Approximate Degree in Classical and Quantum Computing

Mark Bun* Justin Thaler†

May 16, 2023

Abstract

The approximate degree of a Boolean function f captures how well f can be approximated pointwise by low-degree polynomials. This manuscript surveys what is known about approximate degree and illustrates its applications in theoretical computer science.

A particular focus of the survey is a method of proving lower bounds via objects called *dual polynomials*. These represent a reformulation of approximate degree using linear programming duality. We discuss in detail a recent, powerful technique for constructing dual polynomials, called "dual block composition".

*Department of Computer Science, Boston University, Boston, MA 02215, USA. mbun@bu.edu. Supported by NSF grant CCF-1947889.

†Department of Computer Science, Georgetown University, Washington, DC 20057, USA. justin.thaler@georgetown.edu. Supported by NSF CAREER award CCF-184512.

Contents

1 Introduction

The ability (or inability) to represent or approximate Boolean functions by polynomials is a central concept in complexity theory, underlying interactive and probabilistically checkable proof systems, circuit lower bounds, quantum complexity theory, and more. In this manuscript, we survey what is known about a particularly natural notion of approximation by polynomials, capturing pointwise approximation over the real numbers. The *ε-approximate degree* of a Boolean function $f\colon \{-1,1\}^n \to \{-1,1\}$, denoted $\widetilde{\deg}_\varepsilon(f)$, is the least total degree of a real polynomial $p : \{-1,1\}^n \to \mathbb{R}$ such that

$$|f(x) - p(x)| \leq \varepsilon \text{ for all } x \in \{-1,1\}^n. \tag{1}$$

By total degree of p, we refer to the maximum sum of the degrees of all variables appearing in any monomial. For example, $p(x_1, x_2, x_3) = x_1^2 x_2 x_3^2 + x_1 x_2^3$ has total degree 5.

Every Boolean function is approximated to error $\varepsilon = 1$ by the constant 0 function, implying that $\widetilde{\deg}_1(f) = 0$ for all such f. However, whenever ε is strictly less than 1, $\widetilde{\deg}_\varepsilon(f)$ is a fascinating notion with a rich theory and applications throughout theoretical computer science.

Applications of approximate degree lower bounds. The study of approximate degree is itself a "proto-complexity theory" [Aar08], with pointwise approximation by real polynomials serving as a rudimentary model of computation, and degree acting as a measure of complexity. Moreover, when f has large (say, $n^{\Omega(1)}$) approximate degree, it is also hard to compute in a variety of other computational models. Different models correspond to different settings of the error parameter ε with two regimes of particular interest. First, if $\widetilde{\deg}_{1/3}(f)$ is large, then f cannot be efficiently evaluated by *bounded-error* quantum query algorithms [BBC+01].[1] This connection is often referred to as the "polynomial method in quantum computing."

Second, if $\widetilde{\deg}_\varepsilon(f)$ is large *for every* $\varepsilon < 1$, then f is difficult to compute by *unbounded-error* randomized (or quantum) query algorithms (see, e.g., [DGRMT22, Lemma 6]). These are randomized algorithms that are only required to do slightly better than random guessing, and correspond to the complexity class **PP** (short for probabilistic polynomial time) defined by Gill [Gil77]. This connection has been used to answer long-standing questions in relativized complexity, e.g., in studying the power of statistical zero-knowledge proofs (Section 7.2.2), and in communication complexity (Section 10). Approximability of f in this error regime, wherein the error ε is allowed to be arbitrarily close to (but strictly less than) 1,[2] is captured by a notion termed *threshold degree* and denoted $\deg_\pm(f)$.

Applications of approximate degree upper bounds. As just discussed, lower bounds on $\widetilde{\deg}_\varepsilon(f)$ imply hardness results for computing f. There are also many applications of upper bounds on $\widetilde{\deg}_\varepsilon(f)$, typically in the design of fast algorithms in areas such as learning theory [KS04, KKMS08] (see Section 11.2) and differential privacy [TUV12, CTUW14].

[1] The choice of constant 1/3 is made for aesthetic reasons. Replacing $\varepsilon = 1/3$ with any other constant in $(0,1)$ changes the ε-approximate degree of f by at most a constant factor.

[2] Approximate degree is a meaningful notion even for error parameters ϵ that are *doubly-exponentially* close to 1. In particular, for any degree bound d, there are known Boolean functions that can be approximated by degree-d polynomials to error $1 - 2^{-n^{\Theta(d)}}$ but not to smaller error [Pod08, Pod09, BT18a].

In addition to algorithmic applications, approximate degree upper bounds have also been used to prove complexity *lower bounds*. Here is an illustrative example. Suppose one shows that every circuit over n-bit inputs in a class $\mathcal{C}$ can be approximated to error $\varepsilon < 1$ by a polynomial of degree $o(n)$. We know that simple functions f such as Majority and Parity require approximate degree $\Omega(n)$, and therefore cannot be computed by circuits in $\mathcal{C}$. In fact, if $\varepsilon = 1/3$, then one can even conclude that $\mathcal{C}$ is not powerful enough to compute these functions *on average*, meaning that for every circuit $C \in \mathcal{C}$, we have $\Pr_{x \sim \{-1,1\}^n}[C(x) = f(x)] \leq 1/2 + \frac{1}{n^{\omega(1)}}$ [Tal17, BKT21]. This principle underlies several state-of-the-art lower bounds for frontier problems in circuit complexity (Section 11.3.2).

Goals of this survey. This survey covers recent progress on proving approximate degree lower and upper bounds and describes some applications of the new bounds to oracle separations, quantum query and communication complexity, and circuit complexity. On the lower bounds side, progress has followed from an approach called the *method of dual polynomials*, which seeks to prove approximate degree lower bounds by constructing solutions to (the dual of) a certain linear program that captures the approximate degree of any function. This survey explains how several of these advances have been unlocked by a particularly simple and elegant technique—called *dual block composition*—for constructing solutions to this dual linear program. We also provide concise coverage of even more recent lower bound technique based on a new complexity measure called *spectral sensitivity*.

On the upper bounds side, recent explicit constructions of approximating polynomials have been inspired by quantum query algorithms. These constructions also involve new techniques that first express the approximations as sums of exponentially many high-degree terms, and then replace each term with a low-degree approximation that is accurate to exponentially small error.

Roadmap and suggestions for reading the survey. After covering preliminaries (Section 2), we begin in Sections 3 and 4 by covering approximate degree upper bounds, i.e., techniques for constructing low-degree approximations to Boolean functions. We then turn to lower bound techniques, starting with the simpler and older technique of symmetrization (Section 5) before turning to the method of dual polynomials (Section 6). The next two chapters provide progressively more sophisticated developments of this technique, with Section 7 introducing dual block composition as a technique for lower bounding the approximate degree of block-composed functions, and Section 8 moving beyond block-composed functions. Section 9 covers approximate degree lower bounds via spectral sensitivity.

The survey then turns to applications of approximate degree upper and lower bounds. Section 10 covers (a variant of) the so-called pattern matrix method for translating approximate degree lower bounds into approximate-rank and communication lower bounds. Section 11 covers assorted additional applications of both upper and lower bounds on approximate degree.

We have primarily organized the survey by technique. For example, all upper bounds that we cover appear in Sections 3 and 4, with the exception of the approximate degree upper bound for a function called Surjectivity that appears in Section 8.1. This organization maximizes technical and conceptual continuity, but does have some downsides. The results are not covered in increasing order of difficulty, e.g., the easiest lower bounds come after the most challenging upper bounds. It also means that for any specific function or class of functions, the tight upper and lower bounds appear in different parts of the survey.

Readers may wish to skip some of the more technical results that we cover on a first reading. Prominent examples include the upper bound for a function called Element Distinctness in Section 4.4.2, the proof of Theorem 44 in Section 7.2 on a state-of-the-art lower bound for block-composed functions, the entirety of Section 8.5 on lower bounds for problems called Collision and Permutation Testing, and the proof of Theorem 92 in Section 10.5.1, which constructs a dual witness for the high threshold-degree of an AC^0 function with certain "smoothness" properties that are important for applications in communication- and circuit-complexity.

2 Preliminaries

2.1 Terminology and Notation

Boolean functions, Hamming weight, etc. In this manuscript, we primarily model Boolean functions as mapping the domain $\{-1,1\}^n$ to the range $\{-1,1\}$, with -1 interpreted as logical TRUE and $+1$ is interpreted as logical FALSE.

Many authors instead use the domain $\{0,1\}^n$ or the range $\{0,1\}$, with 1 interpreted as TRUE and 0 as FALSE. The domain $\{-1,1\}^n$ and range $\{-1,1\}$ turn out to be the most convenient choice for proving approximate degree lower bounds. and hence we use this convention throughout the vast majority of this survey, The domain $\{0,1\}^n$ and the range $\{0,1\}$ can be more convenient when establishing approximate degree upper bounds; accordingly we do use this domain and range sparingly (see Section 4.5).

We often use a subscript after a function to clarify the number of variables over which it is defined. For example, OR_n denotes the function over domain $\{-1,1\}^n$ that evaluates to -1 if at least one of its inputs equals -1, and otherwise evaluates to 1.

For any input $x \in \{-1,1\}^n$, we let $|x| = \sum_{i=1}^n (1-x_i)/2$ denote the number of coordinates of x equal to -1 and refer to $|x|$ as the *Hamming weight* of x. We let

$$\mathcal{A}(x) = \sum_{i=1}^n x_i.$$

Note that $|x|$ and $\mathcal{A}(x)$ are both degree-1 polynomials in x.

We use $[n]$ to denote the set $\{1,2,\ldots,n\}$, $[n]^*$ to denote $\{0,1,\ldots,n\}$, and $\mathbf{1}_n$ to denote the input in $\{-1,1\}^n$ in which all entries are 1.

For a real number t, we let $\mathrm{sgn}(t)$ equal 1 if t is nonnegative and equal -1 if t is negative. All logarithms in this survey have base 2 unless specified otherwise.

Binomial coefficients. For integers $0 \le k \le n$, $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ denotes the binomial coefficient, and $\binom{n}{\le k}$ denotes the partial sum $\sum_{i=0}^k \binom{n}{i}$. Two standard upper bounds on these quantities are

$$\binom{n}{k} \le \binom{n}{\le k} \le (ne/k)^k$$

and for $n, k > 1$,

$$\binom{n}{\le k} \le n^k.$$

Function composition and product distributions. For two functions f_m, g_b, we denote by $f_m \circ g_b$ the block-composed function over domain $\left(\{-1,1\}^b\right)^m$, i.e.,

$$(f_m \circ g_b)(x_1, \ldots, x_m) \coloneqq f(g(x_1), \ldots, g(x_m)).$$

Given probability distributions $\mu_1, \ldots, \mu_m$ over $\{-1,1\}^b$, we let $\otimes_{i=1}^m \mu_i$ denote the product distribution over $(x_1, \ldots, x_m) \in \left(\{-1,1\}^b\right)^m$ in which x_i is drawn from distribution μ_i. Given a single probability distribution μ, the distribution $\mu^{\otimes n}$ is the product distribution over $(x_1, \ldots, x_n)$ that chooses each x_i independently according to distribution μ. Given a finite set Y, we use $y \sim Y$ to indicate that y is drawn uniformly at random from Y.

Polynomials and their degree. We assume that all polynomials with domain $\{-1,1\}^n$ are multilinear. This means that they have degree at most one in each variable, e.g., $p(x_1, x_2, x_3) = x_1x_2x_3$ is multilinear, but $p(x_1, x_2, x_3) = x_1^2x_2x_3$ is not. This is without loss of generality because $x_i^2 = 1$ whenever $x_i \in \{-1,1\}$. We denote the degree of a univariate polynomial q by $\deg(q)$. For a multivariate polynomial p, we denote by $\deg(p)$ the total degree of p, i.e., the maximum sum of variable degrees over all monomials of p with nonzero coefficients. For example, the polynomial $p(x_1, x_2, x_3) = x_1x_2x_3 + x_1x_3$ has total degree three.

Approximate degree and threshold degree. Recall that in applications of approximate degree, two regimes for the error parameter ε are of particular relevance. The first is $\varepsilon = 1/3$. For brevity, we use $\widetilde{\deg}(f)$ as a shorthand for $\widetilde{\deg}_{1/3}(f)$, and refer to this quantity without qualification as the *approximate degree* of f. The second regime of special interest considers all ε arbitrarily close to, but strictly less than, 1. This regime is equivalent to a notion called the *threshold degree* of f, denoted $\deg_\pm(f)$, which is the least degree of a polynomial p such that

$$p(x) \cdot f(x) > 0 \text{ for all } x \in \{-1,1\}^n. \tag{2}$$

It is not hard to see that the threshold degree of f is greater than d if and only if for *every* $\varepsilon < 1$, f cannot be approximated to error ε by any degree-d polynomial. Any function p that satisfies Condition (2) is said to *sign-represent* f, and p is called a *polynomial threshold function* (PTF) for f. If p has degree 1, then it called a *linear threshold function* (LTF) for f. Another term for an LTF is a *halfspace*. If a nonzero polynomial p satisfies Condition (2) with weak rather than strict inequality, p is said to *weakly* sign-represent f.

Basics of Fourier analysis. For $S \subseteq [n]$, let $\chi_S(x) = \prod_{i \in S} x_i$; we refer to χ_S as the *parity function* over S, or sometimes as the S'th parity function. We will occasionally refer to any such parity function as a *Fourier basis function* or simply as a *monomial*.

For any function $f\colon \{-1,1\}^n \to \mathbb{R}$, there is a unique multilinear polynomial

$$p(x) = \sum_{x \in \{-1,1\}^n} \hat{f}(S) \cdot \chi_S(x)$$

such that $p(x) = f(x)$ for all $x \in \{-1,1\}^n$. The quantity $\hat{f}(S)$ is called the S'th Fourier coefficient of f. The *degree* of f (sometimes referred to as the Fourier degree of f) is the total degree of p, i.e., the size of the largest set S such that $\hat{f}(S) \neq 0$. Sometimes p is referred to as the *multilinear*

extension of f, because, unlike f, it makes sense to evaluate p even at inputs that are not in $\{-1,1\}^n$.

For a real-valued function $f\colon \{-1,1\}^n \to \mathbb{R}$ given by $f(x) = \sum_{S\subseteq[n]} \hat{f}(S)\chi_S(x)$, the *Fourier weight* of f, denoted weight(f), is $\sum_{S\subseteq[n]} |\hat{f}(S)|$. In other words, weight(f) is the ℓ_1-norm of the Fourier coefficients of f.[3] Likewise, the Fourier *sparsity* of f, denoted sparsity(f), is the number of non-zero Fourier coefficients of f.

The following basic fact offers a crude upper bound on the Fourier sparsity and Fourier weight of any low-degree function.

Fact 1. Let $f\colon \{-1,1\}^n \to \mathbb{R}$ satisfy $|f(x)| \leq A$ for all $x \in \{-1,1\}^n$. If f has Fourier degree at most d, then the Fourier sparsity of f is at most $\binom{n}{\leq d}$ and the Fourier weight of f is at most $A \cdot \binom{n}{\leq d} \leq A \cdot n^d$.

Proof. The set $\{\chi_S\colon S \subseteq [n]\}$ forms an orthonormal basis for the vector space of all functions mapping $\{-1,1\}^n \to \mathbb{R}$ under the inner product relation

$$\langle p, q\rangle = 2^{-n} \sum_{x\in\{-1,1\}^n} p(x)\cdot q(x).$$

Accordingly, we can express the Fourier coefficients of f via:

$$\hat{f}(S) = \langle f, \chi_S\rangle = 2^{-n} \sum_{x\in\{-1,1\}^n} f(x)\cdot \chi_S(x).$$

Hence, $|\hat{f}(S)| \leq A$ for each $S \subseteq [n]$. If f has Fourier degree at most d, then $\hat{f}(S) \neq 0$ for at most $\binom{n}{\leq d}$ subsets S. The fact follows. □

We use a special notation for the parity function on all n bits, $\chi_{[n]}$, denoting it by $\oplus_n$. The bit-wise parity of two vectors $x, y \in \{-1,1\}^n$ is denoted $x \oplus y$.

A useful inequality. Several times in this survey, we use the following fact:

Fact 2. For any real number $t > 1$,

$$(1-1/t)^t \leq 1/e,$$

and

$$(1-1/t)^{t-1} \geq 1/e,$$

where $e \approx 2.718$ is Euler's constant.

2.2 The Cast of Characters

We briefly introduce the most important functions that appear throughout this survey. The approximate degree of each of these functions is now mostly or completely understood, and we state what is known for each. Implications of these degree bounds are detailed later in the survey. Beyond their direct consequences, the efforts that led to these results yielded new and broadly applicable techniques for bounding approximate degree.

[3] We remark that many authors use the term "Fourier weight" to refer to the (squared) ℓ_2-norm of the Fourier coefficients of f – this is not the convention used here.

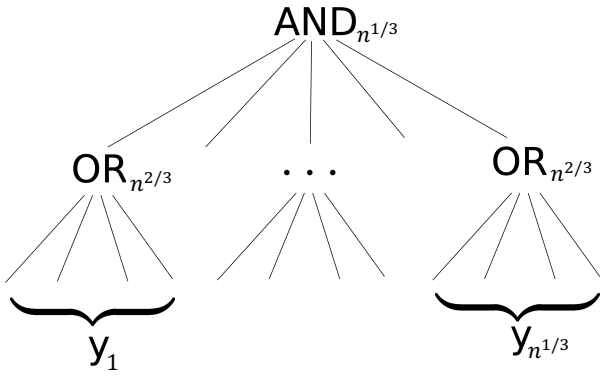


Figure 1: The Minsky-Papert CNF on n variables. The top AND gate has fan-in $n^{1/3}$, while each of the bottom OR gates have fan-in $n^{2/3}$. For each $i = 1, 2, \ldots, n^{1/3}$, $y_i \in \{-1,1\}^{n^{2/3}}$ denotes the input to the ith OR gate.

Symmetric functions. A *symmetric* function $f : \{-1,1\}^n \to \{-1,1\}$ is any function that is invariant under permutations of its input bits. Equivalently, a function is symmetric if its value on any input x depends only on $|x|$, the number of -1s, also known as the Hamming weight of x. Examples of symmetric functions include the OR, AND, Parity, and Majority functions. These ask whether their inputs have Hamming weight that is greater than 0, exactly n, odd, or at least $n/2$, respectively.

The approximate degree and threshold degree of symmetric functions are completely understood, with several proofs of the upper and lower bounds now known. For approximate degree, the upper and lower bounds are covered in Sections 4.3 and 5.2 (see also Section 7.3.2), while for threshold degree they are covered in Sections 3.1 and 5.2.

Theorem 3. Let $f : \{-1,1\}^n \to \{-1,1\}$ be a symmetric function and t be the smallest number such that f is constant on all inputs of Hamming weight between t and $n-t$. Then for $\varepsilon \in (2^{-n}, 1/3)$, we have $\widetilde{\deg}_\varepsilon(f) = \Theta\left(\sqrt{nt} + \sqrt{n\log(1/\varepsilon)}\right)$.

Theorem 4. Let $[n]^* = \{0, 1, \ldots, n\}$. Let $f : \{-1,1\}^n \to \{-1,1\}$ be a symmetric function where $f(x) = F(|x|)$ for a function $F\colon [n]^* \to \{-1,1\}$. Then $\deg_\pm(f) = |\{i : F(i) \neq F(i+1)\}|$.

DNF and CNF formulas. A function that has served a central historical role in the study of approximate degree and threshold degree is the so-called *Minsky-Papert* DNF and CNF.[4] These are, respectively, the read-once DNF and CNF over n variables, with top fan-in $n^{1/3}$ and bottom fan-in $n^{2/3}$. Here, the top fan-in refers to the number of terms or clauses in the DNF or CNF, respectively, while the bottom fan-in refers to the width of each term or clause. See Figure 1 for a depiction. It is known that the Minsky-Papert DNF and CNF have approximate degree $\Theta(n^{1/2})$ (Sections 3.5 and 7.1) and threshold degree $\tilde{\Theta}(n^{1/3})$ (Sections 3.3 and 5.3).[5]

[4] A DNF formula (short for *disjunctive normal form*) of size s is an OR of at most s ANDs, where each AND is evaluated over a set of literals. Here, a literal is an input variable or its negation. So, for example, $\mathsf{OR}(x_1 \wedge \bar{x}_2, x_1 \wedge x_3, x_2 \wedge \bar{x}_3)$ is a DNF of size 3. A CNF is defined analogously with the role of OR and AND reversed.

[5] The tilde notations $\tilde{\Theta}$, $\tilde{\Omega}$, or $\tilde{O}$ hide polylogarithmic, rather than merely constant, factors.

Properties of lists. Several important functions, mainly arising from the quantum computing literature, are best thought of as "properties" of lists of numbers. Let $N \geq R$ with R a power of 2. Each of the following functions takes as input a vector in $x \in \{-1,1\}^n$ with $n = N \log_2 R$. It interprets the vector as a list of (the binary representations of) N numbers $(k_1, \ldots, k_N)$ from range $[R] = \{1, \ldots, R\}$, and it outputs -1 if and only if that list has the stated property. For example, suppose $N = R = 4$, and the range elements $1, \ldots, 4$ are represented in binary by $(1,1)$, $(1,-1)$, $(-1,1)$, and $(-1,-1)$. Then $x = ((1,1),(-1,1),(1,-1),(-1,1))$ would be interpreted as the following list of four numbers: $(1,3,2,3)$.

Two properties of such a list that one might be interested in are: Does every range element appear at least once in the list? Does any range element appear more than once? As we discuss next, the first property is captured by the so-called Surjectivity function, and the second by the Element Distinctness function.

Surjectivity (SURJ) For every $i \in [R]$, does there exist at least one index j such that $k_j = i$? Equivalently, if the input list $(k_1, \ldots, k_N)$ is thought of as the evaluation table of some function, then SURJ evaluates to -1 if the function is surjective.

