$ according to μ. Observe that

$$\sum_{(x_1,\ldots,x_t) \in (\{-1,1\}^m)^t} \zeta(x_1, \ldots, x_t) F(x_1, \ldots, x_t)$$

$$= M_t \|\psi\|_1^t \mathbf{E}_\mu [\Psi(\ldots, \widetilde{\mathrm{sgn}}(\psi(x_i)), \ldots) \, \mathrm{OR}_t(\ldots, f(x_i), \ldots)]$$

$$= 2^{-3t} \|\psi\|_1 \sum_{z \in \{-1,1\}^t} \Psi(z) \left(\sum_{(x_1,\ldots,x_t) \in (\{-1,1\}^m)^t} \mathrm{OR}_t(\ldots, f(x_i), \ldots) \, \mu(x_1, \ldots, x_t | z) \right), \tag{34}$$

where $\mu(\mathbf{x}|z)$ denotes the probability of $\mathbf{x}$ under μ, conditioned on $(\ldots, \widetilde{\mathrm{sgn}}(\psi(x_i)), \ldots) = z$.

Let $A_1 = \{x \in \{-1,1\}^m : \psi(x) \geq 0, f(x) = -1\}$ and $A_{-1} = \{x \in \{-1,1\}^m : \psi(x) < 0, f(x) = 1\}$. Then $2 \sum_{x \in A_1 \cup A_{-1}} |\psi(x)| < \frac{1}{4} \|\psi\|_1 - w$ because ψ has correlation at least $w + \frac{3}{4} \|\psi\|_1$ with f.

As before, for any $z \in \{-1,1\}^t$, the following two random variables are identically distributed:

- The string $(\ldots, f(x_i), \ldots)$ when one chooses $(\ldots, x_i, \ldots)$ from the conditional distribution $\mu(\cdot|z)$.

- The string $(\dots, y_i z_i, \dots)$, where $y \in \{-1,1\}^t$ is a random string whose ith bit independently takes on value -1 with probability $\frac{2}{\|\psi\|_1}\sum_{x\in A_{z_i}}|\psi(x)| < 1/4 - w/\|\psi\|_1$.

Thus, the correlation is

$$2^{-3t}\|\psi\|_1 \sum_{z\in\{-1,1\}^t} \Psi(z)\cdot \mathbf{E}[\mathrm{OR}_t(\dots, y_i z_i, \dots)], \tag{35}$$

where $y \in \{-1,1\}^t$ is a random string whose ith bit independently takes on value -1 with probability $2\sum_{x\in A_{z_i}}|\psi(x)| < 1/4 - w/\|\psi\|_1$. As in the proof of Theorem 1, the one-sided error (30) of the dual witness ψ implies that the input $z = \mathbf{1}$ contributes $\Psi(z) = 1/2$ to Expression (35). All $z \notin \{\mathbf{1}, -\mathbf{1}\}$ are given zero weight by Ψ and hence contribute nothing to the sum. All that remains is to show that the contribution of the term $z = -\mathbf{1}$ to the sum is $\frac{1}{2}(1-2^{-2t+1})$. Since each $y_i = 1$ independently with probability at least $3/4 + w/\|\psi\|_1$, and $\mathrm{OR}_t(\dots, -y_i, \dots) = 1$ as long as there is at least one $y_i \neq -1$, we conclude that $\mathbf{E}[\mathrm{OR}_t(\dots, y_i z_i, \dots)] \geq 1 - 2^{-2t+1}$. It follows that the term corresponding to $z = -\mathbf{1}$ contributes at least $\frac{1}{2}(1 - 2^{-2t+1})$ to the sum. Thus,

$$2^{-3t}\|\psi\|_1 \sum_{z\in\{-1,1\}^t} \Psi(z)\cdot\mathbf{E}[\mathrm{OR}_t(\dots, y_i z_i, \dots)] \geq 2^{-3t}\|\psi\|_1 \left(\frac{1}{2} + \frac{1}{2}(1 - 2^{-2t+1})\right) = 2^{-3t}(1-2^{-2t})\|\psi\|_1.$$

Since ψ is orthogonal to the constant polynomial by Eq. (29), it has expected value 0, and hence the string $(\dots, \widetilde{\mathrm{sgn}}(\psi(x_i)), \dots)$ is distributed uniformly in $\{-1,1\}^t$ when one samples $(x_1, \dots, x_t)$ according to μ. Thus,

$$\sum_{(x_1,\dots,x_t)\in(\{-1,1\}^m)^t} |\zeta(x_1, \dots, x_t)| = 2^{-3t}\|\psi\|_1 \sum_{z\in\{-1,1\}^t} |\Psi(z)| = 2^{-3t}\|\psi\|_1,$$

.

Now the left-hand side of Expression (31) is at least

$$2^{-3t}(1 - 2^{-2t})\|\psi\|_1 - (1 - 2^{-t})\cdot 2^{-3t}\|\psi\|_1 > 2^{-5t}\|\psi\|_1 > 2^{-5t}w,$$

where the last inequality follows from condition (27). This completes the proof.

□

7.1.2 Completing the Proof of Corollary 7

We adapt an argument of Servedio et al. to prove the following one-sided approximate weight lower bound for the function AND_n.

Lemma 20. *Let $d = o(n/\log^2 n)$. Then the function AND_n has one-sided non-constant approximate weight $W^*_{3/4}(\mathrm{AND}_n, d) = 2^{\Omega(n/d)}$.*

Our proof of Lemma 20 follows a symmetrization argument due to Servedio et al. [37]. The key in their proof is the following Markov-type inequality that gives a sharp bound on the derivative of a bounded polynomial in terms of both its degree and weight.

Lemma 21 ([37], Lemma 1)**.** *Let $P : \mathbb{R} \to \mathbb{R}$ be a degree-d polynomial such that*

1. *The coefficients of P each have absolute value at most w, and*

2. $1/2 \leq \max_{x\in[-1,1]} |p(x)| \leq R$.

Then $\max_{x\in[-1,1]} |p'(x)| = O(d \cdot R \cdot \max\{\log W, \log d\})$.

Proof of Lemma 20. Let $p : \mathbb{R}^n \to \mathbb{R}$ be a real polynomial with degree d and non-constant weight w that has one-sided distance at most $3/4$ from AND_n. Specifically, $p(-\mathbf{1}) \leq -1/4$ and $1/4 \leq p(x) \leq 7/4$ at all other Boolean inputs. We will show that $w = 2^{\Omega(n/d)}$. First observe that if $p(-\mathbf{1}) \leq -7/4$, then the polynomial

$$q(x) = \frac{p(x) - 1}{|p(-\mathbf{1}) - 1|} + 1$$

is a true $(3/4)$-approximation to AND_n with weight smaller than $w + 1$, so we can assume without loss of generality that p is in fact a $(3/4)$-approximation to AND_n.

Define the univariate polynomial

$$P(t) := \mathbb{E}_{x\leftarrow\mu_t}[p(x)]$$

where μ_t is the product distribution over $\{-1, 1\}^n$ where each coordinate x_j is independently set to 1 with probability $(1 + t)/2$. Notice that $P(t)$ is obtained from the multivariate expansion of $p(x_1, \ldots, x_n)$ by replacing each variable x_i with t. It is readily verified that P satisfies the following properties.

1. $P(-1) = p(-\mathbf{1})$ and $P(1) = p(\mathbf{1})$,

2. $|P(t)| \leq \frac{7}{4}$ for all $t \in [-1, 1]$, and

3. $\deg P \leq \deg p = d$.

4. P has non-constant weight at most w.

By combining properties (1) and (4), we additionally see that the constant term $P(0)$ has absolute value at most $w + \frac{7}{4}$. We can then verify that P satisfies the conditions of Lemma 21.

1. The coefficients of P each have absolute value at most $w + \frac{7}{4}$ and

2. $1/2 \leq \max_{x\in[-1,1]} |P(t)| \leq \frac{7}{4}$.

Thus we conclude that $|P'(t)| = O(d \max\{\log w, \log d\})$ for $t \in [-1, 1]$. On the other hand, at $t_0 = -1 + 2/n$, we have $\Pr_{x\leftarrow\mu_{t_0}}[x = -1^n] = (1 - \frac{1}{n})^n < 1/e$, so $P(t_0) \geq 1 - \frac{2}{e}$. Since $P(-1) = p(-\mathbf{1}) \leq -\frac{1}{4}$, by the mean value theorem, there is some $t \in [-1, t_0]$ where $P'(t) \geq \frac{n}{4}$. Thus we have $d \max\{\log w, \log d\} = \Omega(n)$, and hence $w = 2^{\Omega(n/d)}$ as long as $d = o(n/\log^2 n)$. □

Finally, we are in a position to prove Corollary 7, restated here for the reader's convenience.