It is known that if $R = N/2$, then the approximate degree of SURJ is $\tilde{\Theta}(n^{3/4})$ (Section 8.1.1) and the threshold degree is $\tilde{\Theta}(n^{1/2})$ (Section 8.1.3).

For a positive integer ℓ, call a list of N numbers ℓ-*to*-1 if ℓ divides N, and N/ℓ range items each appears exactly ℓ times in the list. So for example, if $N = 4$ and $R = 4$, then the list $(4,3,2,1)$ is 1-to-1, and the list $(4,4,2,2)$ is 2-to-1. If $x \in \{-1,1\}^n$ is interpreted as a list $(k_1, \ldots, k_N)$ that is ℓ-to-1, we also refer to x itself as ℓ-to-1. If $N = R$, we refer to 1-to-1 inputs as *permutations*.

Element Distinctness ED**:** Is the list 1-to-1? That is, for every $i \in [R]$, is it the case that there is at most one index j such that $k_j = i$? Or put yet another way, are all numbers in the list distinct?

Element Distinctness generalizes to the k-distinctness function k-ED for integers $k > 2$. This function interprets its input as a list of N numbers from a range of size R and outputs 1 if and only if there is some range item that appears at least k times in the list.

The Collision Problem: Is the list 1-to-1, under the promise that it is either 1-to-1 or 2-to-1?

Permutation Testing (PTP)**:** Is the list a permutation, under the promise that it is either a permutation or *far* from any permutation? By far from any permutation, we mean that at least, say, 10% of range items do not appear even once in the list.

The above three functions all have threshold degree $\tilde{O}(1)$,[6] but their approximate degrees are much more interesting. For $N = R$, ED has $(1/3)$-approximate degree $\tilde{\Theta}(n^{2/3})$ (Sections 4.4.2 and 8.6). Meanwhile, the Collision and Permutation Testing problems have $(1/3)$-approximate degree $\tilde{\Theta}(n^{1/3})$ (Section 8.5).[7] k-ED is known to have approximate degree approaching $n^{3/4}$ from below

[6] ED, the Collision Problem, and PTP each evaluates to FALSE if and only if there is one or more *collisions* in the input list, meaning a pair of list elements $k_j = k_{j'}$ with $j \neq j'$. One can count the number of collisions in the input list by summing over each pair of list elements and checking if the pair collides. This yields a polynomial of degree $\tilde{O}(1)$. An appropriate affine transformation of the polynomial approximates each of the three functions to error $1 - \Theta(1/n^2)$.

[7] See Section 7.3.2 for discussion of how to define the approximate degree of promise problems such as these.

Function	Approximate Degree	Threshold Degree
OR and AND	$\Theta(n^{1/2})$	1
Symmetric t-threshold function ($t \leq n/2$)	$\Theta\left(\sqrt{nt}\right)$	1
Minsky-Papert DNF and CNF	$\Theta(n^{1/2})$	$\tilde{\Theta}(n^{1/3})$
Surjectivity	$\tilde{\Theta}(n^{3/4})$	$\tilde{\Theta}(n^{1/2})$
Element Distinctness	$\tilde{\Theta}(n^{2/3})$	$\tilde{\Theta}(1)$
Collision Problem and Permutation Testing	$\tilde{\Theta}(n^{1/3})$	$\tilde{\Theta}(1)$

Table 1: Approximate degree and threshold degree of the "cast of characters"—specific functions on n bits playing a prominent role in the study of approximate degree. The symmetric t-threshold function evaluates to -1 if and only if the Hamming of the input is at least t.

for large values of k. Exactly how quickly $\widetilde{\deg}_{1/3}(k\text{-ED})$ approaches $n^{3/4}$ as k grows remains open, i.e., for every fixed $k > 2$, there remains a polynomial gap between the known upper and lower bounds on $\widetilde{\deg}_{1/3}(k\text{-ED})$.

3 General Upper Bound Techniques

In this section, we discuss several fundamental techniques for proving approximate degree and threshold degree upper bounds. Notable examples include polynomial interpolation, approximation via Chebyshev polynomials, and sign-representation via rational approximations.

3.1 Interpolation

A simple place to begin is sign representation of symmetric functions. Recall that a symmetric function f mapping $\{-1,1\}^n$ to $\{-1,1\}$ is one that depends only on the Hamming weight of its input, so it is characterized by $f(x) = F(|x|)$ for some univariate function $F\colon [n]^* \to \{-1,1\}$. One example is the OR_n function, which takes the value 1 when $|x| = 0$ and the value -1 otherwise.

It is easy to show that, for $\varepsilon \geq 1 - 1/n$, $\widetilde{\deg}_\varepsilon(\mathsf{OR}_n) = 1$. To see this, let

$$p(x) = \frac{1}{n}(1 - 2|x|). \tag{3}$$

Then $\deg(p) = 1$ since $|x|$ is a linear funtion of x, and moreover

$$|p(x) - \mathsf{OR}_n(x)| \leq 1 - 1/n \text{ for all } x \in \{-1,1\}^n. \tag{4}$$

We can think of the above polynomial p as arising from the following process. Rather than considering F, which is defined over the discrete domain $[n]^*$, consider the piecewise-linear function G that "extends" the domain of F to the continuous interval $[0,n] \subset \mathbb{R}$, i.e., such that

$$G(i) = F(i) \text{ for all } i \in [n]^*. \tag{5}$$

In the case of OR_n, G has a single root, at input $t = 1/2$, and Equation (3) defines p to be the degree-1 function with this root (p is then scaled appropriately to minimize its approximation error, ensuring that Equation (4) holds).

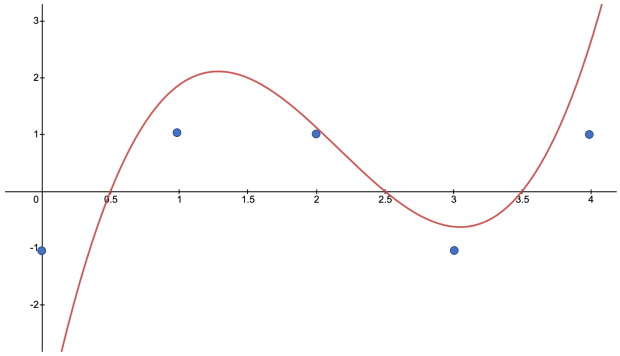


Figure 2: A degree-3 polynomial p defined via Equation (6) to sign-represent the following function $f(x)\colon \{-1,1\}^n \to \{-1,1\}$ with three "sign changes". Define $f(x) = F(|x|)$ where $F(0) = -1$, $F(1) = F(2) = 1$, $F(3) = -1$ and $F(k) = 1$ for all $k \in \{4,\ldots,n\}$. The behavior of F at inputs in $\{0,1,2,3,4\}$ are depicted with blue dots.

In general, let f be a symmetric function that changes sign k times as the Hamming weight of its input increases from 0 to n. Then f can be sign-represented by a polynomial p of degree k, i.e., $\mathrm{sgn}(p(x)) = f(x)$ for all $x \in \{-1,1\}^n$.

More explicitly, let $f(x) = F(|x|)$ where $t_1,\ldots,t_k$ are the Hamming weights where $F(t_i) \neq F(t_i+1)$. Then either the following degree-k polynomial sign-represents f, or its negation does:

$$p(x) = (|x| - (t_1 + 1/2))(|x| - (t_2 + 1/2))\ldots(|x| - (t_k + 1/2)). \tag{6}$$

Indeed, one can check that $p(x) > 0$ if $|x| \geq t_k + 1$ since in this case all terms in the product on the right hand side of Equation (6) are strictly positive. Similarly, $p(x) < 0$ if $|x| \in [t_{k-1}+1, t_k]$, since in this case the final term in the product is negative and all other terms are positive. And $p(x) > 0$ if $|x| \in [t_{k-2}+1, t_k]$, since the final two terms in the product are negative and all other terms are positive. And so on. See Figure 2 for an example.

The polynomial p in Equation (6) is precisely the degree-k polynomial that has the same roots as the natural piecewise-linear function G satisfying Equation (5). We refer to this construction as performing interpolation, since it is defining p via its evaluations at specific points, namely at G's roots.

The threshold degree upper bound given above is in fact tight for any symmetric function, as we show in Section 5.2. Unfortunately, while the behavior of the constructed sign-representing polynomial p is precisely controlled at the interpolation points, p can grow rapidly as one moves away from those points. In particular, $|p(x)|$ can be as large as $n^{\Theta(k)}$ for some inputs $x \in \{-1,1\}^n$ (and it is easily checked that $|p(x)|$ is never less than $1/2$ for any $x \in \{-1,1\}^n$). This means that interpolation-based constructions do not tell us much about the ε-approximate degree of symmetric functions when ε is less than $1 - n^{-O(k)}$. For that, we need more sophisticated tools. The next section describes perhaps the single most important such tool: the Chebyshev polynomials.

3.2 Chebyshev Approximations

Approximate degree considers low-degree point-wise approximations of functions over the discrete, n-dimensional domain $\{-1,1\}^n$. For centuries, the field of approximation theory has studied a

related notion: low-degree point-wise approximations of functions defined over the continuous, unidimensional domain $[-1,1]$. A typical question in this area is as follows: Given a function $F : [-1,1] \to [-1,1]$, what is the least degree of a real polynomial $P : [-1,1] \to \mathbb{R}$ such that $|P(x) - F(x)| \leq \varepsilon$ for all $x \in [-1,1]$? So-called "Jackson-type" theorems give general answers to this question in terms of the continuity and smoothness properties of F. The *Chebyshev polynomials* are central tools for proving results of this type.

Definition 5. For $d \geq 0$, the d'th Chebyshev polynomial of the first kind is the unique polynomial $T_d : \mathbb{R} \to \mathbb{R}$ of degree d such that $T_d(\cos\theta) = \cos(d\theta)$ for every $\theta \in \mathbb{R}$.

One can show using the double-angle formula in trigonometry that the function T_d in Definition 5 is indeed a polynomial of degree at most d. The Chebyshev polynomials have several alternative characterizations, including one given by the recurrence $T_0(x) = 1$, $T_1(x) = x$, and $T_d(x) = 2xT_{d-1}(x) - T_{d-2}(x)$.

Their importance to classical approximation theorem stems from a number of "extremal" properties. For instance, $x^d - 2^{-d+1}T_d(x)$ is the polynomial of degree $d-1$ that provides the best pointwise approximation to x^d. The Chebyshev polynomials are also extremal for a classical result called *Markov's inequality* [Mar90].[8] This inequality states that if G is degree-d polynomial that is bounded over the interval $[-1,1]$, then its derivative $G'(t)$ cannot be too large at any point within the interval.

Theorem 6. (Markov's inequality) Let $G\colon [-1,1] \to [-1,1]$ be a real polynomial of degree at most d. Then $\max_{t\in[-1,1]} |G'(t)| \leq d^2$.

The Chebyshev polynomials are exactly extremal for this result: For any integer $d > 0$, the degree-d Chebyshev polynomial T_d satisfies $|T_d(t)| \leq 1$ for all $t \in [-1,1]$, while $T'_d(t) \geq d^2$ for all $t \in [1,\infty]$, with equality at $t = 1$. In particular, for $d = \lfloor\sqrt{2n}\rfloor$, $T'_d(1) \approx 2n$.

It should come as little surprise, then, that Chebyshev polynomials are useful tools for the discrete approximation problem captured by approximate degree. Their simplest application is to a tight upper bound on the approximate degree of the OR function [NS94].

Lemma 7. $\widetilde{\deg}_{1/3}(\mathsf{OR}_n) = O(\sqrt{n})$.

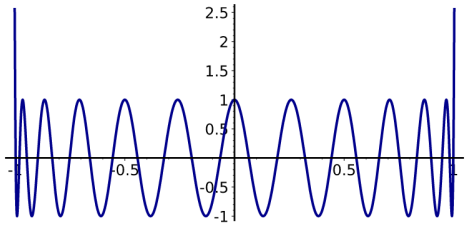
Proof. Our goal is to construct a polynomial $p(x)$ such that $p(1^n) = 1$ and $p(x) = -1$ for all other x. Recalling that $\mathcal{A}(x) = \sum_{i=1}^n x_i$, we may do so by defining $p(x) = q(\mathcal{A}(x)/n)$ where q is a univariate polynomial of degree $O(\sqrt{n})$ such that $q(1) = 1$ and $q(t) \leq -2/3$ for all $t \in [-1, 1-2/n]$. That is, as its input t decreases from 1, the polynomial q "jumps" very quickly from 1 down toward -1, and stays near -1 until t leaves the interval $[-1, 1-2/n]$.

By shifting and scaling T_d without increasing its degree (i.e., by performing an affine transformation), we can obtain a univariate polynomial q with these properties. En route to constructing this polynomial q, consider the polynomial $T_d(t + 4/d^2)$ where $d = \lfloor\sqrt{2n}\rfloor$. While $T_d(t + 4/d^2)$ is bounded by 1 for all $t \in [-1, 1-2/n]$, the fact that $T'_d(t) \geq d^2$ for all $t \geq 1$ ensures that it jumps up to at least 4 for $t = 1$. An additional affine shift produces the following polynomial q with the desired behavior:

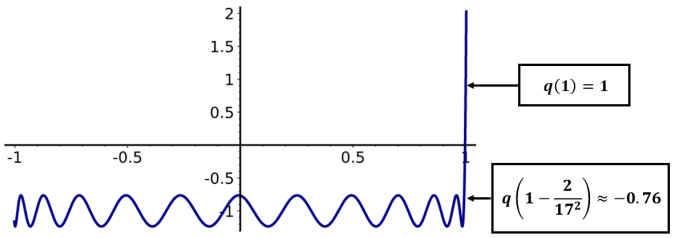
$$q(t) = 2 \cdot \frac{T_d(t + 4/d^2)}{T_d(1 + 4/d^2)} - 1. \tag{7}$$

See Figures (a) and (b) for illustrations of $T_d(t)$ and $q(t)$ when $n = 17^2 = 289$ and $d = \lfloor\sqrt{2n}\rfloor = 24$.

[8] Not to be confused with Markov's inequality from probability theory, but rather a special case of the "Markov brothers' inequality" attributed jointly to A.A. Markov and V.A. Markov.



(a) The degree-24 Chebyshev polynomial T_{24}, which has derivative $24^2 = 576$ at input 1.



(b) The polynomial q from Equation (7) with $d = 24$ (used to approximate OR_n for $n = 17^2 = 289$).

□

A slight generalization of this proof shows that $\widetilde{\deg}_\varepsilon(\mathsf{OR}_n) = O(\sqrt{n}\log(1/\varepsilon))$ for vanishing ε (one can also apply generic error reduction as discussed in Section 3.4). This turns out not to be optimal, as OR_n can be approximated to error ε using degree $O(\sqrt{n\log(1/\varepsilon)})$ by taking advantage of the discrete nature of the problem. We describe this upper bound in detail in Section 4.2.1.

Combining what we know about sign representation and Chebyshev approximation already allows us to construct interesting sign representations for CNF formulas. Specifically, consider the block-composed function $\mathsf{AND}_m \circ \mathsf{OR}_b$. Let $x = (x_1, \ldots, x_m) \in \left(\{-1,1\}^b\right)^m$ denote an arbitrary input to this function.

Lemma 8. $\deg_\pm(\mathsf{AND}_m \circ \mathsf{OR}_b) = O\left(\sqrt{b\log m}\right)$.

Proof. Let p be a polynomial of degree $O\left(\sqrt{b\log m}\right)$ that approximates OR_b to error $1/(3m)$. Letting $x_i \in \{-1,1\}^b$ denote the input to the ith OR gate, the following polynomial of degree $O\left(\sqrt{b\log m}\right)$ sign-represents $(\mathsf{AND}_m \circ \mathsf{OR}_b)(x_1, \ldots, x_m)$:

$$q(x_1, \ldots, x_m) \coloneqq -1 + \sum_{i=1}^{m}(1 + p(x_i)). \tag{8}$$

Indeed, if $\mathsf{OR}(x_i) = -1$ for all i, then $|1 + p(x_i)| \leq 1/(3m)$ for all i, and hence $q(x) \leq -2/3 < 0$. Meanwhile, if $\mathsf{OR}(x_i) = 1$ for even a single i, then $(1 + p(x_i)) \geq 2 - 1/(3m)$ and hence $q(x) > 0$. □

Note that the sign representing polynomial q constructed above is *not* a bounded-error approximation, since it can grow to as large as $\Omega(m)$ on inputs for which $\mathsf{OR}(x_i) = 1$ for $\Omega(m)$ values of i.

3.3 Rational Approximation and Threshold Degree Upper Bounds

Another way to sign-represent a CNF formula is to approximate each OR_b using a *ratio* of low-degree polynomials, combine these as before using a linear sign-representation of AND_m, and then "clear the denominator" to obtain a polynomial. This is a key idea underlying Beigel, Reingold, and Spielman's famous result that **PP** is closed under intersection [BRS95]. Specifically, there are degree-1 polynomials $p(x), q(x)$ over domain $\{-1,1\}^b$ such that the ratio p/q approximates OR_b to

error $1/(3m)$ as follows. For $M \geq 6m$, let $p(x) = 1 - M \cdot |x|$ and $q(x) = 1 + M \cdot |x|$. Then if $x = \mathbf{1}_b$, we have $p(x)/q(x) = \frac{1}{1} = 1$, while if $x \neq \mathbf{1}_b$, we have $\frac{p(x)}{q(x)} \in \left[-1, \frac{1-M}{1+M}\right] \subseteq \left[-1, -1 + \frac{1}{3m}\right]$.

Since $p(x)/q(x)$ approximates OR_b to error $1/(3m)$, by analogy with Equation (8), the following quantity sign-represents $(\mathsf{AND}_m \circ \mathsf{OR}_b)(x_1, \ldots, x_m)$:

$$-1 + \sum_{i=1}^{m} \left(1 + \frac{p(x_i)}{q(x_i)}\right). \tag{9}$$

Unfortunately, Expression (9) is not itself a low-degree polynomial; rather, it is a sum of ratios of degree-1 polynomials. To get a polynomial that sign-represents $\mathsf{AND}_m \circ \mathsf{OR}_b$, we place all terms in the sum of Expression (9) over the common denominator $r(x) = \prod_{j=1}^{m} q(x_j)$. That is, for $i = 1, \ldots, m$, let $s_i(x) \coloneqq p(x_i) \cdot \prod_{j=1,\ldots,m:\, j \neq i} q(x_j)$. Then Expression (9) becomes

$$\frac{1}{r(x)} \cdot \left((m-1)r(x) + \sum_{i=1}^{m} s_i(x)\right).$$

Finally, observe that $r(x) > 0$ for all $x \in \{-1, 1\}^n$, as it is a product of polynomials $q(x_j)$ which are all positive. Hence, multiplication by the denominator $r(x)$ does not alter the sign of the expression. This means that $p^*(x) \coloneqq (m-1)r(x) + \sum_{i=1}^{m} s_i(x)$ sign-represents $\mathsf{AND}_m \circ \mathsf{OR}_b$ and is clearly a polynomial of degree at most m.

Lemma 9. $\deg_{\pm}(\mathsf{AND}_m \circ \mathsf{OR}_b) = O(m)$.

Lemmas 8 and 9 together imply that $\mathsf{AND}_m \circ \mathsf{OR}_b$ has threshold degree at most

$$O(\min\{\sqrt{b \log m}, m\}).$$

Minsky and Papert [MP69] famously showed a (nearly) matching lower bound of $\Omega\left(\min\{\sqrt{b}, m\}\right)$. Later, we describe multiple proofs of Minsky and Papert's lower bound; see Theorem 26 for one based on symmetrization and Theorem 48 for a more general result based on the method of dual polynomials.

3.4 Error Reduction for Approximating Polynomials

Theorem 10. Let $f\colon \{-1,1\}^n \to \{-1,1\}$ be any function. Then for any $\varepsilon > 0$, $\widetilde{\deg}_{\varepsilon}(f) \leq O(\widetilde{\deg}(f) \cdot \log(1/\varepsilon))$.

Proof. We follow the exposition of [DGJ+10, Claim 4.3], which uses a construction called amplifying polynomials. For even integers $\ell > 0$, define the degree-ℓ univariate polynomial

$$A_\ell(u) \coloneqq \sum_{k \geq \ell/2} \binom{\ell}{k} ((1+u)/2)^k \, ((1-u)/2)^{\ell-k}.$$

On any input $u \in [-1, 1]$, $A_\ell(u)$ equals the probability that the majority of tosses of a coin come up Heads, when the probability of seeing Heads is $(1+u)/2$. This polynomial has the following properties (easily proved via the Chernoff bound):

- If $u \in [3/5, 1]$, then $2A_\ell(u) - 1 \in [1 - 2e^{-\ell/6}, 1]$.
- If $u \in [-1, -3/5]$, then $2A_\ell(u) - 1 \in [-1, -1 + 2e^{-\ell/6}]$.

Let p be a 1/3-approximating polynomial for f, and let $\ell = O(\log(1/\varepsilon))$ be a large enough even integer such that $2e^{-\ell/6} \leq \varepsilon$. Then the composition $2A_\ell(p) - 1$ is an ε-approximation for f. □

There is a natural algorithmic interpretation of the construction of Theorem 10. Later, in Section 4.1, we will see that for a polynomial p approximating a Boolean function f, $(1 - p(x))/2$ can often be thought of as outputting the acceptance probability of some (randomized or even quantum) algorithm $\mathcal{A}$ computing f, when $\mathcal{A}$ is run on input x. In this context, $A_\ell(p(x))$ is the acceptance probability of the algorithm that runs ℓ independent copies of $\mathcal{A}$ on x and outputs the majority answer.

3.5 Robust Composition

A beautiful and important result of Sherstov (refining earlier work of Buhrman et al. [BNRdW07]) shows that ε-approximate degree can increase at most multiplicatively under block composition.

Theorem 11 ([She12a])**.** Let $f\colon \{-1,1\}^m \to \{-1,1\}$ and $g\colon \{-1,1\}^n \to \{-1,1\}$ be arbitrary Boolean functions. Then $\widetilde{\deg}(f \circ g) \leq O(\widetilde{\deg}(f) \cdot \widetilde{\deg}(g))$.

Sketch. A natural approach is to attempt to construct a low-degree approximation for $f \circ g$ by simply taking low-degree approximations p and q to f and g respectively, and hoping that the composition $p \circ q$ approximates $f \circ g$. Unfortunately, this does not work directly. The issue is that while $|p(x) - f(x)| \leq 1/3$ for all $x \in \{-1,1\}^m$, this does not grant us any information about p's behavior at *non-Boolean* inputs, while q's outputs may indeed be non-Boolean, because q does not *exactly* compute g, but rather only approximates it at each input in $\{-1,1\}^n$. Hence, understanding the behavior of $p \circ q$ requires us to control p's behavior not only at inputs in $\{-1,1\}^m$, but in fact at real-valued inputs.

Fortunately, q's outputs are not totally arbitrary real numbers: since q approximates a Boolean function g to error at most 1/3, we know that $q(x) \in [-4/3, -2/3] \cup [2/3, 4/3]$ for all $x \in \{-1,1\}^n$ (in particular, q never outputs values in the interval $(-2/3, 2/3)$). What we need is a way to make p *robust* to q's errors, in the sense that, for $i = 1, \ldots, m$, feeding $q(x_i)$ as the i'th input to p rather than $g(x_i)$ does not substantially alter the output of p.

Sherstov [She12a], building on earlier work of [BNRdW07], showed how to accomplish this with only a constant-factor increase in the degree of q. Specifically, he proved the following lemma, whose proof we omit. However, in the next section (Theorem 19), we do prove a weaker result with a far simpler (and algorithmically-inspired) proof, due to Buhrman et al. [BNRdW07].

Lemma 12 (Sherstov [She12a])**.** For any $\delta > 0$ and any polynomial $p\colon \{-1,1\}^m \to [-4/3, 4/3]$ there is a corresponding polynomial $p_{\text{robust}}\colon \mathbb{R}^m \to \mathbb{R}$ of degree $O(\deg p + \log 1/\delta)$ that is robust to noise in the inputs: $|p(x) - p_{\text{robust}}(x + \varepsilon)| < \delta$ for all $x \in \{-1,1\}^m$ and all $\varepsilon \in [-1/3, 1/3]^n$.[9]

Accordingly, one can construct an approximating polynomial for $f \circ g$ by taking any (1/3)-approximating polynomial q for g and any (1/4)-approximating polynomial p for f, applying

[9]Note that p_{robust} may not be multilinear, as Lemma 12 requires control of its behavior at inputs outside of $\{-1,1\}^m$.

Lemma 12 to p (with $\delta = 1/12$) to obtain a polynomial p_{robust} with only a constant-factor blowup in degree. Lemma 12 implies that for all inputs $(x_1, \ldots, x_m) \in \{-1,1\}^{n \cdot m}$ to $f \circ g$, $|(p_{\text{robust}} \circ q)(x_1, \ldots, x_m) - (f \circ g)(x_1, \ldots, x_m)| < (1/4) + (1/12) = 1/3$.

□

In contrast to the situation for related measures such as quantum query complexity, it is still open whether the bound in Theorem 11 is tight for every pair of functions f, g.

Open Problem 13. For every pair of total Boolean functions f, g, is it the case that $\widetilde{\deg}(f \circ g) \geq \Omega(\widetilde{\deg}(f) \cdot \widetilde{\deg}(g))$?

4 Polynomials from Query Algorithms

4.1 A (Very) Brief Introduction to Query Complexity

In (deterministic) query complexity, an algorithm is given oracle access to the bits of an unknown input $x \in \{-1,1\}^n$. Its goal is to evaluate a known function f on x by making as few queries to the oracle as possible. Quantum query complexity is a generalization of this model wherein the algorithm is allowed to make queries in superposition, and must output $f(x)$ with probability at least $2/3$. Here, making queries in superposition roughly means that at each step, the algorithm assigns an amplitude (the quantum analog of a probability) to each possible query that it might make, in the same way a randomized algorithm at each step assigns a probability to each bit to determine which one to query at random. We refer the reader to [Amb18] for details of the model and a recent survey of results.

While quantum query complexity is an information-theoretic model (i.e., the query algorithm is allowed to spend as long as it wants to decide which bits of x to query and to process the oracle's responses to the queries), it turns out to capture much of the power of quantum computing: Most query-efficient quantum algorithms can be realized as time-efficient algorithms and vice versa. While we draw heavily on intuition from quantum query complexity in this survey, we do not require any details of the quantum query model beyond the (limited and vague) description given above.

Beals et al. [BBC$^+$01] proved a result that is central to our understanding of quantum query complexity. They showed that for any quantum query algorithm for f making at most T queries on every input, there is a polynomial $p\colon \{-1,1\}^n \to \mathbb{R}$ of total degree at most $2T$ such that for all $x \in \{-1,1\}^n$, $p(x)$ equals the probability that the algorithm outputs -1 when run on input x.

Since this probability is at least $2/3$ when $f(x) = -1$ and at most $1/3$ when $f(x) = 1$, it follows that $1 - 2p(x)$ approximates f to error $2/3$. In other words, if f is computed by a quantum query algorithm of cost at most T, then the $(2/3)$-approximate degree of f is at most $2T$.

This has two (equivalent) implications. The first is relevant to deriving approximate degree upper bounds: if one can give a T-query quantum algorithm for f, then one can conclude that the approximate degree of f is *at most* $O(T)$. The second is relevant to deriving quantum query lower bounds: if one can show that $\widetilde{\deg}(f) \geq d$, then the quantum query complexity of f is *at least* $\Omega(d)$ (see Theorem 70 in Section 10.1).

4.2 Upper Bounds from Quantum Algorithms

4.2.1 The vanishing-error approximate degree of OR

Theorem 10 takes any (1/3)-approximation to f and turns it into an ε-approximation with an $O(\log(1/\varepsilon))$-factor blowup in degree. For many functions f, this is not the most efficient way to obtain an ε-approximation. For example, recall from Lemma 7 that the approximate degree of OR is $O(\sqrt{n})$. Applying Theorem 10 to $f = \mathsf{OR}$ would give an upper bound of $O(\sqrt{n} \cdot \log(1/\varepsilon))$. The following result shows that the dependence on $\log(1/\varepsilon)$ can be improved quadratically. We show a matching lower bound in Section 7.3.1.

Theorem 14 (Buhrman, Cleve, de Wolf, Zalka ([BCDWZ99]))**.** For any $\varepsilon \in [3^{-n}, 1)$, $\widetilde{\deg}_{\varepsilon}(\mathsf{OR}_n) = O\left(\sqrt{n \log(1/\varepsilon)}\right)$.

Our preferred construction of an ε-approximating polynomial for OR derives the polynomial from an ε-error quantum query algorithm for OR due to [BCDWZ99]. To understand this algorithm, one need not know any details of the quantum query complexity model. All one needs to know is the following fact about (a variant of) Grover's search algorithm.

Fact 15. For any integer $\ell > 0$, there is a quantum query algorithm Grover_ℓ making $O(\sqrt{n/\ell})$ queries to x such that

- If x has Hamming weight 0, the algorithm rejects with probability 1.
- If x has Hamming weight exactly ℓ, the algorithm accepts with probability 1.
- If $\ell < |x| \leq n$, the algorithm accepts with probability at least 2/3.

That is, Grover_ℓ computes $\mathsf{OR}(x)$ with error probability at most 1/3 under the promise that its input has Hamming weight equal to 0 or at least ℓ. And moreover, its error probability is *zero* on the input of Hamming weight 0 or any input of Hamming weight exactly ℓ.

For the reader interested in the underlying quantum algorithms, Fact 15 follows by combining two standard variants of Grover search from [BHT98, BHMT02] that each individually makes $O(\sqrt{n/\ell})$ queries. The first is an "exact" version of search which identifies a marked item in an unsorted list of n elements with probability 1 under the promise that exactly ℓ items are marked. The second is a "promise" version of search that finds such a marked element with probability 2/3 under the promise that at least ℓ items are marked. The guarantee of Fact 15 follows by running the exact version of search, and if it fails, running the promise version of search.

With these facts in hand, we now describe the ε-error quantum algorithm for OR of [BCDWZ99].

Let $t = \log(1/\varepsilon)$. For $i = 1, 2, \ldots, t$, run Grover_i on x and halt and accept if Grover_i accepts. Accept if any run accepts; otherwise reject.

Error Analysis. The above algorithm rejects with probability 1 when run on the input of Hamming weight 0 because each call to Grover_i for $i = 1, \ldots, t$ rejects with probability 1 when run on this input. If run on an input x of Hamming weight between 1 and t, the algorithm accepts with probability 1, because $\text{Grover}_{|x|}$ will be called on x and will accept with probability 1. Finally, if run on an input x of Hamming weight more than t, the algorithm will accept with probability at least $1 - (1/3)^t \geq 1 - \varepsilon$, because each call to Grover_i will independently accept with probability at least 2/3.

Query Cost. The number of queries the algorithm makes to the input is

$$\sum_{i=1}^{t} O\left(\sqrt{n/i}\right) \leq O\left(\sqrt{n \log\left(1/\varepsilon\right)}\right).$$

The inequality here follows from the fact that $\sum_{i=1}^{t} 1/\sqrt{i} \leq 1 + \int_1^t 1/\sqrt{x}\, dx = 1 + 2t^{1/2} - 1 = O(t^{1/2})$.

By the seminal result of Beals et al. [BBC+01] that the ε-error quantum query complexity of any function f is, up to a constant factor, at least its ε-approximate degree, we may immediately conclude from the above algorithm that $\widetilde{\deg}_\varepsilon(\mathsf{OR}_n) \leq O(\sqrt{n \log(1/\varepsilon)})$. However, we find it instructive to explicitly construct the ε-approximating polynomial, below.

Turning the algorithm into a polynomial. Let p_i be the univariate polynomial of degree $O(\sqrt{n/i})$ corresponding to the acceptance probability of Grover_i, i.e., p_i has the following four properties.

- $p_i(0) = 0$.
- $p_i(i) = 1$.
- $p_i(r) \in [2/3, 1]$ for $r \in \{i, i+1, \ldots, n\}$,
- $p_i(r) \in [0, 1]$ for all $r \in \{0, 1, \ldots, n\}$.

The polynomial p_i can be constructed explicitly as follows. The construction of Section 3.2 (after a suitable affine transformation) yields a univariate polynomial Q_n of degree $O(\sqrt{n})$ such that $Q_n(0) = 0$, $Q_n(1) = 1$, $Q_n(k) \in [2/3, 1]$ for all $k \in [1, n]$, and $|Q_n(k)| \leq 1$ for all $k \in [0, n]$. Defining

$$p_i(k) \coloneqq Q_{n/i}(k/i) \tag{10}$$

ensures that p_i has degree $O(\sqrt{n/i})$ and has the four properties above.

Consider

$$p \coloneqq p_1 + p_2 \cdot (1 - p_1) + p_3 \cdot (1 - p_2) \cdot (1 - p_1) + \cdots + p_t \cdot (1 - p_{t-1}) \cdot (1 - p_{t-2}) \cdot \ldots \cdot (1 - p_2) \cdot (1 - p_1)$$

where $t = \log(2/\varepsilon)$.

Intuitively, this captures the acceptance probability of the algorithm that, like [BCDWZ99], first runs Grover_1, and if it returns -1 then it halts and outputs -1, and if not it runs Grover_2 and it halts and outputs -1, and if not it runs Grover_3, and so forth up to Grover_t.

Another way of expressing p is as $p = 1 - \prod_{i \in [t]} (1 - p_i)$. Using this expression, it is easy to check that:

- $p(0) = 0$.
- $p(r) = 1$ for $r \in \{1, \ldots, t\}$. Indeed, $p_r(r) = 1$ implies that $\prod_{i \in [t]} (1 - p_i(r)) = 0$, so $p(r) = 1$.
- $p(r) \in [1 - \varepsilon/2, 1]$ for $r \in \{t+1, \ldots, n\}$. That $p(r) \geq 1 - \varepsilon/2$ follows from the fact that $p_i(r) \geq 2/3$ for every $i \in \{1, \ldots, t\}$, and hence $p(r) \geq 1 - \left(\frac{1}{3}\right)^t \geq 1 - \varepsilon/2$.

  That $p(r) \leq 1$ follows from the fact that $p_i(r) \in [0, 1]$ for every i, and hence $\prod_{i \in [t]} (1 - p_i(r)) \in [0, 1]$.

The degree of p is $\sum_{i=1}^{t} O\left(\sqrt{n/i}\right) = O\left(\sqrt{n \cdot t}\right) = O\left(\sqrt{n \log(1/\varepsilon)}\right)$. Hence, $-1 + 2p(|x|)$ is an ε-approximation to OR_n of degree $O\left(\sqrt{n \log(1/\varepsilon)}\right)$.

Additional discussion of error reduction. As previously discussed, Theorem 14 shows that the generic error reduction result of Theorem 10 is not tight for OR, in that it yields a quadratically suboptimal dependence on $\log(1/\varepsilon)$. In fact, it is not known whether Theorem 10 is tight for *any* total Boolean function. As far as we know, it could even be the case that for every total function f of approximate degree at most $O(n^c)$ for constant $c \in (0,1)$, the ε-approximate degree of f is at most $O(n^c \cdot \log(1/\varepsilon)^{1-c})$.

Open Problem 16. For some constant $c \in (0,1)$ exhibit a total Boolean function f with ε-approximate degree $\Theta(\min\{n^c \log(1/\varepsilon), n\})$ for all $\varepsilon \in [2^{-n}, 1/3]$, or show that no such function exists.

4.3 Consequences of the Vanishing-Error Upper Bound for OR and AND

Theorem 17. Let $f\colon \{-1,1\}^n \to \{-1,1\}$ be any Boolean function with $|f^{-1}(-1)| \leq T$, where $f^{-1}(-1) = \{x\colon f(x) = -1\}$. Then $\widetilde{\deg}_\varepsilon(f) \leq O\left(\sqrt{n \log T} + \sqrt{n \log(1/\varepsilon)}\right)$.

Proof. First, we represent f as a sum of T conjunctions as follows. For each input $y \in \{-1,1\}^n$, let $\mathsf{EQ}_y(x)\colon \{-1,1\}^n \to \{0,1\}$ denote the function that evaluates to 1 if and only if $x = y$. Then

$$f(x) = 1 - 2\sum_{y \in T} \mathrm{EQ}_y(x).$$

Note that $\mathrm{EQ}_y(x)$ can be computed by composing (the negation of) OR with the degree-1 function $x_i \mapsto y_i x_i$.

Next, for each $y \in T$, let p_y denote a polynomial that approximates $\mathrm{EQ}_y(x)$ to error ε/T (this can be done with degree $O(\sqrt{n \log(T/\varepsilon)})$ by Theorem 14), and define $p(x) = 1 - 2\sum_{y \in T} p_y(x)$. By the triangle inequality, p approximates f to error at most $T \cdot (\varepsilon/T) = \varepsilon$. □

Consequences for symmetric functions. Suppose $f\colon \{-1,1\}^n \to \{-1,1\}$ is a symmetric function that is constant at all inputs of Hamming weight between t and $n - t$. Then either f satisfies

$$|f^{-1}(-1)| \leq 2\sum_{i=0}^{t} \binom{n}{t} \leq 2n^t \tag{11}$$

or the negation of f does. If f satisfies Equation (11) then Theorem 17 implies that $\widetilde{\deg}_\varepsilon(f) \leq O(\sqrt{nt \log n} + \sqrt{n \log(1/\varepsilon)})$. This result is tight up to a $\sqrt{\log n}$ factor (the tight bound is $O(\sqrt{nt} + \sqrt{n \log(1/\varepsilon)})$ [dW08]). If the negation of f satisfies Equation (11) then we can draw the same conclusion, as approximate degree is preserved under negation.

The results above approximate f by expressing f as a sum of conjunctions, approximates each term of the sum to exponentially small error, and sums the approximations. The same technique can be used to prove that the approximate degree of the Surjectivity function, SURJ, is $\tilde{O}(n^{3/4})$. We defer coverage of this result to Theorem 53 in Section 8.1.1. The technique arises yet again in the $\tilde{O}(n^{2/3})$ upper bound for Element Distinctness that we prove shortly, in Section 4.4.2.

4.4 More Algorithmically Inspired Polynomials

4.4.1 Collision and PTP

Recall that the Collision and Permutation Testing (PTP) problems for domain size N and range size R were defined in Section 2.2. For example, in the PTP problem, the goal is to distinguish the case that the input list is a permutation, from the case that the list is far from any permutation.

The following quantum algorithm for the Collision and PTP problems, due to Brassard, Hoyer, and Tapp [BHT97], solves both problems with query cost $O(N^{1/3})$. For an input list $k = (k_1, \ldots, k_N) \in [R]^N$, we describe the algorithm assuming it can learn any k_i with a single query; if the list is actually specified via a bit-vector $x \in \left(\{-1,1\}^{\log R}\right)^N$ as per Section 2.2, the query cost would increase by a $\log R$ factor, as each k_i can be learned with $\log R$ queries to bits of x.

For an input list $k = (k_1, \ldots, k_N) \in [R]^N$, randomly sample a set S of $N^{1/3}$ list elements. Let $k|_S$ denote the sampled list elements and $k|_{\bar{S}}$ denote the unsampled list elements. If $k|_S$ contains a collision, i.e., $k_i = k_j$ for distinct $i, j \in S$, then halt and accept. Otherwise, let $\mathcal{R} \subseteq [R]$ denote the range elements appearing in the sample, noting that $\mathcal{R}$ is now known to the algorithm. Apply Grover_ℓ (see Fact 15) to $k|_{\bar{S}}$ with $\ell = |S|/100$ to search for an element of $\mathcal{R}$ appearing in $k|_{\bar{S}}$, accepting if such an element is found, and rejecting otherwise.

Query cost. The number of queries made by the algorithm is $|S|$, plus the number of queries made by Grover_ℓ. In total, this is $O(|S| + \sqrt{N/|S|}) = O(N^{1/3})$ queries.

Correctness analysis. For simplicity, we analyze the algorithm's correctness for the Collision problem, though a nearly identical analysis applies to PTP. If the input list k is 1-to-1, then the algorithm will never find a collision and so it rejects with probability 1. If the input list k is 2-to-1, then the algorithm accepts with probability at least 2/3. This is because, if k is a 2-to-1 input, then either $k|_S$ contains a collision or else there are $|S| > |S|/100$ "cross-collisions" (i.e., $j \notin S$ such that $k_j = k_i$ for at least one $k_i \in S$).

Explicit construction of the resulting polynomials. It is straightforward to translate this quantum query algorithm into an explicit polynomial of degree $O(N^{1/3} \cdot \log R)$ approximating the Collision and PTP problems. For an input $x \in \{-1,1\}^N$, write $x = (x_1, \ldots, x_N) \in (\{-1,1\}^{\log R})^N$. For a set $S \subseteq [N]$, let $x|_S$ denote the subvector $(x_i \colon i \in S) \in (\{-1,1\}^{\log R})^{|S|}$. For a vector of range elements $r = (r_1, \ldots, r_{|S|}) \in [R]^{N^{1/3}}$, let $\mathsf{EQ}_r(x|_S)$ denote that function that equals 1 if $x|_S$ is the binary representation of $r_1, \ldots, r_{N^{1/3}}$. Note that $\mathsf{EQ}_r(x|_S)$ is computed exactly by a polynomial of degree at most $|S| \log R$. Finally, let $\mathsf{inr}_r(x|_{\bar{S}}) \in \{-1,1\}^{N-|S|}$ be the bit-vector with entries indexed by $[N] \setminus S$ such that $(\mathsf{inr}_r(x|_{\bar{S}}))_i$ equals -1 if and only if there is some $j \in [|S|]$ with $x_i = r_j$. Note that each entry of $\mathsf{inr}_r(x|_{\bar{S}})$ depends on only $\log R$ bits of $x|_{\bar{S}}$, and hence is exactly computed by a polynomial of degree $\log R$.

Then consider the polynomial:

$$\frac{1}{\binom{N}{N^{1/3}}} \sum_{S \subseteq [N] \colon |S| = N^{1/3}} p_S(x) + q_S(x), \tag{12}$$

where

$$p_S(x) = \sum_{r=(r_1,\ldots,r_{N^{1/3}})\in[R]^{N^{1/3}}:\ r_i=r_j \text{ for some } i\neq j} \mathsf{EQ}_r(x|_S)$$

and

$$q_S(x) = \sum_{r=(r_1,\ldots,r_{N^{1/3}})\in[R]^{N^{1/3}}:\ \text{all } r_i \text{ are distinct}} \mathsf{EQ}_r(x|_S)\cdot p_\ell(\mathsf{inr}_r\,(x|_{\bar{S}})).$$

Here, p_ℓ is the polynomial capturing the acceptance probability of Grover_ℓ given in Equation (10) of Section 4.2.1.

The polynomial in Equation (12) exactly computes the acceptance probability of the quantum algorithm above. Intuitively, p_S outputs 1 if the sample $x|_S$ contains a collision and 0 otherwise, while $q_S(x)$ outputs the probability that the sample $x|_S$ does not contain a collision, yet Grover search finds a "cross-collision", i.e., an item outside of the sample that collides with an item in the sample. Equation (12) outputs the average of these acceptance probabilities over the random choice of S.

4.4.2 Element Distinctness

Recall from Section 2.2 that the k-distinctness function k-ED (for constant k) interprets its input as a list of N numbers $(k_1,\ldots,k_N)$ from a range of size R and outputs -1 if and only if there is some range item that appears at least k times in the list. The special case $k=2$ is a particularly natural and is referred to as Element Distinctness, or ED for short. Much later in the survey, Section 8.6 establishes a lower bound of $\Omega(N^{2/3})$ on the approximate degree of ED when $R\geq N$. In this section, we sketch a matching upper bound.