Corollary 7. *For each $d = o(n/\log^4 n)$, there is a read-once DNF F satisfying $W(F, d) = \exp\left(\Omega(\sqrt{n/d})\right)$.*

Proof. Set $m = \alpha\sqrt{nd}$ where α is a constant to be determined later, and let $t = n/m = \Omega(\sqrt{n/d})$. Let $F = \mathrm{OR}_t(\mathrm{AND}_m, \ldots, \mathrm{AND}_m)$. By Lemma 20, the inner function AND_m has degree-d one-sided non-constant approximate weight $W^*_{3/4}(\mathrm{AND}_m, d) = 2^{\beta m/d}$ for some constant β. Since $d = o(m/\log^2 m)$, by Theorem 6 the composed function F has degree-d approximate weight

$$W_{1-2^{-t}}(F, d) = 2^{-5t+\beta m/d} = 2^{(-5/\alpha+\beta)\sqrt{n/d}}.$$

Setting $\alpha > 5/\beta$, we get that this approximate weight is greater than 1. By Lemma 17, we have that $W(F, d) > 2^{-t} = 2^{\Omega(\sqrt{n/d})}$. □

7.2 Extending Beigel's Lower Bound to Read-Once DNFs

Corollary 8. *There is an (explicit) read-once DNF $F : \{-1,1\}^n \to \{-1,1\}$ with $\widetilde{\deg}_{1-2^{-n/d^2}}(F) = \Omega(d)$.*

Proof. Let $m = d^2$, $t = n/d^2$, and $f = \mathrm{AND}_m$. Then Theorem 1 guarantees that

$$\widetilde{\deg}_{1-2^{-t}}\left(\mathrm{OR}_t(\mathrm{AND}_m, \ldots, \mathrm{AND}_m)\right) > \widetilde{\mathrm{odeg}}(f).$$

The claim then follows from the fact that $\widetilde{\mathrm{odeg}}(\mathrm{AND}_m) = \Omega(\sqrt{m}) = \Omega(d)$, which can be seen by observing that Nisan and Szegedy's proof that $\widetilde{\deg}(\mathrm{AND}_m) = \Omega(\sqrt{m})$ in fact extends to one-sided approximate degree [33]. Alternatively, it can be directly shown that any dual witness (as defined in Theorem 12) for the fact that $\widetilde{\deg}(\mathrm{AND}_m) = \Omega(\sqrt{m})$ must have one-sided error (cf. [15, Theorem 5.1]). □

7.3 On the Tightness of Corollaries 7 and 8

In Section 4.2, we showed that Corollary 2 is essentially tight by exhibiting a nearly-matching upper bound based on rational approximations. A similar construction shows that any DNF of top fan-in t is computed by a PTF of degree $\tilde{O}(t)$ and weight $\exp\left(\tilde{O}(t)\right)$. This construction immediately shows that Corollary 7 is tight (up to logarithmic factors) for all $d > n^{1/3}$. Indeed, the DNF F for which Corollary 7 demonstrates $W(F, d) \geq \exp(\Omega(\sqrt{n/d}))$ has top fan-in $t = \sqrt{n/d}$, which is less than d for all $d > n^{1/3}$. This construction also reveals a sharp thresholding phenomenon for the read-once DNFs considered in Corollaries 7 and 8 that is similar to the one observed for the depth-three circuit considered in Section 4.3.

However, we can provide an alternative construction that demonstrates the tightness of both Corollaries 7 and 8. Specifically, rather than utilizing rational approximation techniques, we can construct a PTF for a read-once DNF by composing a PTF for the top OR gate with low-degree (polynomial, rather than rational) pointwise approximations to each of the individual terms. We provide this construction because of its power to explain why the lower bounds of Corollaries 7 and 8 take their particular forms.

Fix any function $f : \{-1,1\}^m \to \{-1,1\}$, and let $p : \{-1,1\}^m \to \{-1,1\}$ be a polynomial of degree d and weight w such that $|p(x) - f(x)| < 1/t$ for all $x \in \{-1,1\}^m$. Let $F(x_1, \ldots, x_t) = \mathrm{OR}_t(f(x_1), \ldots, f(x_t))$. Then for $(x_1, \ldots, x_t) \in \{-1,1\}^{m \cdot t}$, the identity $F(x_1, \ldots x_t) = \mathrm{sgn}(1 - t + \sum_{i=1}^t p(x_i))$ yields a PTF for F of degree at most d and weight at most $tw + t + 1$.

Recall that Corollary 7 yields a lower bound of $W(F,d) = \exp\left(\Omega(\sqrt{n/d})\right)$, where F is the read-once DNF with top fan-in roughly $t = \sqrt{n/d}$ and bottom fan-in roughly $m = \sqrt{nd}$. Servedio et al. [37] showed that for any $d > m^{1/2}$, there is a polynomial p of degree $\tilde{O}(d)$ and weight $\exp\left(\tilde{O}(m/d + \log t)\right) = \exp\left(\tilde{O}\left(\sqrt{n/d}\right)\right)$ that approximates the function AND_m to error $1/t^2$. Hence, as long as $d > n^{1/3}$, the polynomial $1 - t + \sum_{i=1}^{t} p(x_i)$ is a PTF for F of degree $\tilde{O}(d)$ and weight $\exp\left(\tilde{O}(\sqrt{n/d})\right)$, showing that Corollary 7 is tight up to logarithmic factors.

Similarly, recall that Corollary 8 yields a lower bound of $\widetilde{\deg}_{1-2^{n/d^2}}(F) = \Omega(d)$, where F is the read-once DNF with top fan $t = n/d^2$ and bottom fan-in $m = d^2$. It is well-known that a transformation of the Chebyshev polynomials yields a polynomial p of degree $\tilde{O}(m^{1/2})$ and weight $\exp\left(\tilde{O}(m^{1/2} + \log t)\right)$ that approximates AND_m to error better than $1/t^2$ (see e.g. [24]). Hence, $1 - t + \sum_{i=1}^{t} p(x_i)$ is a PTF for F of degree $\tilde{O}(d)$ and weight $\exp(\tilde{O}(d + \log t)) = \exp(\tilde{O}(n/d^2))$ when $d < n^{1/3}$. The transformation of Lemma 17 then shows that Corollary 8 is tight up to logarithmic factors in this parameter range.

8 Lower Bounds for AND-OR Trees

The d-level AND-OR tree (respectively, OR-AND tree) on n variables is a function described by a read-once circuit of depth d consisting of alternating layers of AND gates and OR gates, with the root gate being an AND gate (respectively, an OR gate). We assume throughout this section that all gates have fan-in $n^{1/d}$; for example, the two-level AND-OR tree is a read-once CNF in which all gates have fan-in $n^{1/2}$. The assumption on the fan-in is not essential to our analysis in this section, which in fact applies to any read-once Boolean circuit such that all gates at any given layer have the same fan-in. We will let $\text{AND-OR}_{d,n}$ (respectively, $\text{OR-AND}_{d,n}$) denote the d-level AND-OR tree (respectively, OR-AND tree) on n variables.

The current authors [12], and independently Sherstov [38], resolved the approximate degree of $\text{AND-OR}_{2,n}$ by proving an optimal $\Omega(n^{1/2})$ lower bound in this case. However, the techniques of [12, 38] break down for the case of depth three or greater; to the best of our knowledge, the best lower bound that follows from prior work is $\Omega(n^{1/4+1/2d})$, which can be derived by combining the depth-two lower bound [12, 38] with an earlier direct-sum theorem of Sherstov [46, Theorem 3.1].

In this section, we extend the methods of our prior work [12] to prove an $\Omega\left(n^{1/2}/\log^{(d-2)/2} n\right)$ lower bound on the approximate degree of $\text{AND-OR}_{d,n}$ for any constant $d > 0$. This matches an upper bound of Sherstov [40] up to a $\log^{(d-2)/2} n$ factor.

Theorem 9. *Let* $\text{AND-OR}_{d,n}$ *denote the* d*-level AND-OR tree on* n *variables. Then* $\widetilde{\deg}(\text{AND-OR}_{d,n}) = \Omega\left(n^{1/2}/\log^{(d-2)/2} n\right)$ *for any constant* $d > 0$.

8.1 Proof Outline.

To introduce our proof technique, we first describe the method used in [12] to construct an optimal dual polynomial in the case $d = 2$, and we identify why this method breaks down when trying to extend to the case $d = 3$. We then explain how to use our hardness amplification result (Theorem 1) to construct a different dual polynomial that does extend to the case $d = 3$.