The upper bound has been known since the discovery of an $O(N^{2/3})$-query quantum algorithm due to Ambainis [Amb07] (recall that Theorem 70 translates any quantum query algorithm for a function f into an approximating polynomial for f). But an explicit description of an approximating polynomial was not given until work of Sherstov [She18a]. We cover Sherstov's construction, though our presentation is quite different than the treatment in [She18a].

For illustration, we start by giving a simple quantum algorithm for the negation of ED that makes $O(N^{5/6})$ queries. This implies the existence of an approximating polynomial for ED of degree $O(N^{5/6})$. We then explain how to lower the degree bound to the optimal $O(N^{2/3})$.

The $O(N^{5/6})$-query algorithm utilizes a subroutine with the following property. Say that a list item k_i is involved in a collision if there is some $j\neq i$ such that $k_j=k_i$. If there is any list item involved in a collision, the subroutine will find the collision with probability at least $1/N^{2/3}$.

> SUBROUTINE 1: Randomly sample $b=N^{1/3}$ inputs, then determine if any sampled item is involved in a collision. That is, randomly pick a set $S\subseteq[N]$ of size b and query $\{k_i\colon i\in S\}$. Let $\{r_1,\ldots,r_b\}$ denote the multiset of sampled range items. If there exist distinct $i,j\in[b]$ such that $r_i=r_j$, halt and accept. Otherwise, Grover-search over the un-sampled items $\{k_i\colon i\notin S\}$ for an appearance of one of $r_1,\ldots,r_b$, accepting if Grover search finds such an appearance, and rejecting otherwise. This costs $O(b+\sqrt{N})$ queries (b queries for the sample, $O(\sqrt{N})$ for the Grover search).

If there is one or more k_i's involved in a collision, the probability that at least one such k_i is sampled by SUBROUTINE 1 is at least $b/N\geq\Omega(1/N^{2/3})$, and conditioned on this event, the

probability that the subroutine accepts is at least $2/3$. Meanwhile, if there is no k_i involved in a collision, then obviously the probability that SUBROUTINE 1 accepts is 0. Hence, by Theorem 70, SUBROUTINE 1 can be transformed into a polynomial p of degree at most $O(b + \sqrt{N})$ such that $p(x) = 0$ for all $x \in (\mathsf{ED})^{-1}(-1)$ while $p(x) \geq b/N$ for all $x \in (\mathsf{ED})^{-1}(1)$.

Using a technique in quantum algorithm design called amplitude amplification, one can boost the success probability from $\Omega(b/N)$ to $2/3$ while increasing the degree by a factor of $O\left(\sqrt{N/b}\right) = O\left(N^{1/3}\right)$. This yields a query upper bound of $O(N^{1/2} \cdot N^{1/3}) = O(N^{5/6})$. The following theorem captures the analog of amplitude amplification in the context of approximate degree, yielding the claimed $O(N^{5/6})$ approximate degree upper bound for ED.

Theorem 18. Let $f\colon \{-1,1\}^n \to \{-1,1\}$. Suppose that there is a polynomial p of degree at most d, such that $p(x) = 0$ for all $x \in f^{-1}(-1)$ and $\delta \leq p(x) \leq 1$ for all $x \in f^{-1}(1)$. Then the $(1/3)$-approximate degree of f is at most $O\left(d \cdot \sqrt{1/\delta}\right)$.

Proof. Using Chebyshev polynomials (cf. Section 3.2), we can construct a univariate polynomial A of degree $O\left(\sqrt{1/\delta}\right)$ such that $A(0) = 1$ and $A(x) \in [-4/3, -2/3]$ for all $x \in [-1+\delta, 1]$. Then $-A(p(x))$ approximates f to error $1/3$ and has degree $d \cdot \deg(A) \leq O(d \cdot \sqrt{1/\delta})$. □

In summary, we have identified a polynomial approximating ED of the form $-A \circ p$, where A is the polynomial of degree $O(N^{1/3})$ appearing in the proof of Theorem 18 and $p(x)$ is the acceptance probability of SUBROUTINE 1 on input x.

A first attempt at improvement. How can one improve the above approximate degree upper bound of $O(N^{5/6})$? First, let us replace the invocation of Grover search in SUBROUTINE 1, which has query cost only $O(\sqrt{N})$ but has non-zero error probability, with an *exact* search (exact search has trivial query cost $\Omega(N)$, but failure probability zero). Let q denote the acceptance probability of this algorithm. Then just as for $-A \circ p$, $-A \circ q$ is a $(1/3)$-approximation to ED, though its degree is much larger than that of $-A \circ p$.

What actually works. The key idea is that, owing to its composed structure, $A \circ q$ itself can be approximated pointwise by a polynomial of degree just $O(N^{2/3})$, despite the fact that q itself has degree $\Omega(N)$. Since $-A \circ q$ approximates ED, any sufficiently accurate pointwise approximation to $-A \circ q$ will itself approximate ED.

To approximate $A \circ q$, the idea is to express $A \circ q$ as a weighted sum of $2^{O(\deg(A))}$ terms, with each term equal to a power of $(1-q)$, i.e., of the form $c_i \cdot (1 - q(x))^i$ for some $c_i \in \mathbb{R}$ and integer $i \leq \deg(A)$. We will then approximate each term to exponentially small error, ensuring that the sum of the term-by-term approximations is a good approximation to $A \circ q$.

For example, if $A(t) = 2t^3 - 4t^2 + 3t - 1$, then

$$(A \circ q)(x) = 2q(x)^3 - 4q(x)^2 + 3q(x) - 1 = -2(1-q(x))^3 + 2(1-q(x))^2 - (1-q(x)).$$

If we approximate the three terms in this expression, namely $-2(1-q(x))^3$, $2(1-q(x))^2$, and $-(1-q(x))$, individually to error $1/12$ and sum the approximations, we obtain a polynomial Q such that $|Q(x) - (A \circ q)(x)| \leq 3 \cdot 1/12 = 1/4$ for all x in the domain of ED. Note that this technique is identical to that used to approximate any symmetric function (Theorem 17) and to approximate SURJ (Theorem 53). Details follow.

As the polynomial A from Theorem 18 is derived from the Chebyshev polynomial of degree $\deg(A)$, it can be shown that the ℓ_1-norm of the coefficients of A is $2^{O(\deg(A))}$. That is,

$$A(x) = \sum_{i=0}^{\deg(A)} c_i \cdot (1 - q(x))^i$$

for some real numbers $c_0, \ldots, c_{\deg(A)}$ such that

$$\sum_{i=0}^{\deg(A)} |c_i| \leq 2^{O(\deg(A))}.$$

Hence, it suffices to approximate each power $(1 - q(x))^i$ of $q(x)$ to error $\varepsilon = 2^{-\Theta(\deg(A))}$. Summing the approximations yields a polynomial Q such that $|Q(x) - (A \circ q)(x)| \leq \varepsilon \cdot 2^{O(\deg(A))} \leq o(1)$ for all inputs x to ED.

What degree suffices to approximate $(1 - q(x))^i$ to error ε? Since q specifies the acceptance probability of SUBROUTINE 1 (modified to use exact search rather than Grover search), the quantity

$$(1 - q(x))^i$$

captures the probability that i independent runs of SUBROUTINE 1 all output reject. Here, running $i \leq \deg(A)$ independent instances of SUBROUTINE 1 corresponds to taking i independent subsamples of the input list, with each subsample having size b, and then for each subsample (if there is no collision within the subsample), determining whether any range item from the sample appears elsewhere in the input list.

The key insight is that the above probability is unchanged if we do not run $\deg(A)$ independent searches, one for each subsample of size b, but rather run a *single* search, looking for a second occurrence of any range item in the "combined" subsample of size $b \cdot \deg(A)$. If the single search procedure fails to find a second occurrence of a range item from the combined sample, this is equivalent to the failure of all $\deg(A)$ independent searches, one for each subsample.

Performing the single search procedure using ε-error Grover search (Section 4.2.1), the acceptance probability of this algorithm can be approximated to error $\varepsilon = 2^{-\Theta(\deg(A))}$ with degree $O\left(b \cdot \deg(A) + \sqrt{n \cdot \log(1/\varepsilon)}\right) = O(N^{2/3})$ as claimed. Here, $b \cdot \deg(A)$ comes from computing $\deg(A)$ subsamples each of size b, and $\sqrt{n \cdot \log(1/\varepsilon)}$ comes from running an ε-error Grover search procedure to determine if any range item from any of the $\deg(A)$ subsamples appears amongst the $N - b \cdot \deg(A)$ un-sampled items in the input list.

4.5 Algorithmically-Inspired Upper Bound for Composed Functions

We now prove a weaker version of Theorem 11, which upper bounds the approximate degree of a composed function $f \circ g$ by the product of the approximate degree of f and the δ'-approximate degree of g where δ' is much lower than in Theorem 11, namely $O(1/m)$ rather than a positive constant. The reason we include this theorem despite its sub-optimality is that it has a clean, algorithmically-inspired proof.

Theorem 19. Let $\varepsilon, \delta > 0$ and suppose that $f\colon \{-1,1\}^m \to \{-1,1\}$ satisfies $\widetilde{\deg}_\varepsilon(f) \leq d$ and $g\colon \{-1,1\}^n \to \{-1,1\}$ satisfies $\widetilde{\deg}_{\delta/m}(g) \leq d'$. Then $\widetilde{\deg}_{\varepsilon+\delta}(f \circ g) \leq d \cdot d'$.

Proof. For purposes of this proof, it will be convenient to treat f as a function mapping $\{0,1\}^m \to \{0,1\}$ rather than mapping $\{-1,1\}^m \to \{-1,1\}$, and accordingly to treat g as function from $\{0,1\}^n$ to $\{0,1\}$.

Let p be a polynomial of degree that approximates f to error ε, i.e., $|p(x) - f(x)| \leq \varepsilon$ for all $x \in \{0,1\}^m$. Similarly, let q be a polynomial that approximates g to error at most δ/m. We claim that $p \circ q$ approximates $f \circ g$ to error at most $\varepsilon + \delta$.

The intuition for why this is true is as follows. Suppose that the polynomials p and q, when evaluated on a Boolean input, output the acceptance probability of randomized query algorithms $\mathcal{A}_f$ and $\mathcal{A}_g$ when run on that input. That is, one can think of $p(x)$ as the probability that $\mathcal{A}_f$ accepts when run on input x, and similarly $q(y)$ as the probability that $\mathcal{A}_g$ accepts when run on input y.[10] The fact that p approximates f to error ε means that, on each input y in $\{0,1\}^m$, $\mathcal{A}_f$ errs when run on y with probability at most ε (i.e., accepts inputs in $f^{-1}(0)$ or rejects inputs in $f^{-1}(1)$), and similarly for $\mathcal{A}_g$.

Let $(x_1, \ldots, x_m) \in (\{0,1\}^n)^m$ denote an input to $f \circ g$, and let $y = (g(x_1), \ldots, g(x_m)) \in \{0,1\}^m$. Then it turns out that $p \circ q$ is the acceptance probability of the natural "composed algorithm" $\mathcal{A}_{f \circ g}$ that runs $\mathcal{A}_f$ on y, and every time $\mathcal{A}_f$ queries a bit y_i of y, it answers the query by running $\mathcal{A}_g$ on x_i. Since $\mathcal{A}_g$ errs with probability at most δ/m, then by a union bound over all (at most m) calls to $\mathcal{A}_g$, with probability at least $1 - \delta$, $\mathcal{A}_g$ returns $g(x_i)$ for every i on which it is called. In this case, the composed algorithm returns the same answer that $\mathcal{A}_f$ returns when run on input $y = (g(x_1), \ldots, g(x_m))$. Since $\mathcal{A}_f$ itself errs with probability at most ε, the error probability of $\mathcal{A}_{f \circ g}$ is at most $\varepsilon + \delta$. Since $(p \circ q)(x_1, \ldots, x_m)$ is the acceptance probability of $\mathcal{A}_{f \circ g}$ on input $(x_1, \ldots, x_m)$, this means that $p \circ q$ approximates $f \circ g$ to error at most $\varepsilon + \delta$.

Here is the formal analysis. Let us assume that $q(x) \in [0,1]$ for all $x \in \{0,1\}^n$ and $p(y) \in [0,1]$ for all $y \in \{0,1\}^m$.[11] Recall that p is a multilinear polynomial. Since for any $(x_1, \ldots, x_m) \in (\{0,1\}^n)^m$, $(q(x_1), \ldots, q(x_m)) \in [0,1]^m$, Fact 20 below implies that $p(x_1, \ldots, x_m)$ equals the expectation of $p(y)$ under the product distribution over $y \in \{0,1\}^m$ in which $y_i = 1$ with probability $q(x_i)$. Since q approximates g to error at most δ/m, the ith coordinate in this product distribution equals $g(x_i)$ with probability at least $1 - \delta/m$. By a union bound of all m coordinates, this product distribution places mass at least $1 - \delta$ on the point $(g(x_1), \ldots, g(x_m))$. Hence, $p(q(x_1), \ldots, q(x_m))$ lies in the interval

$$[p(g(x_1), \ldots, g(x_m)) - \delta, p(g(x_1), \ldots, g(x_m)) + \delta].$$

Since $|p(y) - f(y)| \leq \varepsilon$ for all $y \in \{0,1\}^m$, this implies that

$$|p(q(x_1), \ldots, q(x_m)) - f(g(x_1), \ldots, g(x_m))| \leq \varepsilon + \delta.$$

□

Fact 20. For any multilinear polynomial $q \colon \mathbb{R}^n \to \mathbb{R}$ and point $(u_1, \ldots, u_n) \in [0,1]^n$, $q(u_1, \ldots, u_n)$ equals the expected value of q under the product distribution in which the ith coordinate is chosen to equal 1 with probability u_i and 0 with probability $(1 - u_i)$.

[10] In general, p and q may not actually correspond to the acceptance probability of any low-query algorithm, even if they output values in $[0,1]$ when evaluated at any Boolean input. That is, although for any T-query randomized algorithm, there exists a degree-T polynomial computing its acceptance probability on each input, the converse is not true [Amb06, She18a] (see Section 8.2 for details). Nonetheless, thinking of p and q as if they do compute acceptance probabilities of query algorithms can be a powerful source of intuition regarding their behavior.

[11] In general, a (δ/m)-approximation q to g can take values in $[-\delta/m, 1+\delta/m]$, but the assumption can be ensured by replacing q with $(q(x) + \delta/m)/(1 + 2\delta/m)$. This maintains the degree of q while increasing the error to at most $2\delta/m$. Similarly, p can be assumed to only take values in $[0,1]$ with at most a doubling of its error.

Proof. By linearity of expectation, it suffices to consider a multilinear polynomial q consisting of a single monomial, say, $q(x) = x_1 \cdot x_2 \dots x_i$. In this case, the statement is immediate from the fact that the expectation of the product of independent random variables equals the product of their expectations. □

5 Lower Bounds by Symmetrization

Prior to the development of the method of dual polynomials (Section 6), approximate degree lower bounds were typically proved via a technique called symmetrization. The ethos of this technique is that univariate polynomials are generally easier to understand than multivariate polynomials. Hence, symmetrization seeks to reduce the task of lower bounding the ε-approximate degree of a multivariate function f to a question about univariate polynomials. This is usually done by generically transforming an n-variate polynomial p into a univariate polynomial q without increasing its degree. One then argues that if p satisfies Condition (1), then q exhibits some behavior that forces it to have large degree. Since $\deg(q)$ lower bounds $\deg(p)$, one concludes that p must have large degree as well.

The transformation giving q is often built from a sequence of probability distributions D_t over $\{-1,1\}^n$, where t ranges over a (finite or infinite) subset S of $\mathbb{R}$. One then shows that for any n-variate polynomial p, its *symmetrization* $q(t) = \mathbb{E}_{x \sim D_t}[p(x)]$ is a univariate polynomial of degree at most $\deg(p)$ over $t \in S$. Here are two classic examples.

- (t-biased symmetrization): Let S be the interval $[-1,1]$. For $t \in S$, let μ_t be the distribution over $\{-1,1\}$ with expected value t. Let B_t be the product distribution $\mu_t^{\otimes n}$ on $\{-1,1\}^n$.
- (Minsky–Papert symmetrization): Let $S = [n]^*$. For each $t \in S$, define H_t to be the uniform distribution over $x \in \{-1,1\}^n$ with Hamming weight t (i.e., exactly t entries equal to -1).

The next two lemmas show that both of these classic symmetrization techniques are indeed degree non-increasing maps from n-variate to univariate polynomials.

Lemma 21. For any polynomial $p\colon \{-1,1\}^n \to \mathbb{R}$ of total degree at most d, the univariate function $q(t) = \mathbb{E}_{x \sim B_t}[p(x)]$ is a polynomial of degree at most d over $[-1,1]$.

Proof. By linearity of expectation, it is without loss of generality to consider a polynomial p consisting of a single monomial, e.g., $p(x) = x_1 x_2 \dots x_d$. Then since B_t is a product distribution,

$$\mathbb{E}_{x \sim B_t}[p(x)] = \mathbb{E}_{x_1 \sim \mu_t}[x_1] \cdot \mathbb{E}_{x_2 \sim \mu_t}[x_2] \cdot \ldots \cdot \mathbb{E}_{x_d \sim \mu_t}[x_d] = t^d.$$

□

Lemma 22. For any polynomial $p\colon \{-1,1\}^n \to \mathbb{R}$ of total degree at most d, the univariate function $q(t) = \mathbb{E}_{x \sim H_t}[p(x)]$ is a polynomial of degree at most d over $[n]^*$.

Proof. Again, it is without loss of generality to assume p is a single monomial, e.g., $p(x) = x_1 x_2 \dots x_d$. Applying the variable transformation $x_i \mapsto (1 - x_i)$ does not alter the degree of p, and hence it is also without loss of generality to assume that $p(x) = (1 - x_1) \cdot (1 - x_2) \cdots (1 - x_d)$. In this case, for $x \in \{-1,1\}^n$ we have $p(x) = 2^d$ if $x_1 = x_2 = \cdots = x_d = -1$ and $p(x) = 0$ otherwise.

For any $t \in [n]^*$, the number of n-bit inputs with Hamming weight exactly t is $\binom{n}{t}$, while the number of such inputs that additionally satisfy $x_1 = x_2 = \cdots = x_d = -1$ is $\binom{n-d}{t-d}$. (We are using the convention that if $t - d$ is negative, then $\binom{n-d}{t-d}$ is 0.) It follows that, for any $t \in [n]^*$,

$$\mathbb{E}_{x \sim H_t}[p(x)] = 2^d \cdot \frac{\binom{n-d}{t-d}}{\binom{n}{t}} = \left(2^d \cdot \frac{(n-d)!}{n!}\right) \cdot t(t-1)(t-2)\cdots(t-d+1)$$

or the zero polynomial. Either way, this is a polynomial in t of degree at most d. □

5.1 Symmetrization Lower Bound for OR

The tight $\Omega(\sqrt{n})$ lower bound for the approximate degree of OR_n follows easily from Lemma 21 and Markov's inequality (Theorem 6). In short, the proof applies Lemma 21 to any polynomial p approximating OR_n to derive a univariate polynomial $q(t)$ that is bounded on the whole interval $[-1, 1]$ but has a large "jump" in the vicinity of $t = 1$ (quantitatively, a derivative of $\Omega(n)$). Markov's inequality then implies that q has degree $\Omega(\sqrt{n})$.

Theorem 23. The approximate degree of OR_n is $\Omega(\sqrt{n})$.

Proof. Let p approximate OR_n to error at most $1/3$, and let q be the univariate polynomial whose existence is guaranteed by Lemma 21. Since the distribution B_1 assigns probability 1 to the input $\mathbf{1}_n$, we may conclude that $q(1) = p(\mathbf{1}_n) \in [2/3, 4/3]$.

Now let $t = 1 - 4/n$. Then B_t assigns probability mass at most $(1 - 2/n)^n < 1/e^2$ to $\mathbf{1}_n$, where the inequality holds by Fact 2. Hence,

$$q(1 - 4/n) = \mathbb{E}_{x \sim B_t}[p(x)] \leq (4/3) \cdot 1/e^2 + (-2/3) \cdot (1 - 1/e^2) \leq -1/3.$$

The Mean Value Theorem now implies that there is some $t^* \in (1 - 4/n, 1)$ such that $q'(t^*) \geq n/4$.

Finally, since it approximates OR_n, the polynomial p has magnitude at most $4/3$ over the entire Boolean hypercube $\{-1, 1\}^n$. Hence $q(t) \in [-4/3, 4/3]$ for all $t \in [-1, 1]$ as well. Applying Markov's inequality to $\frac{3}{4}q$ now implies that $\deg(p) \geq \deg(q) \geq \sqrt{3n/16}$. □

An alternative proof using Minsky-Papert Symmtrization. Our proof of Theorem 23 derived the $\Omega(\sqrt{n})$ lower bound for OR_n from t-biased symmetrization (Lemma 21). The bound can also be derived from Minsky-Papert symmetrization (Lemma 22), but the derivation is complicated by the fact that the univariate polynomial $q(t) = \mathbb{E}_{x \sim H_t}[p(x)]$ in Lemma 22 is *not* a priori guaranteed to be bounded in magnitude at inputs non-integer values between 0 and n. This prevents a direct application of Markov's inequality even to affine transformations of q, as Markov's inequality applies only to polynomials bounded over the real interval $[-1, 1]$.

Our preferred way to address this issue is to invoke a result of Coppersmith and Rivlin [CR92], which states that any degree-d polynomial that is bounded at integer inputs in $[0, n]$ cannot be larger than $2^{O(d^2/n)}$ even at non-integer inputs within the same interval (we do not prove this result in this survey).

Theorem 24 (Coppersmith and Rivlin [CR92])**.** For every univariate polynomial q of degree d such that $|q(x)| \leq 1$ for all integers $x \in [0, n]$, we have $|q(x)| < a \cdot 2^{bd^2/n}$ for all real $x \in [0, n]$, where $a, b > 0$ are universal constants.

Combined with Lemma 22, this rules out an approximating polynomial for OR_n of degree $o(\sqrt{n})$. If such a polynomial existed, then Lemma 22 would turn it into a univariate polynomial q of degree $\ll \sqrt{n}$ such that $q(0) \in [2/3, 4/3]$, $q(1) \in [-4/3, -2/3]$, and $|q(t)| \leq 4/3$ for all $t \in [n]^*$. Moreover, since $d^2/n < 1$, Theorem 24 guarantees that there is a universal constant c such that $|q(x)| \leq c$ for all real numbers $x \in [0, n]$. That is, for degrees smaller than $\sqrt{n}$, boundedness at integer inputs implies boundedness even at non-integer inputs in the same interval. Hence, letting $C = \max\{c, 4/3\}$, $Q(x) \coloneqq (1/C) \cdot q((1-x)/2)$ is a univariate polynomial of degree $\ll \sqrt{n}$ such that $Q'(x) \geq \Omega(n)$ for some $x \in [1 - 2/n, 1]$ (here, we are invoking the Mean Value Theorem just as in the proof of Theorem 23). Moreover, $|Q(x)| \leq 1$ for all real numbers $x \in [-1, 1]$. But this contradicts Markov's inequality.

Application: Optimality of Grover's search algorithm. Recall from Section 4.1 that an approximate degree lower bound for f implies a lower bound on the quantum query complexity of f. Theorem 23 thus implies that the quantum query complexity of OR_n is $\Omega(\sqrt{n})$, matching the upper bound achievable via Grover's search algorithm (Section 4.2.1). Equivalently, to quote the most recent tagline of Scott Aaronson's blog, "quantum computers need $\sim \sqrt{n}$ queries to search a list of size n." While this did not give the first tight quantum query lower bound proof for OR_n—it was first proved by Bennett et al. [BBBV97] using different techniques—the proof via approximate degree has other consequences in quantum complexity. We will see later (Section 10) that approximate degree lower bounds extend in a black-box manner from quantum query to quantum communication lower bounds [She11a, SZ09]. In particular, the version due to [She11a] transfers the lower bound for OR_n to a tight $\Omega(\sqrt{n})$ lower bound on the quantum communication complexity of the Disjointness function, DISJ, an important result (first proved by Razborov [Raz03]) that is not known to follow from other techniques for lower bounding quantum query complexity. We cover this result in this survey (Corollary 82 in Section 10.4.3).

5.2 Arbitrary Symmetric Functions

Threshold degree lower bound via Minsky-Papert symmetrization. Let $f(x)\colon \{-1,1\}^n \to \{-1,1\}$ be a symmetric Boolean function such that $f(x) = F(|x|)$ for all $x \in \{-1,1\}^n$. Suppose that $F(i-1) \neq F(i)$ for k values of $i \in [n]$. We saw in Section 3.1 that $\deg_{\pm}(F) \leq k$. Here we prove a matching lower bound. Let p be a polynomial that sign-represents f and let q be the Minsky-Papert symmetrization of p as per Lemma 22. Then q is a univariate polyonmial of degree at most $\deg(p)$, and $q(i-1) \neq q(i)$ for k values of $i \in [n]$. If $k > 0$, this means that q is a non-zero polynomial with at least k roots, and hence $\deg(q) \geq k$. Since $\deg(q) \leq \deg(p)$, we conclude that the sign-representation p for f has degree at least k.

Approximate degree lower bound via Minsky-Papert symmetrization. Recall from Section 4.3 that if f is a symmetric function that is constant on all inputs of Hamming weight between t and $n-t$ with $t \leq n/2$, then $\widetilde{\deg}(f) = O(\sqrt{nt})$. Here, we sketch a proof of a matching lower bound. For simplicity, let us focus on the symmetric t-threshold function on n bits, $\mathsf{THR}_n^t\colon \{-1,1\}^n \to \{-1,1\}$, for which $\mathsf{THR}_n^t(x) = -1$ if and only if $|x| \geq t$. The lower bound $\widetilde{\deg}(\mathsf{THR}_n^t) \geq \Omega(\sqrt{tn})$ was originally proved by Paturi [Pat92] using symmetrization, and a variant of Markov's inequality (Theorem 6) called the Markov-Bernstein inequality. Our sketch of Paturi's analysis in this section omits some technical details. We give a different proof of the result via the

method of dual polynomials in Section 7.3.2.

Theorem 25 (Markov-Bernstein inequality)**.** Let g be a univariate polynomial of degree at most d such that $|g(x)| \leq 1$ for all $x \in [-1,1]$. Then $|g'(x)| \leq O(d/\sqrt{1-x^2})$ for all $x \in [-1,1]$ such that $\sqrt{1-x^2} \geq 1/d$.

The idea of Paturi's analysis is the following. Let p be a polynomial approximation to THR_n^t of degree d, and let q be its Minsky-Papert symmetrization (Lemma 22). Then $Q(\ell) \coloneqq q(n(\ell+1)/2)$ is a univariate polynomial of degree at most d and satisfies

$$|Q(\ell)| \leq 1 \text{ for all } \ell \in [-1,1] \text{ that are integer multiplies of } 2/n. \tag{13}$$

Moreover, $Q(-1+2(t-1)/n) \in [2/3, 4/3]$, while $Q(-1+2t/n) \in [-4/3, -2/3]$, and hence there is some $x^* \in [-1+2t/n-2/n, -1+2t/n]$ with $Q'(x^*) \geq n/2$.

If Equation (13) actually guaranteed $|Q(\ell)| \leq 1$ for *all real numbers* $\ell \in [-1,1]$, we could apply Theorem 25 to conclude that $d \geq \Omega(\sqrt{nt})$. To see this, observe that:

$$d/\sqrt{1-(x^*)^2} = d/\sqrt{1-(1-O(t/n))^2} = d/\sqrt{O(t/n)} = O\left(d\sqrt{t/n}\right).$$

Since $Q'(x^*) \geq n/2$, d must be large enough to ensure that $d\sqrt{t/n} \geq \Omega(n)$, and hence $d \geq \Omega\left(\sqrt{nt}\right)$.

In the case that $d \leq \sqrt{n}$, Coppersmith-Rivlin's result (Theorem 24) implies that Equation (13) in fact implies that $|Q(\ell)| \leq 2^{O(d^2/n)} = O(1)$ for all real numbers $\ell \in [-1,1]$. However, for larger values of d, Equation (13) does *not* guarantee that $|Q(\ell)| \leq 1$ for all real numbers $\ell \in [-1,1]$. So Paturi's proof involves a tricky case analysis to handle the situation when $Q(\ell)$ may be extremely large at inputs in $[-1,1]$ that are not integer multiples of $2/n$.

5.3 Threshold Degree Lower Bound for the Minsky-Papert CNF

Recall that in Lemma 8 and Lemma 9, we established two different upper bounds on the threshold degree of the Minsky-Papert CNF, $\mathsf{AND}_m \circ \mathsf{OR}_b$, one based on Chebyshev approximations to OR_b and one based on rational approximations. Minsky and Papert [MP69] gave a classic symmetrization argument showing that one of these approximation techniques is always optimal, at least up to an $O(\sqrt{\log m})$ factor.

Theorem 26. $\deg_{\pm}(\mathsf{AND}_m \circ \mathsf{OR}_b) \geq \Omega(\min\{m, b^{1/2}\})$.

At a high level, Minsky and Papert's proof works as follows. They use a generalization of Lemma 22 (Lemma 27 below) to show that if p sign-represents $\mathsf{AND}_m \circ \mathsf{OR}_{4m^2}$, then there exist a polynomial $q : ([4m^2]^*)^m \to \mathbb{R}$ such that $q(t_1, \ldots, t_m) > 0$ iff $t_i = 0$ for some index i. The polynomial q can then be symmetrized once again (while at most doubling its degree) into a univariate polynomial $\tilde{q} : [2m]^* \to \mathbb{R}$ that changes sign m times as its input increases from 0 to $2m$. Such a polynomial requires degree at least m, so (up to a factor of 2) q does as well, and hence so does the original polynomial p. Details follow.

Proof of Theorem 26. We need the following generalization of Lemma 22.

Lemma 27. Let $p\colon \mathbb{R}^{m \cdot b} \to \mathbb{R}$ be any polynomial of total degree most d. Then there is a degree d polynomial $q\colon \mathbb{R}^m \to \mathbb{R}$ such that the following holds. Let $x = (x_1, \ldots, x_m)$ denote an arbitrary input in $(\{-1,1\}^b)^m$. For all integers $\ell_1, \ldots, \ell_m \in [b]^*$,

$$q(\ell_1, \ldots, \ell_m) = \mathbb{E}_{x=(x_1,\ldots,x_m)\colon |x_i|=\ell_i \text{ for } i=1,\ldots,m}[p(x)].$$

Proof. By linearity, it suffices to prove the lemma when $p = \prod_{i=1}^{m} p_i(x_i)$ where $\sum_{i=1}^{m} \deg(p_i) \leq d$. By Lemma 22, for each p_i there is some univariate polynomial q_i of degree at most $\deg(p_i)$ such that for all $\ell_i \in [b]^*$, $q(\ell_i) = \mathbb{E}_{x_i \in \{-1,1\}^b \colon |x_i| = \ell_i}[p_i(x_i)]$. Hence, we can let $q(\ell_1, \ldots, \ell_m) = \prod_{i=1}^{m} q_i(\ell_i)$. □

Without loss of generality, let us assume that $4m^2 \leq b$ (if not, then apply the argument to follow to $\mathsf{AND}_{m'} \circ \mathsf{OR}_b$ for $m' \coloneqq \lceil (b/4)^{1/2} \rceil$, which is a subfunction of $\mathsf{AND}_m \circ \mathsf{OR}_b$). We prove a lower bound of m on the threshold degree of $\mathsf{AND}_m \circ \mathsf{OR}_b$.

Let p be a sign-representing polynomial for $\mathsf{AND}_m \circ \mathsf{OR}_b$ and q be the m-variate polynomial whose existence is guaranteed by applying Lemma 27 to p. Consider the univariate polynomial

$$\tilde{q}(t) = q\left((t-0)^2, (t-2)^2, (t-4)^2, \ldots, (t-2(m-1))^2\right). \tag{14}$$

Clearly, $\deg(\tilde{q}) \leq 2\deg(q) \leq 2\deg(p)$. Observe that if t is an integer in the set $[2m-1]^* = \{0, 1, \ldots, 2m-1\}$, then all inputs to q on the right hand side of Equation (14) are in the set $[b]^*$.

If t is odd, then each of the integers fed into q, namely $(t-2i)^2$ for $i = 0, 1, \ldots, m-1$, is equal to the square of a non-zero integer, and hence *strictly* greater than 0. Since q was obtained by applying Lemma 27 to a polynomial p that sign-represents $\mathsf{AND}_m \circ \mathsf{OR}_b$, it follows that $\tilde{q}(t) < 0$ for odd values of t in $[2m-1]^*$.

Meanwhile, if $t = 2i$ is an even integer in $[2m-1]^*$, then there is one value fed into q on the right hand side of Equation (14) that equals 0 (namely the value $(t-2i)^2 = 0$). Since q was obtained by applying Lemma 27 to a polynomial p that sign-represents $\mathsf{AND}_t \circ \mathsf{OR}_b$, this implies that $\tilde{q}(t) > 0$ for even values of t in $[2m-1]^*$.

The above means that $\tilde{q}$ is a polynomial that changes sign at least $2m-1$ times as its input increases from 0 to $2m-1$. It follows that $\tilde{q}$ has at least $2m-1$ zeros. Since $\tilde{q}$ is not constant, it must have degree at least $2m-1$. Since $\deg(\tilde{q}) \leq 2\deg(p)$, p has degree at least m. □

Preview: Applications to learning, circuits, and communication. In a 1969 book,[12] Minsky and Papert showed that their CNF has threshold degree $\Omega(n^{1/3})$, and also that the Parity function has threshold degree n (this follows from Section 5.2). Equivalently, polynomial threshold functions (introduced in Section 2.1) require degree $\Omega(n^{1/3})$ and $\Omega(n)$ even to compute functions as simple as CNFs and Parity.

Polynomial threshold functions can be thought of as very simple (depth-one or depth-two) neural networks. Hence, these results were interpreted by some researchers as establishing very strong limitations on the expressiveness of neural networks, and are often said to have helped bring about the "first AI winter," which was a sharp decline in research on neural networks in the 1970s.[13]

Despite the limitations on their expressiveness discussed above, low-degree polynomial threshold functions still have applications in computational learning theory. As we discuss in Section 11.2, if $\mathcal{C}$ is a class of functions, all of which have low threshold degree, then there is a learning algorithm for $\mathcal{C}$ in the so-called Probability Approximately Correctly (PAC) model that runs in time exponential in d. Klivans and Servedio [KS04] showed that the threshold degree of *any* polynomial size CNF is at most $\tilde{O}(n^{1/3})$, thereby obtaining an algorithm for PAC-learning CNF formulas in time exponential

[12] Minksy and Papert's book was titled *Perceptrons*, a historic term synonymous with halfspaces, also known as linear threshold functions.

[13] Of course, polynomial threshold functions are an extremely simple kind of neural network—deeper networks (that also differ from Perceptrons in other ways, e.g., by outputting real numbers rather than merely Boolean values) are more expressive and have proven central to progress in machine learning in recent years.

in $n^{1/3}$. This remains the fastest known algorithm for this problem. Up to logarithmic factors, the Minsky–Papert CNF thus has the largest possible threshold degree amongst all CNFs.

As with bounded-error approximate degree (see the end of Section 5.1), generic theorems are known that translate threshold degree lower bounds into communication lower bounds. For example, Sherstov [She09, She11a] showed how Theorem 26 implies an $\Omega(n^{1/3})$ lower bound on the **PP** communication complexity of a problem computed by a polynomial-size depth-3 circuit. As a consequence, such a circuit cannot be computed by depth-2 majority circuits of subexponential size, despite the fact that quasipolynomial-size depth-3 majority circuits can compute all of AC^0 [All89]. This result was later strengthened by Razborov and Sherstov [RS10] to show that the same polynomial-size depth-3 circuit cannot be computed efficiently in the even more powerful **UPP** communication model, answering an old open question of Babai, Frankl, and Simon [BFS86] regarding the relationship between **UPP** and the communication analog of the polynomial hierarchy.

These results and applications (as well as definitions of the **PP** and **UPP** communication models) are covered in detail in Section 10.

6 The Method of Dual Polynomials

Symmetrization arguments are quite powerful and have been used to determine the ε-approximate degree of many important functions. This includes all symmetric functions—those which depend only on the Hamming weight of the input [Pat92]. More sophisticated (and ad hoc) symmetrization arguments have also been applied to classes of non-symmetric functions such as halfspaces [She13b, She13c] and other functions central to quantum computing, cryptography, and circuit complexity [MP69, AS04], including the Minsky–Papert CNF as we have seen (Theorem 26).

Nevertheless, we should not expect symmetrization arguments to yield tight lower bounds for arbitrary functions. Approximating an n-variate function f is inherently a multivariate question. Unless f itself exhibits symmetric structure, it seems unlikely that a univariate function could fully capture the resistance of f to approximation by low-degree n-variate polynomials.

In contrast, a more recent lower bound technique called the *method of dual polynomials* is "lossless" in the sense that for any function f and any setting of the error parameter ε, the method is in principle capable of proving a tight lower bound on $\deg_\varepsilon(f)$. Here is how the method works. Fix a function $f\colon \{-1,1\}^n \to \{-1,1\}$ of interest and a degree bound d. What is the smallest error to which any polynomial of degree less than d can approximate f? The answer to this question is the value of the following linear program. It has $\binom{n}{<d} + 1$ variables, one for each coefficient of p and one for the error parameter ε, and $2 \cdot 2^n$ linear constraints that force p to approximate f to error at most ε at each input $x \in \{-1,1\}^n$.

$$\begin{array}{rll} \min_{p,\varepsilon} & \varepsilon & \\ \text{s.t.} & |p(x) - f(x)| \le \varepsilon & \text{for all } x \in \{-1,1\}^n \\ & \deg p < d & \end{array}$$

Taking the dual yields the following linear program, which has 2^n real-valued variables, one for each $x \in \{-1,1\}^n$.[14] It is helpful to think of these 2^n variables as comprising all evaluations of a

[14] Excellent introductions to linear-programming duality can be found at [O'D11] and [Vaz01, Chapter 12].

real-valued function $\psi\colon \{-1,1\}^n \to \mathbb{R}$.

$$\begin{aligned} \max_{\psi} \quad & \sum_{x\in\{-1,1\}^n} \psi(x)f(x) \\ \text{s.t.} \quad & \sum_{x\in\{-1,1\}^n} |\psi(x)| = 1 \\ & \sum_{x\in\{-1,1\}^n} \psi(x)p(x) = 0 \qquad \text{for all } p \text{ with } \deg p < d \\ & \psi(x) \in \mathbb{R} \text{ for all } x \in \{-1,1\}^n \end{aligned}$$

Weak LP duality implies that in order to prove that $\deg_\varepsilon(f) \geq d$, it suffices to identify a function $\psi\colon \{-1,1\}^n \to \mathbb{R}$ satisfying the following three conditions.

$$\sum_{x\in\{-1,1\}^n} \psi(x)f(x) > \varepsilon \tag{15}$$

$$\sum_{x\in\{-1,1\}^n} |\psi(x)| = 1 \tag{16}$$

$$\sum_{x\in\{-1,1\}^n} \psi(x)p(x) = 0 \text{ for all polynomials } p \text{ of degree less than } d. \tag{17}$$

Strong LP duality, moreover, implies that *every* approximate degree lower bound on f is witnessed by such a ψ. Such a ψ is called a *dual polynomial* for f. We refer to Condition (15) by saying that ψ has correlation at least ε with f, to Condition (16) by saying that ψ has ℓ_1-norm 1, and to Condition (17) by saying that ψ has *pure high degree* at least d, denoting the largest such d by $\mathrm{phd}(\psi)$. This terminology comes from the fact that ψ satisfies Condition (17) if and only if its representation as a multilinear polynomial is a sum only of monomials with degree at least d. We use $\|\psi\|_1 = \sum_{x\in\{-1,1\}^n} |\psi(x)|$ to denote the ℓ_1-norm of ψ and $\langle \psi, \varphi \rangle = \sum_{x\in\{-1,1\}^n} \psi(x)\varphi(x)$ to denote the correlation of any two functions $\psi, \varphi\colon \{-1,1\}^n \to \mathbb{R}$.

We encapsulate the discussion above with the following statement.

Theorem 28. Let $f : \{-1,1\}^n \to \{-1,1\}$. Then for every $\varepsilon > 0$, we have $\deg_\varepsilon(f) \geq d$ if and only if there exists a *dual polynomial* $\psi : \{-1,1\}^n \to \mathbb{R}$ meeting conditions (15), (16), and (17) above.

One may find it helpful to think of ψ as capturing the "component" of f that is "completely missed" by polynomials of degree less than d. Indeed, the pure high degree condition means that every such polynomial p is totally uncorrelated with ψ. If ψ is well-correlated with f, then it means that ψ is a "big part" of f and hence such p must incur a lot of error when approximating f.

Decomposing dual polynomials into pieces. It can be fruitful to think of ψ as consisting of two pieces. There are in fact two natural ways to perform such a decomposition.

- We can think of

$$\psi = \frac{1}{2}(\psi_{+1} - \psi_{-1}) \tag{18}$$

where $\psi_{-1} = 2\max\{-\psi(x), 0\}$ and $\psi_{+1} = 2\max\{\psi(x), 0\}$ are non-negative functions. We think of ψ_{+1} as the "positive part" of ψ and ψ_{-1} as the "negative part". The factor of 2 is chosen to ensure that ψ_{-1} and ψ_{+1} are probability mass functions. Indeed, so long as ψ has pure high degree at least 1 (implying it is uncorrelated with the constant-1 function), then since ψ has ℓ_1-norm 1, it must be the case that $\|\psi_{-1}\|_1 = \|\psi_{+1}\|_1 = 1$. The pure high degree condition ensures that no degree-d polynomial can distinguish the distributions ψ_{-1} and ψ_{+1} with any advantage over random guessing,[15] while the correlation condition guarantees that f *can* distinguish ψ_{-1} and ψ_{+1} with advantage ε. This perspective has been helpful in using approximate degree lower bounds to design low-complexity secret-sharing schemes [BIVW16] (see Section 11.1).

- Alternatively, we can think of $\psi(x)$ as consisting of a sign, $\mathrm{sgn}(\psi(x)) \in \{-1, 1\}$, and a magnitude $|\psi(x)|$, so that $\psi(x) = \mathrm{sgn}(\psi(x)) \cdot |\psi(x)|$. The sign $\mathrm{sgn}(\psi(x))$ can be thought of as ψ's "prediction" for $f(x)$ and the magnitude $|\psi(x)|$ as a measure of ψ's confidence in its prediction. The correlation requirement (Equation (15)) ensures that ψ's predictions, when weighted by its confidence, are accurate on average. With this in mind, we say that ψ *makes an error* at x if $\mathrm{sgn}(\psi(x)) \cdot f(x) < 0$.

  When ψ has ℓ_1-norm 1, we use $|\psi|$ to denote the probability distribution under which x is assigned probability $|\psi(x)|$. Observe that the correlation $\langle \psi, f \rangle$ equals

  $$\Pr_{x\sim|\psi|}[\mathrm{sgn}(\psi(x)) = f(x)] - \Pr_{x\sim|\psi|}[\mathrm{sgn}(\psi(x)) \neq f(x)] = 1 - 2\Pr_{x\sim|\psi|}[\mathrm{sgn}(\psi(x)) \neq f(x)].$$

  If ψ weakly sign-represents f (i.e., ψ *never* makes an error), then $\langle \psi, f \rangle = 1$. In this case we say that ψ is *perfectly correlated* with f. This means that for *every* $\varepsilon < 1$, ψ demonstrates that the ε-approximate degree of f is at least $\mathrm{phd}(\psi)$; equivalently, $\deg_{\pm}(f)$ is at least $\mathrm{phd}(\psi)$.

This discussion leads us to the following explicit characterization of threshold degree.