Let $m = n^{1/2}$ denote the fan-in of all gates in OR-AND$_{2,n}$. In our earlier work [12], we constructed a dual polynomial for OR-AND$_{2,n}$ as follows.[1] We let γ_1 be a dual polynomial witnessing the fact that $\widetilde{\text{odeg}}(\text{AND}_m) = \Omega\left(m^{1/2}\right)$, and we let γ_2 be a dual polynomial witnessing the fact that $\widetilde{\text{deg}}(\text{OR}_m) = \Omega\left(m^{1/2}\right)$. We then combined the dual witnesses γ_1 and γ_2, using the same "combining" technique as in Eq. (24), to obtain a function $\gamma_3 : \{-1,1\}^{m^2} \to \mathbb{R}$ defined via:

$$\gamma_3(x_1, \ldots, x_m) := 2^m \gamma_2(\ldots, \widetilde{\text{sgn}}(\gamma_1(x_i)), \ldots) \prod_{i=1}^{m} |\gamma_1(x_i)|,$$

where $x_i = (x_{i,1}, \ldots, x_{i,m})$. It followed from earlier work [46] that γ_3 has pure high degree equal to the product of the pure high degree of γ_1 and the pure high degree of γ_2, yielding an $\Omega(m)$ lower bound on the pure high degree of γ_3. The new ingredient of the analysis in [12] was to use the one-sided error of the "inner" dual witness γ_1 to argue that γ_3 also had good correlation with OR-AND$_2$.

Extending to Depth Three. Let $M = n^{1/3}$ denote the fan-in of all gates in AND-OR$_{3,n}$. In constructing a dual witness for AND-OR$_{3,n}$ = AND$_M$(OR-AND$_{2,M^2}, \ldots,$ OR-AND$_{2,M^2}$), it is natural to try the following approach. Let γ_4 be a dual polynomial witnessing the fact that the approximate degree of AND$_M = \Omega(\sqrt{M})$. Then we can combine γ_3 and γ_4 in the same manner as above to obtain a dual function γ_5:

$$\gamma_5(x_1, \ldots, x_M) := 2^M \gamma_4(\ldots, \widetilde{\text{sgn}}(\gamma_3(x_i)), \ldots) \prod_{i=1}^{M} |\gamma_3(x_i)|, \tag{36}$$

where $x_i = (x_{i,1}, \ldots, x_{i,M^2})$. The difficulty in establishing that γ_5 is a dual witness to the high approximate degree of AND-OR$_{3,n}$ is in showing that γ_5 has good correlation with AND-OR$_3$. In our earlier work, we showed γ_3 has large correlation with OR-AND$_{2,n}$ by exploiting the fact that the inner dual witness γ_1 had one-sided error, i.e., $\gamma_1(y)$ agrees in sign with AND$_M$ whenever $y \in \text{AND}_M^{-1}(-1)$. However, γ_3 itself does not satisfy an analogous property: there are inputs $x_i \in \text{OR-AND}_{2,M^2}^{-1}(-1)$ such that $\gamma_3(x_i) > 0$, *and* there are inputs $x_i \in \text{OR-AND}_{2,M^2}^{-1}(1)$ such that $\gamma_3(x_i) < 0$.

To circumvent this issue, we use a different inner dual witness γ_3' within Eq. (36). Our construction of γ_3' will utilize our hardness amplification analysis to achieve the following: while γ_3' will have error "on both sides", the error from the "wrong side" will be very small. The hardness amplification step will cause γ_3' to have pure high degree that is lower than that of the dual witness γ_3 constructed in [12] by a $\sqrt{\log n}$ factor. However, the hardness amplification step will permit us to prove the desired lower bound on the correlation of γ_5 with AND-OR$_{3,n}$.

8.2 Proof of Theorem 9

Proof. We begin by proving the claimed lower bound for AND-OR$_{3,n}$ before explaining how to extend the argument to AND-OR$_{d,n}$ for an arbitrary depth $d > 0$.

[1] We actually constructed a dual polynomial for AND-OR$_{2,n}$, but the analysis for the case of OR-AND$_{2,n}$ is entirely analogous.

Notation. There will be a total of seven intermediate dual witnesses that arise in our construction of a dual witness ψ_7 for AND-OR$_{3,n}$. We will denote these seven dual witnesses as $\psi_1, \dots, \psi_7$. Let $M = n^{1/3}$ denote the fan-in of all gates in AND-OR$_{3,n}$. Our goal is to construct a dual witness ψ_7 to demonstrate that $\widetilde{\deg}(\text{AND-OR}_{3,n}) = \Omega\left(n^{1/2}/\log^{1/2} n\right)$.

To this end, define ψ_6 to be a dual polynomial witnessing the fact that $\widetilde{\text{odeg}}_{.99}(\text{AND}_M) = \Omega(\sqrt{M})$. By Theorem 12, there is some $d_6 = \Omega(\sqrt{M})$ such that ψ_6 satisfies:

$$\sum_{a \in \{-1,1\}^M} \psi_6(a)\,\text{AND}_M(a) > .99, \tag{37}$$

$$\sum_{a \in \{-1,1\}^M} |\psi_6(a)| = 1, \tag{38}$$

$$\sum_{a \in \{-1,1\}^M} \psi_6(a)\chi_S(a) = 0 \text{ for each } |S| \leq d_6 \text{ and} \tag{39}$$

$$\psi_6(-\mathbf{1}) < 0. \tag{40}$$

As stated in the proof outline, we are ultimately going to construct a function $\psi_5 : \{-1,1\}^{M^2} \to \mathbb{R}$ that serves as a dual witness to the high approximate degree of OR-AND$_{2,M^2}$ while having "almost no error on the wrong side". We will then define our final dual witness ψ_7 via

$$\psi_7(x_1, \dots, x_M) := 2^M \psi_6(\dots, \widetilde{\text{sgn}}(\psi_5(x_i)), \dots) \prod_{i=1}^{M} |\psi_5(x_i)|, \tag{41}$$

where $x_i = (x_{i,1}, \dots, x_{i,M^2})$.

Construction of ψ_5. Consider the function OR-AND$_{2,M^2}$. Let $t = 100 \log n$. We view the root OR gate as an OR of ORs, where the top OR has fan-in M/t and the bottom OR gates each have fan-in t. Thus, we are now thinking of the two-level OR-AND tree as a three-level circuit, where the top two levels consist of OR gates, and the bottom level consists of AND gates. Consider the function $F = \text{OR}_t(\text{AND}_M, \dots, \text{AND}_M)$. Corollary 8 constructs a dual witness ψ_3 demonstrating that there is some $d_3 = \Omega(\sqrt{M})$ such that $\widetilde{\text{odeg}}_{1-2^{-t}}(F) \geq d_2$ (see the Remark following the proof of Theorem 1). This dual witness ψ_3 was defined via:

$$\psi_3(b_1, \dots, b_t) := 2^t \psi_2(\dots, \widetilde{\text{sgn}}(\psi_1(b_i)), \dots) \prod_{i=1}^{M} |\psi_1(b_i)|,$$

where $b_i = (b_{i,1}, \dots, b_{i,M})$, ψ_1 was a dual witness to the high one-sided approximate degree of AND$_M$, and ψ_2 was defined such that $\psi_2(\mathbf{1}) = 1/2$, $\psi_2(-\mathbf{1}) = -1/2$, and ψ_2 evaluates to 0 for all other inputs in $\{-1,1\}^t$.

The proof of Theorem 1 showed that ψ_3 satisfies:

$$\sum_{b \in \{-1,1\}^{t \cdot M}} \psi_3(b)F(b) > 1 - 2^{-t} = 1 - 1/n^{100}, \tag{42}$$

$$\sum_{b \in \{-1,1\}^{t \cdot M}} |\psi_3(b)| = 1, \tag{43}$$

$$\sum_{b\in\{-1,1\}^{t\cdot M}} \psi_3(b)\chi_S(b) = 0 \text{ for each } |S| \le d_3 \text{ and} \tag{44}$$

$$\psi_3(b) < 0 \text{ for each } b \in F^{-1}(-1). \tag{45}$$

Now let ψ_4 denote a dual witness to the fact that $\widetilde{\deg}_{.99}(\mathrm{OR}_{M/t}) = \Omega(\sqrt{M/t})$. As observed in [15, Theorem 5.1], any dual witness for this fact will have one-sided error, but on the side opposite from the one we used to define $\widetilde{\mathrm{odeg}}$. Thus there is some $d_4 = \Omega(\sqrt{M/t})$ such that the following equations hold:

$$\sum_{w\in\{-1,1\}^{M/t}} \psi_4(w)\,\mathrm{OR}_{M/t}(w) > .99, \tag{46}$$

$$\sum_{w\in\{-1,1\}^{M/t}} |\psi_4(w)| = 1, \tag{47}$$

$$\sum_{w\in\{-1,1\}^{M/t}} \psi_4(w)\chi_S(w) = 0 \text{ for each } |S| \le d_4 \text{ and} \tag{48}$$

$$\psi_4(\mathbf{1}) > 0. \tag{49}$$

Finally, we combine the dual witnesses ψ_4 and ψ_3 to obtain the desired function ψ_5:

$$\psi_5(z_1,\dots,z_{M/t}) := 2^{M/t}\psi_4(\dots,\widetilde{\mathrm{sgn}}(\psi_3(z_i)),\dots)\prod_{i=1}^{M}|\psi_3(z_i)|, \tag{50}$$

where $z_i = (z_{i,1},\dots,z_{i,t\cdot M})$.