Theorem 29. Let $f : \{-1, 1\}^n \to \{-1, 1\}$. Then $\deg_{\pm}(f) \geq d$ if and only if there exists a dual polynomial $\psi : \{-1, 1\}^n \to \mathbb{R}$ such that

$$\psi(x)f(x) \geq 0 \text{ for all } x \in \{-1, 1\}^n \tag{19}$$

$$\sum_{x\in\{-1,1\}^n} |\psi(x)| = 1 \tag{20}$$

$$\sum_{x\in\{-1,1\}^n} \psi(x)p(x) = 0 \text{ for all polynomials } p \text{ of degree less than } d. \tag{21}$$

[15] A polynomial p *distinguishes* two probability distributions μ and μ' with advantage ε if $|\mathbb{E}_{x\sim\mu}[p(x)] - E_{x\sim\mu'}[p(x)]| \geq \varepsilon$, i.e., if p's expected behavior under a random draw from the first distribution differs significantly from its behavior under a random draw from the second distribution. "Random guessing" refers to the procedure that ignores its input x and outputs a random bit. Since random guessing ignores its input, it fails to distinguish any two probability distributions μ and μ' (i.e., its advantage is 0).

A simple example of a dual polynomial. Consider the parity function function $\oplus_n$ on n bits. Minsky and Papert famously[16] used symmetrization to prove that $\deg_{\pm}(\oplus_n) = n$. A dual polynomial for this fact is simply $\psi \coloneqq 2^{-n} \cdot \oplus_n$. Clearly ψ has perfect correlation with $\oplus_n$ (since it is just a rescaling of $\oplus_n$ itself) and has ℓ_1-norm 1. Finally, as $\oplus_n$ is a monomial of degree n, it is uncorrelated with any polynomial of degree at most $n-1$.

6.1 A Dual Polynomial for OR_n

A more complicated example is to construct a dual polynomial for the fact that $\widetilde{\deg}(\mathsf{OR}_n) \geq \Omega(\sqrt{n})$. Here is a construction from [BT15a], slightly refining an earlier dual polynomial of Špalek [Špa08] and in turn building on ideas of Kahn et al. [KLS96]. For any subset $S \subseteq [n]^*$, define the univariate polynomial $q_S(t) = \prod_{i \in [n]^*, i \notin S}(t-i)$. We refer to S as the set of *unkilled points*, since $q_S(t) = 0$ for all $t \in [n]^* \setminus S$. Let c be a sufficiently large constant, and let

$$S = \{0,1\} \cup \{ci^2 \colon i = 1, 2, \ldots, \lfloor \sqrt{n/c} \rfloor\}. \tag{22}$$

Define $\psi \colon \{-1,1\}^n \to \mathbb{R}$ as $\psi(x) = (-1)^{|x|} \cdot q_S(|x|)$, and finally define the dual polynomial for OR_n to be $\psi_{\mathsf{OR}}(x) = \psi(x)/\|\psi\|_1$. By design, ψ_{OR} has ℓ_1-norm 1, so to show that it is a dual polynomial for OR_n, we must show it has pure high degree $\lfloor \sqrt{n/c} \rfloor$ and that it has correlation at least $1/3$ with OR_n. The former holds by combining the following fact and lemma.

Fact 30. If Q is any univariate polynomial of degree at most $n-1$, then $\sum_{t=0}^{n} (-1)^t \binom{n}{t} Q(t) = 0$.

Proof. We again use the fact that the parity function on n bits is uncorrelated with every polynomial of total degree at most $n-1$. The n-variate polynomial $Q(|x|)$ has degree at most $n-1$ and its correlation with the parity function is $\sum_{t=0}^{n} (-1)^t \binom{n}{t} Q(t)$ (here, we use the fact that there are exactly $\binom{n}{t}$ inputs of Hamming weight t). □

Lemma 31. ψ_{OR} has pure high degree at least $d = \lfloor \sqrt{n/c} \rfloor$.

Proof. Let $p \colon \{-1,1\}^n \to \mathbb{R}$ be a polynomial of degree less than d. Then Lemma 22 guarantees that $Q(t) = \mathbb{E}_{x \sim H_t}[p(x)]$ is a univariate polynomial of degree less than d over $[n]^*$. The correlation of p with ψ_{OR} is

$$\sum_{t=0}^{n} \sum_{x \in H_t} \psi_{\mathsf{OR}}(x) p(x) = \frac{1}{\|\psi\|_1} \sum_{t=0}^{n} (-1)^t \binom{n}{t} q_S(t) \cdot Q(t).$$

This evaluates to 0 using Fact 30 and the fact that $q_S \cdot Q$ is a univariate polynomial of degree at most $\deg(q_S) + \deg(Q) \leq (n-d-1) + d = n-1$. □

The first conceptual step to showing that ψ_{OR} has correlation at least $1/3$ with OR_n is the following fact.

Fact 32. The correlation of ψ_{OR} with OR_n is $\langle \psi_{\mathsf{OR}}, \mathsf{OR}_n \rangle = 2 \cdot \psi_{\mathsf{OR}}(\mathbf{1}_n)$.

Proof. Since ψ_{OR} has pure high degree at least 1, it is uncorrelated with the constant-1 function. Hence, $\sum_{x \in \{-1,1\}^n} \psi_{\mathsf{OR}}(x) \cdot \mathsf{OR}_n(x) = 2 \cdot \psi_{\mathsf{OR}}(\mathbf{1}_n) + \sum_{x \in \{-1,1\}^n} \psi_{\mathsf{OR}}(x) \cdot (-1) = 2 \cdot \psi_{\mathsf{OR}}(\mathbf{1}_n)$. □

[16] Or perhaps infamously, given that this result contributed to the first "AI winter" for neural network research, see Section 5.3.

Hence, to show that $\langle \psi_{\mathsf{OR}}, \mathsf{OR} \rangle \geq 1/3$, it suffices to show that $\psi_{\mathsf{OR}}(\mathbf{1}_n) \geq 1/6$. In other words, ψ_{OR} places a constant fraction of its mass on this single input. This follows from an elementary calculation, albeit a lengthy one, upon which we now embark. We attempt to ease the pain by breaking the calculation into steps and providing intuition for each step as we go. (This calculation can also be skipped with no loss of continuity in this survey.)

Proof that $\psi_{\mathsf{OR}}(\mathbf{1}_n) \geq 1/6$. Recall that $\psi_{\mathsf{OR}}(\mathbf{1}_n) = \|\psi\|_1^{-1} \cdot \psi(\mathbf{1}_n)$. By construction, $\psi(x) = 0$ unless $|x| \in S$. Hence, letting $A = \psi(\mathbf{1}_n)$ and

$$B = \sum_{t \in S \setminus \{0\}} \sum_{|x|=t} |\psi(x)|,$$

we have that

$$\psi_{\mathsf{OR}}(\mathbf{1}_n) = \frac{A}{A+B}.$$

So showing that $\psi_{\mathsf{OR}}(\mathbf{1}_n) \geq 1/6$ is equivalent to showing that $B \leq 5A$.

For $t \in S$, let

$$B_t = \sum_{|x|=t} |\psi(x)| = \binom{n}{t} |q_S(t)|.$$

Then $B = \sum_{t \in S \setminus \{0\}} B_t$. We will show that for any $t \in S \setminus \{0,1\}$ and a large enough constant $c > 0$ in the definition of S (Equation (22)), $B_t/A \lesssim 2/t$ (later, we also explain that $B_1/A \leq 2$). This ensures the desired result that

$$B/A \lesssim B_1/A + \cdot \sum_{i=1}^{\sqrt{n/c}} 2/(ci^2) \leq 2 + A \sum_{i=1}^{\infty} 2/(ci^2) = 2 + \pi^2/(3c),$$

which is less than 5 if $c \geq 2$.

For any $t \in S \setminus \{0\}$, here is the key calculation bounding B_t/A:

$$B_t/A = \binom{n}{t} \frac{|q_S(t)|}{|q_S(0)|} = \binom{n}{t} \left| \frac{\prod_{j \in [n]^* \setminus S} (j-t)}{\prod_{j \in [n]^* \setminus S} (j-0)} \right| =$$

$$\binom{n}{t} \left(\frac{\prod_{j \in S:\, j \neq 0} (j-0)}{n!} \right) \cdot \left(\frac{t!(n-t)!}{\prod_{j \in S:\, j \neq t} |j-t|} \right) = \frac{\prod_{j \in S:\, j \neq 0} j}{\prod_{j \in S:\, j \neq t} |j-t|}. \tag{23}$$

Before continuing the calculation to upper bound B_t/A, let us attempt to give some intuition for why B_t/A is small if t is large. The denominator in Equation (23) is the product of the distances of t to every other unkilled point, while the numerator is the same but with t replaced by 0. The key intuition is that the numerator is smaller than the denominator because 0 is close to a lot of unkilled points (namely 1 and ci^2 for small values of i). While large values of $t \in S$ are not close to *any* other unkilled points, because the distance between any two perfect squares ci^2 and $c(i+1)^2$ is $2ci+1$, which grows linearly with i. For example, whereas the closest point in S to 0 is 1, the closest point in S to $t = c\lfloor\sqrt{n/c}\rfloor^2 \approx n$ has distance $\approx \sqrt{n/c} \gg 1$ to t.

The formal calculation upper bounding Expression (23) is simpler if we do *not* include the point 1 in S. So for illustration, we next show that Expression (23) is at most 2 if

$$S = \{ci^2 \colon i = 0, 1, \ldots, \lfloor\sqrt{n/c}\rfloor\},$$

and then explain how the tighter bound of $1/t$ can be obtained if 1 is included in S.

Bounding Expression (23) **if** 1 **is not included in** S**.** Letting $m = \lfloor \sqrt{n/c} \rfloor$ and $t = ci^2$ for $i \geq 1$, we have:

$$\begin{aligned}
&\frac{\prod_{j \in S:\, j \neq 0} j}{\prod_{j \in S:\, j \neq t} |j - t|} \\
&= \frac{\prod_{j=1}^{m} (cj^2)}{\left(\prod_{j=0}^{i-1} |ci^2 - cj^2|\right)\left(\prod_{j=i+1}^{m} |ci^2 - cj^2|\right)} \\
&= \frac{(m!)^2}{\left(\prod_{j=0}^{i-1} |i - j|(i+j)\right)\left(\prod_{j=i+1}^{m} |i - j|(i+j)\right)} \\
&= \frac{(m!)^2}{\left(i! \prod_{j=0}^{i-1} (i+j)\right)\left((m-i)! \prod_{j=i+1}^{m} (i+j)\right)} \\
&= \frac{2(m!)^2}{(m+i)!(m-i)!}.
\end{aligned}$$

Finally, observe that $\frac{(m!)^2}{(m+i)!(m-i)!} = \frac{m}{m+i} \cdot \frac{m-1}{m+i-1} \cdot \cdots \cdot \frac{m-i+1}{m+1}$ is a product of terms smaller than 1, and hence

$$\frac{2(m!)^2}{(m+i)!(m-i)!} \leq 2. \tag{24}$$

Bounding Expression (23) **if** 1 **is included in** S**.** One gets a tighter bound on Expression (23) if 1 is included in S because 0 is much closer to 1 than is any other point in S. Specifically, for $t = ci^2$ with $i \geq 1$, if we include 1 in S, then the numerator of Expression (23) does not change, while the denominator increases by a factor of $(t-1)$. This gives us the desired bound that $B_t \leq 2/(t-1)$.

Of course, including 1 to S does have the effect of causing $B_1 \neq 0$, so we must separately show that $|B_1|$ is not too large. One can show that B_1/A is only slightly larger than 1 because when $t = 1$, each term of the numerator of Expression (23) is exceedingly close to the corresponding denominator (namely, within additive distance 1).

6.1.1 Where did this dual come from?

A common complaint about dual polynomial constructions is that their definitions appear as if by magic, with lengthy calculations needed to show they are well-correlated with the target function f. But there is one source of intuition regarding their construction: complementary slackness. One can think of a dual polynomial ψ as assigning weights to the *constraints* of the primal linear program, with $\psi(x)$ being the weight assigned to the constraint $|p(x) - f(x)| \leq \varepsilon$. Complementary slackness asserts that if p is an optimal solution to the primal linear program, there must be an optimal solution ψ^* to the dual that only assigns nonzero weight to the constraints *made tight* by p, i.e., $\psi^*(x) \neq 0$ only for those x such that $|p(x) - f(x)| = \varepsilon$.

For the function $f = \mathsf{OR}_n$, we know roughly what an optimal solution to the primal looks like—see Equation (7), which gave an approximation $p(x) = q(\mathcal{A}(x)/n)$ for OR_n, where q is the transformed degree-d Chebyshev polynomial from Equation (7). The values of $\mathcal{A}(x)/n$ where

$|q(\mathcal{A}(x)/n) - \mathsf{OR}_n(x)|$ is maximized are closely approximated by the extreme points of the degree-d Chebyshev polynomial. These extreme points are well-known to be given by the *Chebyshev nodes*, equal to $\cos\left(\frac{i\pi}{d}\right)$ for $i = 1, 2, \ldots, d$.

Taking the Taylor-series expansion $\cos(x) = 1 - x^2/2! + x^4/4! - x^6/6! + \ldots$ and truncating it after the quadratic term shows that

$$\cos\left(\frac{i\pi}{d}\right) \approx 1 - \frac{1}{2} \cdot \left(\frac{i\pi}{d}\right)^2.$$

When $d = \Theta(\sqrt{n})$ we have $1 - \frac{1}{2} \cdot \left(\frac{i\pi}{d}\right)^2 \approx 1 - 2ci^2/n$ for some constant c. Inputs x for which $\mathcal{A}(x)/n = 1 - 2ci^2/n$ are precisely those inputs with Hamming weight $|x| = ci^2$. And these in turn are exactly those inputs (other than those with $|x| = 1$) in S that are assigned nonzero values by ψ per Equation (22).

6.1.2 Two additional properties of ψ_{OR}

The dual polynomial ψ_{OR} we constructed satisfies additional properties beyond what is needed (Conditions (15)-(17)) to ensure that $\widetilde{\deg}(\mathsf{OR}) \geq \Omega(\sqrt{n})$. As we will see later, these properties play essential roles in constructing and analyzing dual polynomials for functions derived from OR_n via composition, e.g., $\mathsf{AND}_m \circ \mathsf{OR}_n$.

First, any dual polynomial for ψ_{OR} has an important one-sided error property [GS10]. Fact 32 implies that $\psi_{\mathsf{OR}}(\mathbf{1}_n)$ must be positive if ψ_{OR} is to have positive correlation with OR_n. Since $\mathsf{OR}_n^{-1}(+1) = \{\mathbf{1}_n\}$, this means that the only inputs on which ψ_{OR} makes an error are in $\mathsf{OR}_n^{-1}(-1)$ (recall that we say ψ *makes an error* at x if $\operatorname{sgn}(\psi(x)) \cdot f(x) < 0$). This is stated in the following corollary.

Corollary 33. $\{x \colon \psi_{\mathsf{OR}}(x) \cdot \mathsf{OR}(x) < 0\} \subseteq \mathsf{OR}^{-1}(-1)$.

Second, as shown in [BT19b, BKT18], the calculation used to show that $\psi_{\mathsf{OR}}(\mathbf{1}_n) \geq 1/6$ in fact establishes the following stronger property, showing that the total mass that $|\psi_{\mathsf{OR}}|$ places on inputs of Hamming weight t decreases very rapidly with t, especially once $t \gg \sqrt{n}$.

Theorem 34. There are constants $c_1, c_2 > 0$ such that for all $t \in [n]^*$,

$$\sum_{|x|=t} |\psi_{\mathsf{OR}}(x)| \leq c_1 \cdot \exp(-c_2 \cdot t/\sqrt{n})/t.$$

Proof. The calculation is identical to that used to show that $\psi_{\mathsf{OR}}(\mathbf{1}_n) \geq 1/6$, with the additional observation that Expression (24) is exponentially small if $t = ci^2$ for $i \geq n^{1/4}$. Specifically, recalling the $m = \lfloor \sqrt{n/c} \rfloor$, Expression (24) equals $\frac{2(m!)^2}{(m+i)!(m-i)!} = 2 \cdot \binom{2m}{m}^{-1} \binom{2m}{m+i}$. By standard results about anti-concentration of the binomial distribution, this is at most $2 \cdot \exp(-\Omega(i^2/m)) = \exp\left(-\Omega(t/\sqrt{n})\right)$ as required to conclude that $\sum_{|x|=t} |\psi_{\mathsf{OR}}(x)| \leq c_1 \cdot \exp(-c_2 \cdot t/\sqrt{n})/t$ for some universal constants $c_1, c_2 > 0$. □

The extra properties satisfied by the dual polynomial ψ_{OR} captured in Theorem 34 and Corollary 33 both have natural "primal" interpretations, which readers might find more intuitive.

Primal interpretation of Corollary 33: One-sided approximate degree. Let ψ be a dual polynomial for the ε-approximate degree of f, such that ψ satisfies the additional property that

$$\{x \colon \psi(x) \cdot f(x) < 0\} \subseteq f^{-1}(-1). \tag{25}$$

Then ψ in fact witnesses that the *one-sided* approximate degree of g is at least $d = \mathrm{phd}(\psi)$. Here, one-sided approximate degree is an intermediate notion between approximate degree and threshold degree, defined below.

Definition 35. A real polynomial p is a *one-sided* ε-approximation for f if

$$|p(x) - (-1)| \leq \varepsilon \quad \forall x \in f^{-1}(-1) \qquad \text{and} \qquad p(x) \geq 1 - \varepsilon \quad \forall x \in f^{-1}(1).$$

The one-sided approximate degree of f, denoted $\widetilde{\mathrm{odeg}}_\varepsilon(g)$, is the minimum degree of a one-sided ε-approximation for f.

Note that $\deg_\pm(f) \leq \widetilde{\mathrm{odeg}}_\varepsilon(f) \leq \widetilde{\deg}_\varepsilon(f)$ for every $\varepsilon > 0$, but there can be huge gaps in either inequality. For instance, we've seen that OR_n has one-sided approximate degree equal to its approximate degree (namely, $\Theta(\sqrt{n})$), which is vastly larger than its threshold degree, which is 1. Meanwhile $\widetilde{\mathrm{odeg}}_{1/3}(\mathsf{AND}_n) = 1$, with the one-sided approximation being $\mathcal{A}(x) + (n-1)$. This equals the threshold degree of AND_n and is vastly smaller than its approximate degree $\Theta(\sqrt{n})$.

Claim 36. For every $\varepsilon > 0$ and degree d, we have $\widetilde{\mathrm{odeg}}_\varepsilon(f) \geq d$ if and only if there exists a dual polynomial ψ satisfying Conditions (15)-(17) as well as Condition (25).

One can prove Claim 36 by expressing one-sided approximate degree as a linear program analogous to approximate degree, and observing that a ψ satisfying the assumptions of Claim 36 is equivalent to a solution to the dual linear program with value ε.

Primal interpretation of Theorem 34. Suppose f has a dual polynomial ψ of pure high degree at least d that places very little mass on a subset $S \subseteq \{-1,1\}^n$, i.e., $|\psi(S)| \coloneqq \sum_{x \in S} |\psi(x)|$ is small. Then f cannot be approximated by any degree-d polynomial p, even if p is allowed to be *very* large on inputs in S. This is formalized in the following claim.

Claim 37. Let $0 < \delta < 1$. Suppose that ψ satisfies Conditions (15)-(17) and additionally that $|\psi(S)| \leq \varepsilon\delta/3$. Then for any polynomial p such that

$$|p(x) - f(x)| \leq \varepsilon/3 \text{ for all } x \notin S \quad \text{and} \quad |p(x)| \leq 1/\delta \text{ for all } x \in S, \tag{26}$$

we have $\deg(p) \geq d$.

Proof. Let p be a polynomial of degree less than d satisfying Condition (26). Then because ψ has pure high degree at least d, we have $\langle \psi, p \rangle = 0$. On the other hand,

$$\begin{aligned}\langle \psi, p \rangle &= \sum_{x \notin S} \psi(x)p(x) + \sum_{x \in S} \psi(x)p(x) \geq \left(\sum_{x \notin S} \psi(x)f(x) - |\psi(x)| \cdot \frac{\varepsilon}{3} \right) - \sum_{x \in S} |\psi(x)| \cdot \frac{1}{\delta} \\ &\geq \langle \psi, f \rangle - \varepsilon\delta/3 - \varepsilon/3 - \varepsilon/3 > 0.\end{aligned}$$

Here, the first inequality used Condition (26), the second used that the ℓ_1-norm of ψ is 1 (Equation (16)) and that $|\psi(S)| \leq \varepsilon\delta/3$, and the final inequality used that $\langle \psi, f \rangle > \varepsilon$ (Condition (15)). $\square$

A similar argument to Claim 37 shows that Theorem 34 implies[17] that there is some constant $c > 0$ such that no polynomial of degree $d \leq c\sqrt{n}$ can satisfy the following condition:

$$|p(x) - \mathsf{OR}_n(x)| \leq \exp(c \cdot |x|/\sqrt{n}) \text{ for all } x \in \{-1,1\}^n.$$

7 Dual Lower Bounds for Block-Composed Functions

Prior to 2012, Problem 13 was open even for the special case that $f = \mathsf{AND}$ and $g = \mathsf{OR}$. This case was eventually resolved via the method of dual polynomials [She13a, BT15a] using a simple yet powerful technique called dual block composition. Dual block composition tries to take dual polynomials witnessing the high approximate degrees of f and g individually, and combine them in a very specific manner to obtain a dual polynomial for the (even higher) approximate degree of $f \circ g$. The combining technique was proposed by several authors [She13b, Lee09, SZ09]. Here it is:

Definition 38. Given dual polynomials $\psi\colon \{-1,1\}^m \to \mathbb{R}$ and $\phi\colon \{-1,1\}^b \to \mathbb{R}$ such that ϕ has pure high degree at least 1, define the dual block composition $\psi \star \phi$ by

$$(\psi \star \phi)(x_1, \ldots, x_m) = \psi(\operatorname{sgn}(\phi(x_1)), \ldots, \operatorname{sgn}(\phi(x_m))) \cdot \prod_{i=1}^{m} (2|\phi(x_i)|).$$

Intuition for Definition 38. There are two ways to think about Definition 38, corresponding to the two ways of decomposing dual polynomials as discussed in Section 6. The first way to view $\psi \star \phi$ is as half the difference between two distributions $(\psi \star \phi)_{+1}$ and $(\psi \star \phi)_{-1}$ constructed as follows. To sample from $(\psi \star \phi)_{+1}$, first choose z from ψ_{+1} and then choose $x = (x_1, \ldots, x_m)$ from the product distribution $\otimes_{i=1}^m \phi_{z_i}$. Similarly, to sample from $(\psi \star \phi)_{-1}$, first choose z from ψ_{-1} and then choose $x = (x_1, \ldots, x_m)$ from the product distribution $\otimes_{i=1}^m \phi_{z_i}$.