Analysis of ψ_5. The analysis in [12] immediately implies that ψ_5 has L_1-norm equal to 1, has pure high degree at least $d_3\cdot d_4 = \Omega\left(M/\sqrt{t}\right) = \Omega\left(M/\sqrt{\log n}\right)$, and that the correlation of ψ_5 with $\text{OR-AND}_{2,M^2}$ is at least $.99 - 2^{-t} \ge .98$. We claim that ψ_5 satisfies an additional property, which formalizes the notion that ψ_5 has "almost no error on the wrong side". Let $A_{-1} = \{z \in \{-1,1\}^{M^2} : \psi_5(z) < 0, \text{OR-AND}_{2,M^2}(z) = 1\}$. We will show that:

$$\sum_{z\in A_{-1}} |\psi_5(z)| \le 1/n^{100}. \tag{51}$$

To establish Eq. (51), we first collect some observations. Let $B_{-1} = \{z_i \in \{-1,1\}^{M\cdot t} : \psi_3(z_i) < 0, F(z_i) = 1\}$.

- Observation 1: For every $z = (z_1,\dots,z_{M/t}) \in \left(\{-1,1\}^{t\cdot M}\right)^{M/t}$ in A_{-1}, the following property must hold: $z_i \in B_{-1}$ for every i such that $\psi_3(z_i) < 0$. This holds because $F(z_i) = 1$ for *all* $i \in \{1,\dots,M/t\}$, since $\text{OR-AND}_{2,M^2}(z) = 1$.

- Observation 2: For every $z = (z_1,\dots,z_{M/t}) \in \left(\{-1,1\}^{t\cdot M}\right)^{M/t} \in A_{-1}$, there must exist a z_i such that $\psi_3(z_i) < 0$. This is because, if $\psi_3(z_i) \ge 0$ for all $i \in \{1,\dots,M/t\}$, then $\psi_5(z)$ agrees in sign with $\psi_4(\mathbf{1}) > 0$ (see Eq. (49)), contradicting the assumption that $z \in A_{-1}$.

- Observation 3: Let μ be the distribution on $\{-1,1\}^{M^2}$ defined via: $\mu(z_1,\ldots,z_{M/t}) = \prod_{i=1}^{M/t} |\psi_3(z_i)|$. Since ψ_3 is balanced, the string $(\ldots,\widetilde{\text{sgn}}(\psi_3(z_i)),\ldots)$ is distributed uniformly in $\{-1,1\}^{M/t}$ when one samples $z=(z_1,\ldots,z_{M/t})$ according to μ.
- Observation 4: Because ψ_3 has correlation $1-1/n^{100}$ with F (see Eq. (42)), the following equation holds:
$$\sum_{z_i \in B_{-1}} |\psi_3(z_i)| \leq 1/2n^{100}.$$
- Observation 5: As in the proof of Theorem 1, let $\mu(z|w)$ denote the probability of z under μ, conditioned on $(\ldots,\widetilde{\text{sgn}}(\psi_3(z_i)),\ldots) = w$. If $z \sim \mu(\cdot|w)$ for some string w where $w_i = -1$, then the probability that $F(z_i)=1$ when $\widetilde{\text{sgn}}(\psi_3(z_i)) = w_i$ is $2\sum_{z_i \in B_{-1}} |\psi_3(z_i)|$.

Thus, we may write:

$$\sum_{z \in A_{-1}} |\psi_5(z)| = \sum_{z \in A_{-1}} 2^{M/t} |\psi_4(\ldots,\widetilde{\text{sgn}}(\psi_3(z_i)),\ldots)| \prod_i |\psi_3(z_i)|$$

$$\leq \sum_{w \in \{-1,1\}^{M/t}, w \neq \mathbf{1}} |\psi_4(w)| \cdot \Pr_{z \sim \mu(\cdot|w)}[z_i \in B_{-1}\ \forall i : w_i = -1]$$

$$\leq \sum_{w \in \{-1,1\}^{M/t}} |\psi_4(w)| \cdot 1/n^{100} \leq 1/n^{100}.$$

Here, the equality holds by definition of ψ_5 (see Eq. (50)), the first inequality holds by Observations 1, 2 and 3, the second inequality holds by Observations 4 and 5, and the fourth inequality holds because the L_1 norm of ψ_4 is 1 (see Eq. (47)).

Bounding the Correlation of ψ_7 with AND-OR$_{3,n}$. Using Equation Eq. (51), it is possible to adapt the analysis of [12] to show that $\sum_x \psi_7(x)\,\text{AND-OR}_{3,n}(x) > .95$. The goal of the analysis is to show that

$$\sum_x \psi_7(x)\,\text{AND-OR}_{3,n}(x) \approx \sum_{a \in \{-1,1\}^M} \psi_6(a)\,\text{AND}_M(a) > .99. \tag{52}$$

To this end, let $A_{-1} = \{z \in \{-1,1\}^{M^2} : \psi_5(z) < 0, \text{OR-AND}_{2,M^2}(z) = 1\}$ as above, and let $A_1 = \{z \in \{-1,1\}^{2,M^2} : \psi_5(z) \geq 0, \text{OR-AND}_{2,M^2}(z) = -1\}$. Notice that $A_1 \cup A_{-1}$ is the set of all inputs z where the sign of $\psi_5(z)$ disagrees with OR-AND$_{2,M^2}(z)$. Notice that $\sum_{z \in A_1 \cup A_{-1}} |\psi_5(z)| \leq .01$ because ψ_5 has correlation at least .98 with OR-AND$_{2,M^2}$.

Let ν be the distribution on $\left(\{-1,1\}^{M^2}\right)^M$ given by $\nu(x_1,\ldots,x_M) = \prod_{i=1}^M |\nu(x_i)|$. Since ν is orthogonal to the constant polynomial, it has expected value 0, and hence the string $(\ldots,\widetilde{\text{sgn}}(\psi_5(x_i)),\ldots)$ is distributed uniformly in $\{-1,1\}^M$ when one samples $(x_1,\ldots,x_M)$ according to ν. Let $\nu(x_i|a)$ denote the probability of x_i under ν, conditioned on $(\ldots,\widetilde{\text{sgn}}(\psi_5(x_i)),\ldots) = a$.

For any given $a \in \{-1,1\}^M$, the following two random variables are identically distributed:

- The string $(\dots, \text{OR-AND}_{2,M^2}(x_i), \dots)$ when one chooses $(\dots, x_i, \dots)$ from the conditional distribution $\nu(\cdot|a)$.
- The string $(\dots, y_i a_i, \dots)$, where $y \in \{-1,1\}^M$ is a random string whose ith bit independently takes on value -1 with probability $2\sum_{x_i \in A_{a_i}} |\nu(x_i)| \leq .02$.

Thus, the left hand side of Expression (52) equals

$$\sum_{a \in \{-1,1\}^M} \psi_7(a) \cdot \mathbf{E}[\text{AND}_M(\dots, y_i a_i, \dots)], \tag{53}$$

where $y \in \{-1,1\}^M$ is a random string whose ith bit independently takes on value -1 with probability $2\sum_{x_i \in A_{a_i}} |\psi(x_i)| \leq .02$.

All $a \neq -\mathbf{1}_M$ can be handled exactly as in [12] and [46] to argue that they contribute at least $(1-.02)\psi_6(a)$ to the sum. The key property exploited here is that AND_M has low *block-sensitivity* on these points, allowing us to apply the following proposition.

Proposition 22 ([46])**.** *Let $f : \{-1,1\}^M \to \{-1,1\}$ be a given Boolean function. Let $y \in \{-1,1\}^M$ be a random string whose ith bit is set to -1 with probability at most $\gamma \in [0,1]$, and to $+1$ otherwise, independently for each i. Then for every $a \in \{-1,1\}^M$,*

$$\mathbf{P}_y[f(a_1, \dots, a_M) \neq f(a_1 y_1, \dots, a_M a_M)] \leq 2\gamma \, \text{bs}_a(f).$$

In particular, since $\text{bs}_a(\text{AND}_M) = 1$ for all $a \neq -\mathbf{1}_M$, Proposition 22 implies that for all $a \neq -\mathbf{1}_M$, and $a = \text{AND}_M$, $\mathbf{P}_y[f(a_1, \dots, a_M) = f(a_1 y_1, \dots, a_M y_M)] \geq 1 - .02$.

We next argue that the term corresponding to $a = -\mathbf{1}_M$ contributes at least $(1-2M/n^{100})\psi_6(a)$ to Expression (53). By Eq. (51) and a union bound, for $a = -\mathbf{1}_M$, the y_i's are *all* -1 with probability $1 - 2M/n^{100}$, and hence $\mathbf{E}_y[\text{AND}_M(\dots, y_i z_i, \dots)] \geq (1 - 2M/n^{100})\text{AND}_M(-\mathbf{1}_M) = -(1 - 2M/n^{100})$. By Eq. (40), $\widetilde{\text{sgn}}(\psi_6(-\mathbf{1}_M)) = -1$, and thus the term corresponding to $a = -\mathbf{1}_M$ contributes at least $(1 - 2M/n^{100})\psi_6(a)$ to Expression (26) as claimed. We conclude that $\sum_x \psi_7(x)\,\text{AND-OR}_{3,n} \geq .97$.

Completing the proof for $d = 3$. The proof that ψ_7 has L_1-norm 1 and has pure high degree at least $d_5 \cdot d_6 = \Omega\left(n^{1/2}/\log^{1/2}(n)\right)$ is identical to prior work [46] (see also Appendix B). Combined with the fact that $\sum_x \psi_7(x)\,\text{AND-OR}_{3,n} \geq .97$, we conclude that ψ_7 is a dual witness to the fact that $\widetilde{\deg}_{.95}(\text{AND-OR}_{3,n}) = \Omega\left(n^{1/2}/\log^{1/2}(n)\right)$.

The case of general d. For ease of exposition, we focus on the case where d is odd; the case of even d is similar. To construct a dual witness proving that $\widetilde{\deg}(\text{AND-OR}_{d,n}) = \Omega(n^{1/2}/\log^{(d-2)/2}(n))$, we inductively assume that there exists a dual witness ψ_1' for the function $G = \text{AND-OR}_{d-2,n^{1-2/d}}$ satisfying the following properties for some $d_1' = \Omega(n^{(1-2/d)/2}/\log^{(d-3)/2}(n))$.

$$\sum_{y \in \{-1,1\}^{n^{1-2/d}}} \psi_1'(y)G(y) > .99, \tag{54}$$

$$\sum_{y \in \{-1,1\}^{n^{1-2/d}}} |\psi_1'(y)| = 1, \text{ and} \tag{55}$$

$$\sum_{y \in \{-1,1\}^{n^{1-2/d}}} \psi_1'(y)\chi_S(y) = 0 \text{ for each } |S| \leq d_1'. \tag{56}$$

In addition, define $C_1 = \{y : \psi_1'(y) > 0, G(y) = -1\}$. We assume inductively that

$$\sum_{y \in C_1} |\psi_1'(y)| \leq 2n^{1-2/d}/n^{100}. \tag{57}$$

Eq. (57) intuitively captures the property that ψ_1' has "almost no error on the wrong side". (We clarify that when lower bounding the approximate degree of OR-AND$_{d,n}$ rather than AND-OR$_{d,n}$, we replace the inductive hypothesis of Eq. (57) with the equivalent bound on $\sum_{y \in C_{-1}} |\psi_1'(y)|$, where $C_{-1} = \{y : \psi_1'(y) < 0, \text{OR-AND}_{d-2,n^{1-2/d}}(y) = 1\}$.)

As a base case of the induction, the dual witness ψ_1 that we used in the case $d = 3$ clearly satisfies the above properties (in fact, ψ_1 had one-sided error, and therefore satisfied an even stronger condition than Eq. (57)).

Now we set $M = n^{1/d}$, and define $\psi_2, \psi_3, \dots, \psi_7$ exactly as in the case $d = 3$, but with the dual witness ψ_1' in place of the dual witness ψ_1. That is, we let $\psi_2 : \{-1,1\}^t \to \mathbb{R}$ be defined via $\psi_2(\mathbf{1}) = 1/2$, $\psi_2(-\mathbf{1}) = -1/2$, and $\psi_2(b_i) = 0$ for all other $b_i \in \{-1,1\}^t$. We define

$$\psi_3(b_1, \dots, b_t) := 2^t \psi_2(\dots, \widetilde{\text{sgn}}(\psi_1'(b_i)), \dots) \prod_{i=1}^{M} |\psi_1'(b_i)|,$$

where $b_i = (b_{i,1}, \dots, x_{i,M})$. We define ψ_4 to be a dual witness to the fact that $\widetilde{\deg}_{.99}(\text{OR}_{M/t}) = \Omega(\sqrt{M/t})$ for $t = 100 \log n$. We define ψ_5 exactly as in Eq. (50). We define ψ_6 to be a dual witness to the high one-sided approximate degree of AND$_M$, and we define ψ_7 exactly as in Eq. (41).

As above, ψ_7 has L_1-norm 1 and pure high degree at least $d_1' \cdot d_4 \cdot d_6 = \Omega\left(n^{1/2}/\log^{(d-2)/2}(n)\right)$. Here, $d_1' = \Omega\left(n^{(1-2/d)/2}/\log^{(d-3)/2}(n)\right)$ denotes the pure high degree of ψ_1', $d_4 = \Omega\left((M/t)^{1/2}\right)$ denotes the pure high degree of ψ_4, and $d_6 = \Omega\left(M^{1/2}\right)$ denotes the pure high degree of ψ_6.

The analysis that ψ_7 has large correlation with AND-OR$_{d,n}$ proceeds identically to the above, with one modification. In the case of $d = 3$, ψ_1 had one-sided error, so we could directly invoke our hardness amplification result (Theorem 1) to conclude that ψ_3 also had one-sided error, as well as correlation $1 - 2^{-t}$ with the target function $\text{OR}_t(G, \dots, G)$. In the case of general d, ψ_1' does not have one-sided error. However, ψ_1' "almost" has one-sided error, as formalized by Eq. (57). It is straightforward to modify the proof of Theorem 1 to show though ψ_1' satisfies a weaker condition than did ψ_1, the dual witness ψ_3 nonetheless satisfies the following properties.

Let $B_{-1} = \{z_i \in \{-1,1\}^{n^{1-2/d} \cdot t} : \psi_3(z_i) < 0, \text{OR}_t(G, \dots, G)(z_i) = 1\}$, and let $B_1 = \{z_i \in \{-1,1\}^{n^{1-2/d} \cdot t} : \psi_3(z_i) > 0, \text{OR}_t(G, \dots, G)(z_i) = -1\}$. Then:

- $\sum_{z_i \in B_{-1}} |\psi_3(z_i)| \leq 2^{-t}$.
- $\sum_{z_i \in B_1} |\psi_3(z_i)| \leq t \cdot 2n^{1-2/d}/n^{100}$.

That is, ψ_3 has error exponentially small in t on one side, and the error on the other side blows up by at most a factor of t relative to ψ_1. This permits us to obtain a variant of Eq. (51), namely:

$$\sum_{z \in A_{-1}} |\psi_5(z)| \leq 2tM/n^{100}, \tag{58}$$

where as above A_{-1} is defined to via:

$$A_{-1} = \{z \in \{-1,1\}^{n^{1-1/d}} : \psi_5(z) < 0, \text{OR-AND}_{d-1,n^{1-1/d}}(z) = 1\}.$$

The remainder of the analysis establishing that ψ_7 has high correlation with AND-OR$_{d,n}$ is now the same as in the case $d = 3$. In addition, Eq. (58) ensures that the inductive hypothesis holds for depth $d+1$ (by setting ψ_1' within the construction at depth $d+1$ equal to ψ_5). This completes the induction and the proof. □

9 Conclusion

Approximate degree is an important measure of the complexity of a Boolean function, and as highlighted above, it has numerous applications throughout theoretical computer science. We have established a generic form of hardness amplification for approximate degree: a way of taking a Boolean circuit that cannot be pointwise approximated by low-degree polynomials to within constant error in a certain one-sided sense, and constructing a deeper circuit that cannot be pointwise approximated even with very high error. We used this hardness amplification result to obtain new bounds on the discrepancy and threshold weight of AC^0, as well as to obtain new lower bounds for read-once DNFs and AND-OR trees of constant depth. Moreover, our hardness amplification techniques pave the way for further progress – they will automatically translate new lower bounds on the one-sided approximate degree of AC^0 into new bounds on the threshold weight and discrepancy of AC^0. For example, our techniques show that an $\tilde{\Omega}(n)$ lower bound on the one-sided approximate degree of AC^0 would imply an $\exp\left(\tilde{\Omega}(\sqrt{n})\right)$ lower bound on the threshold weight of AC^0 and an $\exp\left(-\tilde{\Omega}(\sqrt{n})\right)$ upper bound on the discrepancy of AC^0.