The second interpretation is to get a prediction $\operatorname{sgn}((\psi \star \phi)(x))$ for $(f \circ g)(x)$ as follows. First, construct the vector $z = (\operatorname{sgn}(\phi(x_1)), \ldots, \operatorname{sgn}(\phi(x_m)))$ consisting of ϕ's predictions for each evaluation of g on $x_1, \ldots, x_m$. The final prediction $\operatorname{sgn}(\psi(z))$ for $(f \circ g)(x)$ is then simply ψ's prediction on input z. The confidence assigned to this prediction is proportional to the product of the confidences of all of the constituent predictions, namely $|\psi(z)| \cdot \prod_{i=1}^m |\phi(x_i)|$.

We remark that in the special case that $|\phi|$ is the uniform distribution, i.e., $|\phi(x)| = 2^{-b}$ for all $x \in \{-1,1\}^b$, then $\psi \star \phi$ is (up to scaling) the (non-dual) block composition of ψ with $2^b \cdot \phi$. This observation is particularly relevant in Section 10, when we prove communication and matrix-analytic lower bounds by applying dual block composition with $|\phi|$ uniform.

When is $\psi \star \phi$ a good dual witness? The hope is that if ψ is a dual witness to the fact that $\widetilde{\deg}(f) \geq d_f$ and ϕ is a dual witness to $\widetilde{\deg}(g) \geq d_g$, then $\psi \star \phi$ is a dual witness to the fact that $\widetilde{\deg}_\varepsilon(f \circ g) \geq d_f \cdot d_g$ for some constant $\varepsilon \in (0,1)$. This requires showing that $\psi \star \phi$ satisfies Conditions (15)-(17) for $d = d_f \cdot d_g$. In fact, as we prove below (Lemmas 39 and 40), $\psi \star \phi$ does always satisfy the second and third (Conditions (16) and (17)).

[17] One actually needs a slight strengthening of the upper bound in Theorem 34, in which the final factor $1/t$ is replaced with $1/t^2$.

Lemma 39. If ψ has pure high degree d_f and ϕ has pure high degree d_g, then the pure high degree of $\psi \star \phi$ is at least $d_f \cdot d_g$.

Proof. Let us consider the representation of $\psi \star \phi \colon \{-1,1\}^{m \cdot b} \to \mathbb{R}$ as a multilinear polynomial. (We remind the reader of the discussion of "Basics of Fourier analysis" in Section 2.1.) The lemma is equivalent to showing that the coefficient of every monomial of degree less than $d_f \cdot d_g$ is 0 (i.e., all Fourier coefficients of $\psi \star \phi$ of degree less than $d_f \cdot d_g$ are 0).

By linearity, it is without loss of generality to assume that $\psi(z_1, \ldots, z_m)$ is itself a monomial. By assumption, the degree of this monomial is at least d_f; say, $\psi(z_1, \ldots, z_m) = z_1 z_2 \ldots z_{d_f}$ (larger degree can be handled similarly). Then

$$2^{-m} \cdot (\psi \star \phi)(x) = \left(\prod_{i=1}^{d_f} \operatorname{sgn}(\phi(x_i)) \right) \left(\prod_{i=1}^{m} |\phi(x_i)| \right) = \left(\prod_{i=1}^{d_f} \phi(x_i) \right) \left(\prod_{i=d_f+1}^{m} |\phi(x_i)| \right).$$

By assumption, all monomials of ϕ have degree at least d_g. Since $x_1, \ldots, x_{d_f}$ are disjoint blocks of variables, every monomial appearing in $\prod_{i=1}^{d_f} \phi(x_i)$ has degree at least $d_f \cdot d_g$. For example, if $\phi(x_i) = \prod_{j=1}^{d_g} x_{i,j}$, then $\prod_{i=1}^{d_f} \phi(x_i) = \prod_{i=1}^{d_f} \prod_{j=1}^{d_g} x_{i,j}$. Since the blocks of variables $x_{d_f+1}, \ldots, x_m$ are disjoint from $x_1, \ldots, x_{d_f}$, multiplying this expression by $\prod_{i=d_f+1}^{m} |\phi(x_i)|$ (or any other function of $x_{d_f+1}, \ldots, x_m$ for that matter) does not decrease the degree of any appearing monomial. This proves the lemma. □

Lemma 40. If ϕ has pure high degree at least 1, then the ℓ_1-norm of $\psi \star \phi$ is 1.

Proof. Since ψ has ℓ_1-norm 1, $|\psi|$ is a probability distribution. Recall that we can think of $|\psi \star \phi|$ as first choosing z according to the probability distribution $|\psi|$, and then choosing $x = (x_1, \ldots, x_m) \in (\{-1,1\}^b)^m$ from the product distribution $\otimes_{i=1}^{m} \phi_{z_i}$. Hence, $|\psi \star \phi|$ is a convex combination[18] of probability distributions, and thus is itself a probability distribution.

□

Unfortunately, it is *not* always true that $\psi \star \phi$ satisfies Condition (15). An example is when $f = \mathsf{AND}_m$ and $g = \mathsf{AND}_b$. That is, if ψ is a dual witness for

$$\widetilde{\deg}(\mathsf{AND}_m) \geq \Omega(\sqrt{m})$$

and

$$\widetilde{\deg}(\mathsf{AND}_b) \geq \Omega(\sqrt{b}),$$

then $\psi \star \phi$ is *not* a dual witness for the fact that $\widetilde{\deg}(\mathsf{AND}_m \circ \mathsf{AND}_b) \geq \Omega(\sqrt{mb})$. While the latter statement is true (since $\mathsf{AND}_m \circ \mathsf{AND}_b$ is simply AND_{mb}), the function $\psi \star \phi$ is sadly not a dual witness to this fact. However, there are a variety of special cases in which $\psi \star \phi$ *is* known to witness that $\widetilde{\deg}_\varepsilon(f \circ g) \geq d_f \cdot d_g$ for some constant $\varepsilon \in (0,1)$. Section 7.1 describes the proof for $\mathsf{AND}_m \circ \mathsf{OR}_b$.

[18] A convex combination of objects is a linear combination (i.e., weighted sum) of objects in which the the coefficients (i.e., weights) of the sum are non-negative and sum to 1.

7.1 The Approximate Degree of $\mathsf{AND}_m \circ \mathsf{OR}_b$ is $\Omega(\sqrt{m \cdot b})$

We've seen that whenever ψ and ϕ are dual witnesses to the high approximate degrees of f and g, respectively, then $\psi \star \phi$ has two of the three properties needed to prove that $f \circ g$ has high approximate degree (large pure high degree, and ℓ_1-norm 1). We now sketch why the third property, namely high correlation with $f \circ g$, holds in the special case of $f = \mathsf{AND}_m$ and $g = \mathsf{OR}_b$.

Lemma 41. Let ψ have correlation at least $7/8$ with AND_m and ϕ have correlation at least $7/8$ with OR_b. Then $\psi \star \phi$ has correlation at least $1/3$ with $\mathsf{AND}_m \circ \mathsf{OR}_b$.

Proof. Recall that to sample from $|\psi \star \phi|$, one chooses a vector $z \in \{-1,1\}^m$ according to $|\psi|$ and then chooses an input $x = (x_1, \ldots, x_m) \in (\{-1,1\}^b)^m$ from the product distribution $\otimes_{i=1}^m \phi_{z_i}$. Taking this perspective, a short calculation shows that $\langle \psi \star \phi, \mathsf{AND}_m \circ \mathsf{OR}_b \rangle$ equals

$$\sum_{z \in \{-1,1\}^m} \psi(z) \cdot \mathbb{E}_{x \sim \otimes_{i=1}^m \phi_{z_i}}[(\mathsf{AND}_m \circ \mathsf{OR}_b)(x)]$$

$$= \sum_{z \in \{-1,1\}^m} \psi(z) \cdot \mathsf{AND}_m(z) \cdot \left(1 - 2 \cdot \underbrace{\Pr_{x \sim \otimes_{i=1}^m \phi_{z_i}}[(\mathsf{AND}_m \circ \mathsf{OR}_b)(x) \neq \mathsf{AND}_m(z)]}_{:= E(z)} \right). \tag{27}$$

In other words, $\langle \psi \star \phi, \mathsf{AND}_m \circ \mathsf{OR}_b \rangle$ is the same as $\langle \psi, \mathsf{AND}_m \rangle$, but each term in the sum is adjusted by an error term $E(z)$. Since we know that ψ has high correlation with AND_m, it is enough to show that these error terms are small. Quantitatively, it will be enough to show that $E(z) \leq 1/8$ for every z.

Case 1: $z \neq -\mathbf{1}_m$. In this case, $(\mathsf{AND}_m \circ \mathsf{OR}_b)(x) = \mathsf{AND}_m(z)$ so long as there is *at least* one x_i such that $\mathsf{OR}_b(x_i) = 1$. Let i be any index with $z_i = 1$. Then Fact 32 combined with the assumption that ϕ has correlation at least $7/8$ with OR_b implies that $\phi_{+1}(\mathbf{1}_b) \geq 7/8$ and hence $E(z) \leq 1/8$.

Case 2: $z = -\mathbf{1}_m$. In this case, $(\mathsf{AND}_m \circ \mathsf{OR}_b)(x) = \mathsf{AND}_m(z)$ only if $\mathsf{OR}_b(x_i) = -1$ for *all* $i = 1, 2, \ldots, m$, i.e., if $x_i \neq \mathbf{1}_b$ for all $i = 1, 2, \ldots, m$. In this case, Corollary 33 implies that $\phi_{-1}(\mathbf{1}_b) = 0$. It follows that for all x in the support[19] of $\otimes_{i=1}^m \phi_{-1}$, we have $x_i \neq \mathbf{1}_b$ for all $i = 1, 2, \ldots, m$. Hence, $E(-\mathbf{1}_m) = 0$. □

Lemmas 39-41, together with $\widetilde{\deg}_{7/8}(\mathsf{AND}_m) = \Theta(\sqrt{m})$ and $\widetilde{\deg}_{7/8}(\mathsf{OR}_b) = \Theta(\sqrt{b})$, imply:

Theorem 42. $\widetilde{\deg}(\mathsf{AND}_m \circ \mathsf{OR}_b) \geq \Omega(\sqrt{mb})$.

The key to the proof of Lemma 41 was Case 2, which exploited the fact that the dual witness ϕ for the inner function $g = \mathsf{OR}_b$ had one-sided error: $\{x \colon \phi(x) \cdot g(x) < 0\} \subseteq g^{-1}(-1)$ (Corollary 33), i.e., ϕ is actually a dual witness for $\widetilde{\mathrm{odeg}}_{7/8}(\mathsf{OR}_b) \geq \Omega(\sqrt{b})$ (see Definition 35 and Claim 36). In fact, the proof of Theorem 42 shows more generally that $\widetilde{\deg}(\mathsf{AND}_m \circ g) \geq \Omega(\sqrt{m} \cdot \widetilde{\mathrm{odeg}}_{1/3}(g))$.

[19] The phrase "x is in the support of probability distribution μ" means that μ assigns non-zero mass to x, i.e. $\mu(x) > 0$.

In contrast, recall that $\widetilde{\mathrm{odeg}}_{7/8}(\mathsf{AND}_b) = 1$. This explains why dual block composition yields a good dual witness for $\mathsf{AND}_m \circ \mathsf{OR}_b$ but not for $\mathsf{AND}_m \circ \mathsf{AND}_b$, even though both functions have approximate degree $\Theta(\sqrt{mb})$.

7.2 Hardness Amplification via Dual Block Composition

Hardness amplification theorems for approximate degree show that the block composition $f \circ g$ is harder to approximate by low-degree polynomials than is g alone.

7.2.1 Increasing degree via composition

Theorem 42 is an example of such a result, with $f = \mathsf{AND}_m$ and $g = \mathsf{OR}_b$, showing that the degree required to approximate $f \circ g$ to error $1/3$ is larger than the degree required to approximate g to the same error. Open Problem 13 above asks whether a vast generalization holds: is it the case that for every pair of functions f, g, $\widetilde{\deg}(f \circ g) \geq \Omega(\widetilde{\deg}(f) \cdot \widetilde{\deg}(g))$. This question has been resolved in a number of important special cases, described below.

The following theorem resolves the question in the case that the approximate degree and threshold degree of g happen to coincide (as is the case, e.g., for g equal to the parity function, $\oplus_b$). This result will be useful later (Section 10) when developing applications to sign-rank.

Theorem 43. [She13b, Lee09] For any $\varepsilon > 0$ let $f\colon \{-1,1\}^m \to \{-1,1\}$ and $g\colon \{-1,1\}^b \to \{-1,1\}$ be Boolean functions with $\deg_\varepsilon(f) \geq d$ and $\deg_\pm(g) \geq D$. Then $\widetilde{\deg}_\varepsilon(f \circ g) \geq D \cdot d$. Moreover, this is witnessed by $\psi \star \phi$, where ψ is any dual witness for the fact that $\widetilde{\deg}_\varepsilon(f) \geq d$ and ϕ is any dual witness for the fact that $\deg_\pm(g) \geq D$.

Proof. Suppose without loss of generality that $D \geq 1$. Lemma 39 guarantees that $\psi \star \phi$ has pure high degree at least $d \cdot D$ and Lemma 40 guarantees it has ℓ_1-norm 1. It remains to show that $\langle \psi \star \phi, f \circ g \rangle = \varepsilon$. As per the analysis in Lemma 41,

$$\sum_{z \in \{-1,1\}^m} \psi(z) \cdot f(z) \cdot \left(1 - 2 \cdot \underbrace{\Pr_{x \sim \otimes_{i=1}^m \phi_{z_i}} [(f \circ g)(x) \neq f(z)]}_{:= E(z)} \right). \tag{28}$$

Hence, it is enough to show that $E(z) = 0$. Since ϕ witnesses that $\deg_\pm(g) \geq D$, it follows that $\phi(y) \cdot g(y) \geq 0$ for all $y \in \{-1,1\}^b$. (See Theorem 29.) Hence, the support of ϕ_{+1} is contained in $g^{-1}(+1)$, and the support of ϕ_{-1} is contained in $g^{-1}(-1)$. Accordingly, for all $z \in \{-1,1\}^m$ and all x in the support of $\otimes_{i=1}^m \phi_{z_i}$, we have that $(f \circ g)(x) = f(z)$. Hence, $E(z) = 0$ as claimed. □

The following theorem resolves Open Problem 13 in a different special case, namely when the outer function f has linear approximate degree. Its proof is more complicated because the dual witness constructed to prove it is not obtained via "vanilla" dual block composition $\psi \star \phi$, but rather by multiplying $\psi \star \phi$ by a function p that is meant to "kill" the witness on problematic inputs without ruining its pure high degree. By killing a problematic input x, we mean that p is designed so that it evaluates to 0 at x. This proof is particularly technical and may be skipped with no loss of continuity in this survey.

Theorem 44 ([She12b])**.** For any Boolean functions $f : \{-1,1\}^m \to \{-1,1\}$ and $g\colon \{-1,1\}^b \to \{-1,1\}$, let $d = \widetilde{\deg}(f)$ and $D = \widetilde{\deg}_{1-d/(16m)}(g)$. Then

$$\widetilde{\deg}(f \circ g) \geq \Omega\left(d \cdot D\right).$$

Detailed proof sketch. As indicated before the theorem statement, the idea is to construct a dual witness for the claimed lower bound via a variant of dual block composition. Specifically, let ψ be a dual witness for $\widetilde{\deg}(f) \geq d$ and ϕ be a dual witness for $\widetilde{\deg}_{1-d/(16m)}(g) \geq D$. In general, $\psi \star \phi$ may not be well-correlated with $f \circ g$. Indeed, recall Equation (28). While we know that $\sum_{z\in\{-1,1\}^m} \psi(z) \cdot f(z) = \langle \psi, f\rangle \geq 1/3$, the additional error term $E(z)$ may be large for each $z \in \{-1,1\}^m$, for reasons we now explain.

Let $\mathcal{E}_\phi \coloneqq \{y\colon g(y) \neq \mathrm{sgn}(\phi(y))\}$ denote the error set for the witness ϕ for g, and let $E_\phi \coloneqq \sum_{y\in\mathcal{E}_\phi} |\phi(y)|$ denote the total mass placed on $\mathcal{E}_\phi$. We know that $1-d/(16m) = \langle \phi, g\rangle = 1-2E_\phi$, and hence E_ϕ may be as large as $d/(32m)$. For a fixed bit $z_i \in \{-1,1\}$, conditioning on $\mathrm{sgn}(\phi(x_i)) = z_i$ as done by ϕ_{z_i} could even raise the mass placed on $\mathcal{E}_\phi$ by another factor of 2, to $d/(16m)$. This means that if $x = (x_1,\ldots,x_m) \sim \otimes_{i=1}^m \phi_{z_i}$, the expected number of blocks $x_i \in \mathcal{E}_\phi$ may be as large as $m \cdot d/(16m) = d/16$, which is much larger than 1 if $d \geq \omega(1)$.

Let us refer to blocks $x_i \in \mathcal{E}_\phi$ as *error blocks*. If there is even *one* error block, then it is possible that $(f\circ g)(x) \neq \mathrm{sgn}((\psi\star\phi)(x))$. So unless the expected number of error blocks is vastly below its expectation, we cannot rule out the possibility that $(f\circ g)(x) \neq \mathrm{sgn}((\psi\star\phi)(x))$.

Accordingly, call an input $x = (x_1,\ldots,x_m) \in (\{-1,1\}^b)^m$ "bad" if there are one or more error blocks $x_i \in \mathcal{E}_\phi$, and otherwise call x "good". The idea is to modify $\psi\star\phi$ by multiplying it by a polynomial p designed to "kill" as many bad inputs x as possible. We need to do this without substantially decreasing the pure high degree of the dual witness. This constraint will prevent p from killing *all* bad inputs. Hence, we must also guarantee that p avoids "amplifying" the values assigned to bad inputs that are *not* killed by p.

Let $\mathbb{I}_{\mathcal{E}_\phi}$ denote the indicator function of $\mathcal{E}_\phi$. Let $Q(t) = \prod_{i=1}^{d/2}(t-i)$ be the univariate polynomial that "kills" all integers between 1 and $d/2$, and define p such that for all $(x_1,\ldots,x_m) \in \left(\{-1,1\}^b\right)^m$

$$p(x_1,\ldots,x_m) = Q\left(\sum_{j=1}^m \mathbb{I}_{\mathcal{E}_\phi}(x_j)\right).$$

Intuitively, when applied to an input $(x_1,\ldots,x_m) \in \left(\{-1,1\}^b\right)^m$, p counts the number of errors made by the m copies of ϕ on inputs $(x_1,\ldots,x_m)$, and evaluates to 0 if this number is between 1 and $d/2$. Note that, assuming $d/2$ is even, $p(x)$ is non-negative for all $x \in \left(\{-1,1\}^b\right)^m$.

An assumption to simplify the construction. To simplify the construction of the appropriate dual witness for $f\circ g$, let us assume that for any $z_i \in \{-1,1\}$, conditioning on $\mathrm{sgn}(\phi(x_i)) = z_i$ does not alter the probability that x_i is an error block. That is, we assume that the mass ϕ assigns to false-negative errors and false-positive errors is the same, i.e.,

$$\sum_{x_i\in\{-1,1\}^b\colon \phi(x_i)<0, f(x_i)=1} |\phi(x_i)| = \sum_{x_i\in\{-1,1\}^b\colon \phi(x_i)>0, f(x_i)=-1} |\phi(x_i)|. \tag{29}$$

The construction under the simplifying assumption. Define

$$\gamma'(x) \coloneqq (\psi \star \phi)(x) \cdot p(x) \tag{30}$$

and $\gamma = \|\gamma'\|_1^{-1} \cdot \gamma'$. By design, γ has ℓ_1-norm 1. We now explain that it also has pure high degree at least $D \cdot (d/2)$, and has correlation at least $1/3 - o(1)$ with $f \circ g$.

Pure high degree analysis. The key insight for the pure high degree assertion is that the analysis in Lemma 39 guarantees something a little stronger than the mere fact that $\mathrm{phd}(\psi \star \phi) \geq d \cdot D$. Specifically, recall the Fourier representation $\psi \star \phi = \sum_S \widehat{\psi \star \phi}(S) \cdot \chi_S(x_1, \ldots, x_m)$. For every parity function $\chi_S(x_1, \ldots, x_m)$ with a non-zero Fourier coefficient, there are least d *blocks* $x_{i_1}, \ldots x_{i_d}$, such that *every block* x_{i_j} has at least D variables in S. For example, if $m = 3$, $b = 2$, and $d = D = 2$, then $x_{1,1} \cdot x_{1,2} \cdot x_{2,1} \cdot x_{2,2}$ is a possible parity function with non-zero coefficient in the Fourier representation of $\psi \star \phi$, because there are two blocks that each contribute degree two to the parity. But $x_{1,1} \cdot x_{1,2} \cdot x_{2,1} \cdot x_{3,2}$ is not, because while its total degree is four, there is only one block x_1 contributing two or more variables to the parity.