Our results naturally open a number of important directions for future work. In this paper, we exhibited a depth-three circuit F (consisting of an OR of disjoint copies of Element Distinctness) with threshold weight $W(F, n^{2/5}) = \exp\left(\tilde{\Omega}\left(n^{2/5}\right)\right)$. This bound is tight in the sense that there exists a PTF of degree $\tilde{O}(n^{2/5})$ and weight $\exp\left(\tilde{O}(n^{2/5})\right)$ that computes F. However, we conjecture that F in fact has *threshold degree* $\tilde{\Omega}(n^{2/5})$; that is, for a sufficiently small constant c, we conjecture that $W(F, cn^{2/5}/\log n) = \infty$. Such a lower bound would represent the first super-polylogarithmic improvement over Minsky and Papert's seminal $\Omega(n^{1/3})$ lower bound on the threshold degree of AC^0 from 1968 [31, 34].

Another interesting problem is to determine the discrepancy of polynomial-size DNF formulas. We showed an $\exp\left(-\tilde{\Omega}(n^{2/5})\right)$ upper bound for the discrepancy of polynomial-size depth-three circuits, but for DNFs the best known upper bound remains $\exp\left(-\Omega(n^{1/3})\right)$, while the best known lower bound is $\exp\left(-\tilde{O}(n^{1/2})\right)$ (this follows from an intermediate result of Klivans and Servedio [23]). Closing this gap would settle O'Donnell and Servedio's question of whether the Generalized Winnow or Perceptron algorithms can learn DNF formulas in time $\exp\left(\tilde{O}(n^{1/3})\right)$.

Acknowledgements. We are grateful to Sasha Sherstov and Li-Yang Tan for valuable feedback on an earlier version of this manuscript.

References

[1] S. Aaronson. The polynomial method in quantum and classical computing. In *Proc. of Foundations of Computer Science* (FOCS), page 3, 2008. Slides available at `www.scottaaronson.com/talks/polymeth.ppt`.

[2] S. Aaronson and Y. Shi. Quantum lower bounds for the collision and the element distinctness problems. *Journal of the ACM*, 51(4):595-605, 2004.

[3] , E. Allender. A Note on the Power of Threshold Circuits. In *Proc. of FOCS*, pages 580-584, 1989.

[4] A. Ambainis. Polynomial degree and lower bounds in quantum complexity: Collision and element distinctness with small range. *Theory of Computing*, 1(1):37-46, 2005.

[5] L. Babai, P. Frankl, and J. Simon. Complexity classes in communication complexity (preliminary version). *FOCS*, pages 337–347, 1986.

[6] R. Beals, H. Buhrman, R. Cleve, M. Mosca, and R. de Wolf. Quantum lower bound by polynomials. *Journal of the ACM*, 48(4):778-797, 2001.

[7] P. Beame, W. Machmouchi. The quantum query complexity of AC^0. *Quantum Information & Computation* 12(7-8): 670-676, 2012.

[8] R. Beigel. The polynomial method in circuit complexity. *Proc. of the Conference on Structure in Complexity Theory*, pages 82-95, 1993.

[9] R. Beigel. Perceptrons, PP, and the polynomial hierarchy. *Computational Complexity*, 4:339-349, 1994.

[10] A. Blum. Learning Boolean functions in an infinite attribute space (extended abstract). *STOC*, pages 64–72. ACM, 1990.

[11] H. Buhrman, N. K. Vereshchagin, and R. de Wolf. On computation and communication with small bias. *CCC*, pages 24–32, 2007.

[12] M. Bun and J. Thaler. Dual lower bounds for approximate degree and Markov-Bernstein inequalities. *ICALP*, pages 303-314, 2013.

[13] J. Forster, M. Krause, S. V. Lokam, R. Mubarakzjanov, N. Schmitt, and H. Simon. Relations between communication complexity, linear arrangements, and computational complexity. *FSTTCS*, pages 171-182, 2001.

[14] Y. Freund. Boosting a weak learning algorithm by majority. *Inf. Comput.*, 121(2):256-285, 1995.

[15] D. Gavinsky and A. A. Sherstov. A separation of NP and coNP in multiparty communication complexity. *Theory of Computing*, 6(1):227-245, 2010.

[16] M. Goldmann, J. Håstad, A. A. Razborov. Majority gates vs. general weighted threshold gates. *Computational Complexity* 2: 277-300, 1992.

[17] A. Hajnal, W. Maass, P. Pudlák, M. Szegedy, and G. Turán. Threshold circuits of bounded depth. *J. of Comput. and System Sci.*, 46(2):129-154, 1993.

[18] P. Høyer, M. Mosca, and R. de Wolf. Quantum search on bounded-error inputs. In *Proc. of International Colloquium on Automata, Languages, and Programming*, pages 291–299, 2003.

[19] A. Kalai, A. Klivans, Y. Mansour, and R. Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing* 37(6):1777-1805, 2008.

[20] H. Klauck. On Arthur Merlin games in communication complexity. *CCC*, pages 189–199, 2011.

[21] H. Klauck. Lower bounds for quantum communication complexity. *FOCS*, pages 288-297, 2001.

[22] H. Klauck, R. Špalek, and R. de Wolf. Quantum and classical strong direct product theorems and optimal time-space tradeoffs. SIAM Journal on Computing, 36(5): 1472-1493, 2007.

[23] A. R. Klivans and R. A. Servedio. Learning DNF in time $2^{\tilde{O}(n^{1/3})}$. *J. of Comput. and System Sci.*, 68(2):303-318, 2004.

[24] A. R. Klivans and R. A. Servedio. Toward attribute efficient learning of decision lists and parities. *Journal of Machine Learning Research*, 7:587–602, 2006.

[25] A. R. Klivans and A. A. Sherstov. Lower bounds for agnostic learning via approximate rank. *Computational Complexity*, 19(4):pages 581–604, 2010.

[26] M. Krause. On the computational power of Boolean decision lists. *Computational Complexity* 14(4):362-375, 2005.

[27] M. Krause and P. Pudlák. On the computational power of depth-2 circuits with threshold and modulo gates. *Theor. Comput. Sci.*, 174(1-2): 137-156, 1997.

[28] T. Lee. A note on the sign degree of formulas. *CoRR abs/0909.4607*, 2009.

[29] T. Lee and A. Shraibman. Disjointness is hard in the multi-party number-on-the-forehead model. *CCC*, pages 81–91, 2008.

[30] N. Littlestone. From online to batch learning. *COLT*, pages 269–284, 1989.

[31] M. L. Minsky and S. A. Papert. Perceptions: An Introduction to Computational Geometry. MIT Press, Cambridge, MA., 1969.

[32] N. Nisan. The Communication Complexity of Threshold Gates. *Combinatorics, Paul Erdos is Eighty*, pages 301-315, 1994.

[33] N. Nisan and M. Szegedy. On the degree of boolean functions as real polynomials. *Computational Complexity*, 4: pages 301–313, 1994.

[34] R. O'Donnell and R. Servedio. New degree bounds for polynomial threshold functions. *Combinatorica*, 30(3):327-358, 2010.

[35] R. Paturi. On the degree of polynomials that approximate symmetric Boolean functions (Preliminary Version). *STOC*, pages 468-474, 1992.

[36] A. A. Razborov and A. A. Sherstov. The Sign-Rank of AC^0. *SIAM J. Comput.*, 39(5):1833-1855, 2010.

[37] R. A. Servedio, L.-Y. Tan, and J. Thaler. Attribute-Efficient learning and weight-degree tradeoffs for polynomial threshold functions. *Journal of Machine Learning Research - Proceedings Track*, 23: 14.1-14.19, 2012.

[38] A. A. Sherstov. Approximating the AND-OR Tree. *Theory of Computing*, 2013.

[39] A. A. Sherstov. Communication lower bounds using dual polynomials. *Bulletin of the EATCS*, 95:59-93, 2008.

[40] A. A. Sherstov. Making polynomials robust to noise. *STOC*, pages 747-758, 2012.

[41] A. A. Sherstov. Optimal bounds for sign-representing the intersection of two halfspaces by polynomials. *STOC*, pages 523-532, 2010.

[42] A. A. Sherstov. The pattern matrix method. *SIAM J. Comput.*, 40(6):pages 1969-2000, 2011.

[43] A. A. Sherstov. Separating AC^0 from depth-2 majority circuits. *SIAM J. Comput.*, 28(6):2113-2129, 2009.

[44] A. A. Sherstov. Strong direct product theorems for quantum communication and query complexity. *SIAM J. Comput.*, 41(5):1122–1165, 2012.

[45] A. A. Sherstov. The multiparty communication complexity of set disjointness. *STOC*, pages 525–524, 2012.

[46] A. A. Sherstov. The intersection of two halfspaces has high threshold degree. *Proc. of Foundations of Computer Science (FOCS)* pages 343-362, 2009. To appear in *SIAM J. Comput. (special issue for FOCS 2009)*

[47] A. A. Sherstov. The unbounded-error communication complexity of symmetric functions. *Combinatorica*, 31(5):583-614, 2011.

[48] Y. Shi. Approximating linear restrictions of Boolean functions. Manuscript, 2002. Available online at `web.eecs.umich.edu/~shiyy/mypapers/linear02-j.ps`.

[49] Y. Shi and Y. Zhu. Quantum communication complexity of block-composed functions. *Quantum Information & Computation* 9(5): 444-460, 2009.

[50] R. Špalek. A dual polynomial for OR. Manuscript, 2008. Available online at: `http://arxiv.org/abs/0803.4516`

[51] L. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134-1142, 1984.

A One-Sided Approximate Degree of Element Distinctness

Improving on results of Aaronson and Shi [2], Ambainis [4] showed that the Element Distinctness problem with small range has approximate degree $\tilde{\Omega}(n^{2/3})$. Recall that the Element Distinctness problem on input size $n = N \log N$, where N is a power of 2, takes as input N blocks of length $\log N$ and evaluates to -1 if and only if the blocks are distinct. We will show that there is a dual witness Ψ for the high approximate degree of Element Distinctness having one-sided error. Hence, this dual witness actually demonstrates that Element Distinctness has high *one-sided* approximate degree.

The idea is that any dual witness for Element Distinctness can be "symmetrized" to produce a new dual witness Ψ that is constant on inputs $x \in T$, where T is the set of inputs for which Element Distinctness evaluates to true. We then use the fact that Ψ is balanced to argue that the *total* correlation of Ψ with Element Distinctness is a constant multiple of the correlation restricted to inputs in T. Since Ψ has positive correlation with Element Distinctness, it follows that Ψ must have the correct sign on all inputs in T, as desired.

Formally, let ψ be a dual witness for the fact that $f =$ Element Distinctness has ε-approximate degree $d = \tilde{\Omega}(n^{2/3})$ for some constant ε. By Theorem 12,

$$\sum_{x \in \{-1,1\}^n} f(x)\psi(x) > \varepsilon, \tag{59}$$

$$\sum_{x \in \{-1,1\}^n} |\psi(x)| = 1, \tag{60}$$

and

$$\sum_{x \in \{-1,1\}^n} \psi(x)\chi_S(x) = 0 \text{ for each } |S| \leq d. \tag{61}$$

For any permutation $\sigma \in S^N$, and $x = (x_1, \ldots, x_N) \in \{-1,1\}^n$, define

$$\sigma(x) = (x_{\sigma(1)}, \ldots, x_{\sigma(N)}).$$

That is, σ acts on $\{-1,1\}^n$ by permuting the N blocks of length $\log N$. Observe that for every $\sigma \in S^N$ and every $x \in \{-1,1\}^n$,

$$f(\sigma(x)) = f(x). \tag{62}$$

Now define the symmetrized dual witness

$$\Psi(x) = \mathbb{E}_{\sigma \in S^N}[\psi(\sigma(x))].$$

We will show that Ψ is a dual witness for f with one-sided error by checking the conditions of Theorem 13. First,

$$\begin{aligned}
\sum_{x \in \{-1,1\}^n} \Psi(x)f(x) &= \mathbb{E}_{\sigma \in S^N}\left[\sum_x \psi(\sigma(x))f(x)\right] \\
&= \mathbb{E}_{\sigma \in S^N}\left[\sum_x \psi(x)f(x)\right] && \text{by Eq. (62)} \\
&> \epsilon && \text{by (59)},
\end{aligned}$$

verifying (4). Condition (5) is immediate from (60). Condition (6) follows because

$$\sum_{x\in\{-1,1\}^n} \Psi(x)\chi_S(x) = \mathbb{E}_{\sigma\in S^N}\left[\sum_x \psi(x)\chi_{\sigma(S)}(x)\right]$$

where $\sigma(S) = \{\sigma(i) : i \in S\}$ and from (61).

Finally, we check the one-sided error condition (7). We will first show that Ψ is constant on $f^{-1}(-1)$. Let $x^* = (x_1^*, \ldots, x_N^*)$ where x_i^* is the binary encoding of i. Since there are only N distinct strings of length $\log N$, $f(x) = -1$ if and only if $x = \sigma_x(x^*)$ for some $\sigma_x \in S^N$. Therefore, if $f(x) = -1$, then

$$\Psi(x) = \mathbb{E}_{\sigma\in S^N}[\psi(\sigma(x))] = \mathbb{E}_{\sigma\in S^N}[\psi((\sigma\circ\sigma_x)(x^*))] = \Psi(x^*),$$

so Ψ is constant on $f^{-1}(-1)$.

By condition (4) it holds that

$$\sum_{x\in f^{-1}(1)} \Psi(x) - \sum_{x\in f^{-1}(-1)} \Psi(x) > \varepsilon,$$

and by condition (5) applied to χ_S for $S = \emptyset$ it holds that

$$\sum_{x\in f^{-1}(1)} \Psi(x) + \sum_{x\in f^{-1}(-1)} \Psi(x) = 0.$$

Subtracting the second equation from the first, we conclude that

$$-2\sum_{x\in f^{-1}(-1)} \Psi(x) > \varepsilon.$$

Since Ψ is constant on $f^{-1}(-1)$, this implies that $\Psi(x) < 0$ whenever $x \in f^{-1}(-1)$, proving (7).

B Final Details of the Proof of Theorem 1

B.1 Proof of Equation 21

Let μ be the distribution on $(\{-1,1\}^m)^t$ given by $\mu(x_1,\ldots,x_t) = \prod_{i=1}^t |\psi(x_i)|$. Since ψ is orthogonal to the constant polynomial, it has expected value 0, and hence the string $(\ldots, \widetilde{\mathrm{sgn}}(\psi(x_i)), \ldots)$ is distributed uniformly in $\{-1,1\}^t$ when one samples $(x_1,\ldots,x_t)$ according to μ. Thus,

$$\sum_{(x_1,\ldots,x_t)\in(\{-1,1\}^m)^t} |\zeta(x_1,\ldots,x_t)| = \sum_{z\in\{-1,1\}^t} |\Psi(z)| = |\Psi(\mathbf{1})| + |\Psi(-\mathbf{1})| = 1,$$

proving Eq. (21). □

B.2 Proof of Equation 22

We prove that the polynomial ζ defined in Eq. (24) satisfies Eq. (22), reproduced here for convenience.

$$\sum_{(x_1,\ldots,x_t)\in(\{-1,1\}^m)^t} \zeta(x_1,\ldots,x_t)\chi_S(x_1,\ldots,x_t) = 0 \text{ for each } |S| \le d. \qquad (22)$$

To prove Eq. (22), notice that since Ψ is orthogonal on $\{-1,1\}^t$ to constant functions, we have the Fourier representation

$$\Psi(z) = \sum_{\substack{T\subseteq\{1,\ldots,t\}\\ T\neq\emptyset}} \hat{\Psi}(T)\chi_T(z)$$

for some reals $\hat{\Psi}(T)$. We can thus write

$$\zeta(x_1,\ldots,x_t) = 2^t \sum_{T\neq\emptyset} \hat{\Psi}(T) \prod_{i\in T} \psi(x_i) \prod_{i\notin T} |\psi(x_i)|.$$

Given a subset $S \subseteq \{1,\ldots,t\}\times\{1,\ldots,m\}$ with $|S| \le d$, partition $S = (\{1\}\times S_1)\cup\cdots\cup(\{t\}\times S_t)$ where each $S_i \subseteq \{1,\ldots,m\}$. Then

$$\sum_{(x_1,\ldots,x_t)\in(\{-1,1\}^m)^t} \zeta(x_1,\ldots,x_t)\chi_S(x_1,\ldots,x_t)$$
$$= 2^t \sum_{T\neq\emptyset} \hat{\Psi}(T) \prod_{i\in T} \underbrace{\left(\sum_{x_i\in\{-1,1\}^m} \psi(x_i)\chi_{S_i}(x_i)\right)}_{} \prod_{i\notin T} \left(\sum_{x_i\in\{-1,1\}^m} |\psi(x_i)|\chi_{S_i}(x_i)\right).$$

Since $|S| \le d$, we have that $|S_i| \le d$ for every index $i \in \{1,\ldots,t\}$. Thus for each set T, at least one of the underbraced factors is zero, as χ_{S_i} is orthogonal to ψ whenever $|S_i| \le d$.

C Degree Independent Threshold Weight Bounds via Duality

In this section, we use the dual characterization of threshold weight to give a new proof of a version of Krause's result translating degree-d threshold weight lower bounds for a function F into degree independent threshold weight lower bounds for a related function F'. Specifically, we prove the lemma

Lemma 19. *Let $F : \{-1,1\}^n \to \{-1,1\}$ be a Boolean function, and define $F' : \{-1,1\}^{3n} \to \{-1,1\}$ by*

$$F'(x_1,\ldots,x_n,y_1,\ldots,y_n,z_1,\ldots,z_n) := F(\ldots,(\bar{z}_i \wedge x_i)\vee(z_i\wedge y_i),\ldots).$$

Then

$$W(F')^2 \ge \min\left\{\frac{W(F,d)}{2n}, 2^d\right\}.$$

Proof. By Theorem 16 (condition (14)), it suffices to exhibit a distribution μ' over $\{-1,1\}^{3n}$ for which

$$|\mathbb{E}_{(x,y,z)\sim\mu'}[F'(x,y,z)\chi_S(x,y,z)]| \leq \max\left\{\left(\frac{2n}{W(F,d)}\right)^{1/2}, 2^{-d/2}\right\} \text{ for all } S \subseteq \{1,\dots,3n\}.$$

We construct the distribution μ' as follows. By condition (15) of Theorem 16, there is a probability distribution μ over $\{-1,1\}^n$ such that

$$|\mathbb{E}_{w\sim\mu}[F(w)\chi_S(w)]| \leq \left(\frac{2n}{W(F,d)}\right)^{1/2} \text{ for each } |S| \leq d. \tag{63}$$

Define $\mu'(x,y,z) = 2^{-2n}\mu(\mathrm{Sel}_z(x,y))$, where $\mathrm{Sel}_z(x,y) = (\dots,(\bar{z}_i \wedge x_i)\vee(z_i\wedge y_i),\dots)$ selects for each index in $[n]$ a bit from either x or y according to z. The distribution μ' has a natural interpretation as follows: it first selects the string z uniformly at random from $\{-1,1\}^n$. Next, it sets the values of the variables in (x,y) that are selected by z so that they are distributed according to the distribution μ. Finally, it sets the values of the unselected variables in (x,y) uniformly at random.

Note that μ' is indeed a probability distribution, as for every string $w \in \{-1,1\}^n$, there are exactly 2^{2n} strings (x,y,z) for which $\mathrm{Sel}_z(x,y) = w$. Moreover, this observation allows us to write

$$\mathbb{E}_{(x,y,z)\sim\mu'}[F'(x,y,z)\chi_S(x,y,z)] = 2^{-2n}\sum_{w\in\{-1,1\}^n} F(w)\mu(w) \sum_{(x,y,z):\mathrm{Sel}_z(x,y)=w} \chi_S(x,y,z).$$

Write S as the disjoint union $(\{1\}\times S_1)\cup(\{2\}\times S_2)\cup(\{3\}\times S_3)$ where S_1, S_2, S_3 correspond to indices in x, y, z respectively. Then the expectation becomes

$$2^{-2n}\sum_{z\in\{-1,1\}^n}\chi_{S_3}(z)\underbrace{\sum_{w\in\{-1,1\}^n} F(w)\mu(w)\sum_{(x,y):\mathrm{Sel}_z(x,y)=w}\chi_{S_1}(x)\chi_{S_2}(y)}$$

Let $G(z)$ denote the underbraced sum.

Suppose there is an index $i \in S_3$ that is not contained in $S_1 \cup S_2$. Then for every $z \in \{-1,1\}^n$, the string z^i obtained from z by flipping the bit at index i satisfies $\chi_{S_3}(z^i) = -\chi_{S_3}(z)$. On the other hand, for any $(x,y) \in \{-1,1\}^{2n}$, if we set $x' = (x_1,\dots,x_{i-1},y_i,x_{i+1},\dots,x_n)$ and analogously set $y' = (y_1,\dots,y_{i-1},x_i,y_{i+1},\dots,y_n)$, then $\mathrm{Sel}_z(x,y) = \mathrm{Sel}_{z^i}(x',y')$. Moreover, because $i \notin S_1 \cup S_2$, it holds that $\chi_{S_1}(x')\chi_{S_2}(y') = \chi_{S_1}(x)\chi_{S_2}(y)$. It follows that $G(z) = G(z^i)$, as each term (x,y) in the underbraced sum defining $G(z)$ is "matched" by term (x',y') in the underbraced sum defining $G(z^i)$. When combined with the fact that $\chi_{S_3}(z^i) = -\chi_{S_3}(z)$, we see that the terms corresponding to z and z^i in the outer sum cancel out, and hence the entire outer sum evaluates to zero. We conclude that for the expectation to be nonzero, we must have $S_3 \subseteq S_1 \cup S_2$, and we assume this holds for the remainder of the proof.

Consider any $i \in S_1$. Then we claim that $G(z) = 0$ whenever z_i selects y_i, i.e., for any z such that $z_i = -1$. This can be seen by another pairing argument: If $\mathrm{Sel}_z(x,y) = w$ but z_i selects y_i, then $\mathrm{Sel}_z(x^i,y) = w$ as well. However, $\chi_{S_1}(x) = -\chi_{S_1}(x^i)$ because $i \in S_1$. This ensures that the innermost sum is zero and hence $G(z) = 0$. The analogous statement holds also for any $i \in S_2$, so for $G(z)$ to be nonzero, it must hold that $z_i = 1$ for all $i \in S_1$ and $z_i = -1$ for all $i \in S_2$. Below, we refer to such a z as a "contributing" z, and all other values of z as "non-contributing". In particular, we must have $S_1 \cap S_2 = \emptyset$ for z to be contributing.

For any fixed contributing z, it holds that

$$\sum_{(x,y):\mathrm{Sel}_z(x,y)=w} \chi_{S_1}(x)\chi_{S_2}(y) = 2^n \chi_{S_1 \cup S_2}(w).$$

Therefore, it holds that

$$\begin{aligned}
|\mathbb{E}_{(x,y,z)\sim\mu'}[F'(x,y,z)\chi_S(x,y,z)]| &= 2^{-2n}\left|\sum_{z\in\{-1,1\}^n} \chi_{S_3}(z)G(z)\right| \\
&\leq 2^{-n}\sum_{z:G(z)\neq 0}\left|\sum_{w\in\{-1,1\}^n} F(w)\mu(w)\chi_{S_1\cup S_2}(w)\right| \\
&\leq 2^{-|S_1|-|S_2|}\left|\sum_{w\in\{-1,1\}^n} F(w)\mu(w)\chi_{S_1\cup S_2}(w)\right|, \qquad (64)
\end{aligned}$$

where inequality (64) used the fact that $G(z) = 0$ for any non-contributing z.

Now we consider two cases for the size of S. First suppose $|S| \leq d$, so in particular, $|S_1 \cup S_2| \leq d$. Then Eq. (63) and inequality (64) implies that

$$|\mathbb{E}_{(x,y,z)\sim\mu'}[F'(x,y,z)\chi_S(x,y,z)]| \leq \left(\frac{2n}{W(f,d)}\right)^{1/2}.$$

Second, suppose that $|S| > d$. We have argued that if $\mathbb{E}_{(x,y,z)\sim\mu'}[F'(x,y,z)\chi_S(x,y,z)] \neq 0$, then $S_3 \subseteq S_1 \cup S_2$. Hence, it must be the case that $|S_1| + |S_2| \geq |S|/2 > d/2$. Therefore, inequality (64) implies that $\mathbb{E}_{(x,y,z)\sim\mu'}[F'(x,y,z)\chi_S(x,y,z)] \leq 2^{-d/2}$. This completes the proof.

□

ECCC ISSN 1433-8092

http://eccc.hpi-web.de