Meanwhile, p is "relatively low-degree" in the following sense. Since the univariate polynomial Q has degree only $d/2$, the multivariate polynomial $p(x) = (Q \circ \mathbb{I}_{\mathcal{E}_\phi})(x)$ clearly has total degree at most $b \cdot (d/2)$. But in fact something even stronger holds: for every parity function $\chi_S(x_1, \ldots, x_m)$ in the Fourier representation of p with a non-zero coefficient, S involves variables from *at most* $d/2$ blocks. So in the example above, $x_{1,1} \cdot x_{1,2}$ is a possible parity function with non-zero coefficient in the Fourier representation of p, since it involves variables from only $1 = d/2$ blocks, but $x_{1,1} \cdot x_{2,1}$ is not as it involves variables from $2 > d/2$ blocks.

To summarize this discussion, we have that the non-zero Fourier coefficients of $\psi \star \phi$ correspond to parities involving at least d blocks, while the non-zero Fourier coefficients of p correspond to parities involving at most $d/2$ blocks. Therefore, when computing the Fourier representation of $(\psi \star \phi) \cdot p$ via the distributive law, each monomial of p only destructively interferes with at most $d/2$ blocks of each monomial of $\psi \star \phi$. Hence, $(\psi \star \phi) \cdot p$ has pure high degree at least $(d - (d/2)) \cdot D \geq (d/2) \cdot D$. For example, if $\psi \star \phi$ and p are as per the examples above, then

$$(\psi \star \phi)(x) \cdot p(x) = (x_{1,1} \cdot x_{1,2} \cdot x_{2,1} \cdot x_{2,2}) \cdot (x_{1,1} \cdot x_{1,2}) = x_{2,1} \cdot x_{2,2},$$

which has pure high degree $2 \geq 1 \cdot 2 = (d - (d/2)) \cdot D$.

Correlation analysis. As usual, the goal is to show that $\langle \gamma, f \circ g \rangle \approx \langle f, \psi \rangle$. Recall from Section 7 that to sample from $|\psi \star \phi|$, one chooses a vector $z \in \{-1, 1\}^m$ according to $|\psi|$ and then chooses an input $x = (x_1, \ldots, x_m) \in (\{-1, 1\}^b)^m$ from the product distribution $\otimes_{i=1}^m \phi_{z_i}$ (see Equation (18) for the definition of ϕ_{z_i}).

As γ is obtained by multiplying $\psi \star \phi$ by p, let μ_z be the distribution obtained from $\otimes_{i=1}^m \phi_{z_i}$ by multiplying the probability assigned to each input x by $p(x)$, and then renormalizing to ensure μ_z has ℓ_1-norm 1. By our simplifying assumption (Equation (29)), the distribution of $p(x)$ when $x \sim \otimes_{i=1}^m \phi_{z_i}$ is independent of z, and hence this normalization factor is the same for all $z \in \{-1, 1\}^m$. It follows that, in analogy with Equation (28),

$$\langle \gamma, f \circ g \rangle = \sum_{z \in \{-1,1\}^m} \psi(z) \cdot f(z) \cdot \left(1 - 2 \cdot \underbrace{\Pr_{x \sim \mu_z} [(f \circ g)(x) \neq f(z)]}_{\coloneqq E'(z)} \right).$$

For each $z \in \{-1,1\}^m$, it is possible to establish that $E'(z) \leq 2^{-\Omega(d)}$ via the ideas above. Specifically, under the product distribution $x \sim \otimes_{i=1}^m \phi_{z_i}$, the number of error blocks is, in expectation, at most $d/16$, and $p(x) = 0$ if this number is between 1 and $d/2$. Hence, the total mass assigned by the product distribution $\otimes_{i=1}^m \phi_{z_i}$ to bad inputs x with T error blocks is 0 if $T \leq d/2$ and, by a standard Chernoff bound, is at most $e^{-4T^2/d}$ if $T > d/2$. Meanwhile, $p(x)$ maps good inputs to $(d/2)!$, and maps a bad input with T error blocks to $(T-1)(T-2)\cdot\cdots\cdot(T-(d/2))$. This ensures that, relative to good inputs, p amplifies that mass placed on bad inputs by a factor of at most $\frac{(T-1)!}{(d/2)!(T-1-(d/2))!} = \binom{T-1}{d/2} \leq \binom{T}{d/2} \leq (2Te/d)^{d/2}$.

Since the probability that μ_z assigns to x is the product of $p(x)$ and the mass assigned to x by $\otimes_{i=1}^m \phi_{z_i}$, this implies that $E'(z)$ is at most

$$\sum_{T=d/2+1}^{m} e^{-4T^2/d} \cdot (2Te/d)^{d/2} \leq m \cdot e^{-d} \cdot e^{d/2} \leq e^{-d/3}.$$

□

A last state-of-the-art result resolves Open Problem 13 (up to a logarithmic factor) in the special case that the outer function f is symmetric.

Theorem 45 ([BBGK18])**.** Let $f : \{-1,1\}^m \to \{-1,1\}$ be a symmetric Boolean function and g be an arbitrary function. Then $\widetilde{\deg}(f \circ g) \cdot \log m \geq \Omega(\widetilde{\deg}(f) \cdot \widetilde{\deg}(g))$.

Theorem 45 is not proved using the method of dual polynomials, but rather indirectly relies on a sophisticated quantum algorithm for combinatorial group testing, due to Belovs [Bel15].

7.2.2 Increasing error via composition

Sherstov [She12b] proved an XOR Lemma for approximate degree showing that $\oplus_m \circ g$ requires both higher degree *and larger error* to approximate than g itself. His proof technique was essentially identical to Theorem 44, using a refinement of dual block composition to construct a dual witness for the claim.

Theorem 46. ([She12b]) Let g be a Boolean function with $\widetilde{\deg}_{1/2}(g) \geq d$ and $F = \oplus_m \circ g$. Then $\widetilde{\deg}_{1-2^{-m}}(F) \geq \Omega(m \cdot d)$.

Approximate degree turns out to be a powerful tool for studying the fundamental circuit class AC^0, consisting of constant-depth $\{\mathsf{AND}, \mathsf{OR}, \mathsf{NOT}\}$-circuits of polynomial size. We will see later in this survey (Section 8) that AC^0 contains functions that cannot be approximated well by low-degree polynomials, and this implies that AC^0 contains functions that are hard to compute in a variety of models, such as quantum and small-bias communication complexity.

When using approximate degree to study AC^0, one would like the "hardness-amplified" function F to be a constant-depth circuit whenever g is. Theorem 46 is not useful in this context because the parity function $\oplus_m$ is not in AC^0. More recent work has shown that error amplification within AC^0 is possible by taking the outer function to be AND, so long as the inner function has high *one-sided* approximate degree (Section 6.1.2).

Theorem 47. ([BT15b]) Let g be a Boolean function with $\widetilde{\mathrm{odeg}}_{1/2}(g) \geq d$ and $F = \mathsf{AND}_m \circ g$. Then $\widetilde{\deg}_{1-2^{-m}}(F) \geq d$.

Theorem 48. ([She18b]) Let g be a Boolean function with $\widetilde{\mathrm{odeg}}_{1/2}(g) \geq d$ and $F = \mathsf{AND}_m \circ g$. Then $\deg_{\pm}(F) \geq \min\{d, m\}$.

Note that since $\widetilde{\mathrm{odeg}}_{1/2}(\mathsf{OR}_b) \geq \Omega(\sqrt{b})$, Theorem 26 is a special case of Theorem 48. That is, Minsky and Papert's threshold degree lower bound for their CNF is a special case of a far more general result that can be proved using dual block composition as opposed to symmetrization.

Theorems 47 and 48 are easily seen to be false if the assumption that $\widetilde{\mathrm{odeg}}_{1/2}(g) \geq d$ is replaced with $\widetilde{\deg}_{1/2}(g) \geq d$, as can be seen by setting $g = \mathsf{AND}$. Indeed, $\mathsf{AND}_m \circ \mathsf{AND}_b = \mathsf{AND}_{m \cdot b}$, and $\mathsf{AND}_{m \cdot b}$ can be approximated to error $1 - 1/(m \cdot b)$ with degree 1.

Proof of Theorem 47. Here, we define a simple dual witness ψ for the fact that AND_m has approximate degree at least 1 by taking $\psi(\mathbf{1}_m) = 1/2$, $\psi(-\mathbf{1}_m) = -1/2$, and $\psi(x) = 0$ otherwise. Let ϕ be any dual witness to the fact that $\widetilde{\mathrm{odeg}}_{1/2}(f) \geq d$. We claim that $\psi \star \phi = \frac{1}{2} \cdot \left(\phi_{+1}^{\otimes m} - \phi_{-1}^{\otimes m}\right)$ witnesses that $\widetilde{\deg}_{1-2^{-m}}(F) \geq d$. Note that $\psi \star \phi$ has ℓ_1-norm 1 by Lemma 40, and pure high degree d by Lemma 39 and the fact that $\mathrm{phd}(\psi) \geq 1$ and $\mathrm{phd}(\phi) \geq d$.

To show that $\langle \psi \star \phi, \mathsf{AND}_m \circ g \rangle \geq 1 - 2^{-m}$, recall from the proof of Lemma 41 (Equation (27)) that the key to showing that $\langle \psi \star \phi, \mathsf{AND}_m \circ g \rangle \approx \langle \psi, \mathsf{AND}_m \rangle = 1$ is to upper bound

$$E(z) = \Pr_{x \sim \otimes_{i=1}^m \phi_{z_i}} [(\mathsf{AND}_m \circ g)(x) \neq \mathsf{AND}_m(z)] \tag{31}$$

for the two points $z = -\mathbf{1}_m, \mathbf{1}_m$ in the support of $|\psi|$.

Case 1: $z = \mathbf{1}_m$. In this case, $(\mathsf{AND}_m \circ g)(x) \neq \mathsf{AND}_m(z)$ only if $g(x_1) = g(x_2) = \cdots = g(x_m) = -1$. It can be seen that since ϕ has correlation at least $1/2$ with g, $\phi_{+1}(g^{-1}(-1)) \leq 1/2$. Hence, for $z = \mathbf{1}_m$, Expression (31) is at most 2^{-m}.

Case 2: $z = -\mathbf{1}_m$. Since ϕ is a dual witness for the *one-sided* approximate degree of g, the support of ϕ_{-1} is a subset of $g^{-1}(-1)$, and hence the support of $\otimes_{i=1}^m \phi_{-1}$ is a subset of $(\mathsf{AND}_m \circ g)^{-1}(-1)$. Hence, for $z = -\mathbf{1}_m$, Expression (31) is 0. □

Theorem 48 can be proved by building on this construction, adding to $\psi \star \phi$ an additional "correction term" ζ of pure high degree m such that $\psi \star \phi - \zeta$ is perfectly correlated with $\mathsf{AND}_m \circ g$. We do not prove Theorem 48 in this survey, but closely related ideas can be found in the proof of Theorem 92 in Section 10, which constructs an explicit dual solution for the high threshold degree of the Minsky-Papert CNF $\mathsf{AND}_{n^{1/3}} \circ \mathsf{OR}_{n^{2/3}}$, with additional properties that are useful in applications to communication and circuit complexity.

Application: Oracle separations for statistical zero knowledge. Certain applications require the "hardness-amplifying function" to be still simpler than AND_m. Define $\mathsf{GAPMAJ}_m \colon \{-1,1\}^m \to \{-1,1\}$ to be the partial function that equals -1 if at least $2/3$ of its inputs are -1, equals $+1$ if at least $2/3$ of its inputs are $+1$, and is undefined otherwise.

Theorem 49 ([BCH+19])**.** Let f be a Boolean function with $\widetilde{\deg}_{1/2}(f) \geq d$. Let $F = \mathsf{GAPMAJ}_m \circ f$. Then $\widetilde{\deg}_{1-2^{-\Omega(m)}}(F) \geq d$ and $\deg_{\pm}(F) \geq \Omega(\min\{d, m\})$.

Bouland et al. [BCH+19] used this result to exhibit an oracle $\mathcal{O}$ relative to which $\mathbf{SZK}^{\mathcal{O}} \not\subseteq \mathbf{PP}^{\mathcal{O}}$. Here **SZK** is the class of languages with efficient statistical zero knowledge proofs—interactive proofs of language membership such that the proofs reveal no information other than their own validity.[20] As **PP** is a very powerful complexity class, this separation gives some evidence for the prevailing belief that **SZK** contains intractable problems. The proof of the oracle separation proceeds as follows.

A very brief and rough overview of **SZK** *and* **PP** *query complexity.* Recall that we briefly discussed the notion of query complexity in Section 4.1. Using a standard diagonalization argument (see [FSS84b, Ko89, RST15] for details of the technique), it suffices to establish a separation in the analogous query complexity models:

Fact 50. To obtain an oracle $\mathcal{O}$ such that $\mathbf{SZK}^{\mathcal{O}} \not\subseteq \mathbf{PP}^{\mathcal{O}}$, it suffices to identify an F such that $\mathbf{SZK}^{\mathbf{dt}}(F) = O(\log n)$ and $\mathbf{PP}^{\mathbf{dt}}(F) = n^{\Omega(1)}$.

Here $\mathbf{SZK}^{\mathbf{dt}}(F)$ denotes the least *cost* of a statistical zero knowledge *query* protocol computing F. The cost of a statistical zero-knowledge query protocol for F refers to the the length of the proof (i.e., the number of bits exchanged by the prover and verifier), plus the number of queries to the input string x made by the verifier.

Similarly, $\mathbf{PP}^{\mathbf{dt}}(F)$ is the least d for which a randomized algorithm making at most d queries computes $F(x)$ with probability at least $1/2 + 2^{-d}$ (see Section 10.1 for further details). Since the acceptance probability of any d-query randomized algorithm is a polynomial of degree at most d, we have that if $\mathbf{PP}^{\mathbf{dt}}(F) \leq d$, then $\widetilde{\deg}_{\varepsilon}(F) \leq d$ for $\varepsilon = 1 - 2^{-d}$. So to prove a $\mathbf{PP}^{\mathbf{dt}}$ lower bound, it is enough to prove an approximate degree lower bound for an error parameter that is exponentially close to 1.

Recall from Section 2.2 that the Permutation Testing Problem (PTP) is a partial function that interprets its input x as a list of (the binary representations of) $N = \Theta(n/\log n)$ numbers from range $[N]$. The list can itself be interpreted as a function $\pi\colon [N] \to [N]$. The function $\mathsf{PTP}(x) = -1$ if π is a permutation and $\mathsf{PTP}(x) = -1$ if π is "far" from every permutation.

As we show in Section 8.5, PTP has large (1/3)-approximate degree, namely $\Omega(N^{1/3})$ [Aar12, AS04]. Meanwhile, Permutation Testing has a zero-knowledge protocol with logarithmic cost: A common random string samples a range item $i \in [N]$, and the prover is required to provide a pre-image j of i under π. The verifier can confirm that $\pi(j) = i$ by querying $\log N$ bits of x. This protocol is perfectly complete, because if π is one-to-one, then any range element j has some pre-image i. And it has soundness error bounded away from 1, because if π is far from any permutation, then a constant fraction of range elements $j \in [R]$ have *no* pre-image under π, and if the common random string selects such a range element, there is no possible response the prover can send that would convince the verifier to accept.

The protocol is also perfect zero knowledge because, when the input is a permutation, the verifier learns only a random pair (i, j) such that $\pi(j) = i$; the verifier could compute this information on its own by picking j at random from $[N]$ and making $O(\log N)$ queries to learn $i = \pi(j)$.

To get a $\mathbf{PP}^{\mathbf{dt}}$ lower bound, we need a function with low **SZK** query complexity, yet with high ε-approximate degree even for ε exponentially close to 1. PTP itself does not have high

[20] The precise definition of statistical zero-knowledge proof systems is beyond the scope of this survey. Roughly speaking, these are interactive proof systems satisfying standard notions of completeness and soundness, such that the verifier runs in polynomial time and moreover learns nothing from the prover beyond the validity of the statement being proven.

ε-approximate degree if ε is larger than $1 - 1/n^2$ (see Footnote 6 of Section 2.2). However, we can transform PTP into such a function by composing it with a function that preserves **SZK** query complexity, yet amplifies hardness against polynomial approximation. Specifically, let $F = \mathsf{GAPMAJ}_{n^{1/4}} \circ \mathsf{PTP}_{n^{3/4}}$. One can show that composition with GAPMAJ preserves logarithmic **SZK** query complexity. Meanwhile, Theorem 49 implies that $\widetilde{\deg}_{1-2^{-n^{1/4}}}(F) = \Omega(n^{1/4})$.

7.3 Some Unexpected Applications of Dual Block Composition

While dual block composition was introduced as a way to understand how approximate degree behaves under function composition, it has found applications to lower bounding the approximate degree even of "non-block-composed" functions. In this section, we describe two initial such applications. The first is a clean proof of the tight lower bound on the ε-approximate of OR. The second is a clean proof of the tight lower bound on the $(1/3)$-approximate degree of any symmetric Boolean function.

Roughly speaking, the first result exploits that OR_n can in fact be written in as a block-composed function $\mathsf{OR}_n = \mathsf{OR}_t \circ \mathsf{OR}_{n/t}$. The second result exploits that for any symmetric Boolean function f, there is a block-composed function $g = \mathsf{MAJ}_{2t} \circ \mathsf{OR}_{n/(2t)}$ for some integer t such that g "agrees with" f on all of the inputs that "are responsible for" the large approximate degree of g. Hence, f "inherits" the hardness of g.

Section 8 describes much more sophisticated applications of dual block composition to non-block-composed functions.

7.3.1 Lower bound on the vanishing-error approximate degree of OR

Recall that in Theorem 14 of Section 4.2.1 we showed that the ε-approximate degree of OR is $O(\sqrt{n \log(1/\varepsilon)})$. We now establish a matching lower bound. This result was originally proved by [BCDWZ99] using the theorem of Coppersmith and Rivlin (Theorem 24). We prove it via a clean application of dual block composition due to Sherstov and Thaler [ST19].

Theorem 51. For any $\varepsilon \in [2^{-n}, 1/3]$, $\widetilde{\deg}_{\varepsilon}(\mathsf{OR}_n) \geq \Omega\left(\sqrt{n \log(1/\varepsilon)}\right)$.

Proof. Let $t = \log_2(1/\varepsilon)/3$, and let ϕ be a dual witness for $\widetilde{\deg}(\mathsf{OR}_{n/t}) \geq \Omega(\sqrt{n/t})$. Let $\psi = 2^{-t} \cdot \oplus_t$, which we interpret as a dual witness for the fact that the exact (i.e., $(\varepsilon = 0)$-approximate degree) of OR_t is t. Clearly ψ has pure high degree t and ℓ_1-norm 1. Consider $\psi \star \phi$.

By Lemmas 39 and 40, $\psi \star \phi$ has pure high degree at least $\Omega\left(t \cdot \sqrt{n/t}\right) = \Omega(\sqrt{nt})$, and $\|\psi \star \phi\|_1 = 1$. It remains to show that $\langle \psi \star \phi, \mathsf{OR}_n \rangle > \varepsilon$. By Fact 32, $\langle \psi \star \phi, \mathsf{OR}_n \rangle = 2\,|(\psi \star \phi)(\mathbf{1}_n)|$. Let $w = 2|\phi(\mathbf{1}_{n/t})| > 1/3$. Then

$$|(\psi \star \phi)(\mathbf{1}_n)| = 2^{-t} \cdot w^t > 2^{-t} \cdot 3^{-t} \geq 2^{-3t} \geq \varepsilon.$$

□

7.3.2 Lower bound on the approximate degree of symmetric functions

Recall from Section 5.2 that if $\mathsf{THR}^t_n \colon \{-1,1\}^n \to \{-1,1\}$ denotes the function for which $\mathsf{THR}^t_n(x) = -1$ if and only if $|x| \geq t$, then $\widetilde{\deg}(f) = \Theta(\sqrt{nt})$. Here, we present a proof of the $\Omega(\sqrt{nt})$ lower

bound using dual block composition, which we find to be cleaner than the symmetrization-based analysis sketched in Section 5.2 (we believe this dual witness is also cleaner than the only previous dual witness for symmetric Boolean functions in the literature [BT15a]). The proof does take for granted that $\widetilde{\deg}(\mathsf{MAJ}_{2t}) \geq \Omega(t)$; a reasonably simple dual witness for this fact is given in [BT15a, Section 4.3]. A similar argument was used in [BBGK18] to show a lower bound on the approximate degree of composed functions when the outer function is symmetric.

Lower bound via dual block composition. Unless $t = 0$ or $t = n$, THR_n^t cannot itself be written as a composition of two functions defined on smaller domains, so it may seem strange that we plan to use dual block composition to analyze its approximate degree. The idea is to identify a *partial* function F that *is* a composed function, such that $F(x) = \mathsf{THR}_n^t(x)$ for all inputs x in the domain of F. We then prove (via an explicit construction of a dual polynomial for F) that F requires degree $\Omega(\sqrt{nt})$ to approximate over its domain, and thereby conclude that $\mathsf{THR}_n^t(x)$ requires the same degree to approximate.

Approximate degree of partial functions. To give the details of the proof, we must introduce a natural notion of the ε-approximate degree of a partial function f_n defined over some strict subdomain S of $\{-1,1\}^n$. Specifically, the notion relevant to this section requires the approximating polynomial p for f to be bounded even at inputs in $\{-1,1\}^n$ outside of the promise S:

- $|p(x) - f_n(x)| \leq \varepsilon$ for all $x \in S$.
- $|p(x)| \leq 1 + \varepsilon$ for all $x \in \{-1,1\}^n \setminus S$.

We stress that this a *different* notion of approximate degree for partial functions than the one that is relevant to Section 8. That section relates the approximate degree of a total function such as $\mathsf{SURJ}_{R,N}$ to that of a partial function such as $\mathsf{AND}_R \circ \mathsf{OR}_N$ under the promise that the input has Hamming weight at most N; the notion of approximation relevant there places no restrictions on p's behavior outside of the promise set. That notion would not suffice in the setting of this section (see Footnote 21).

An application of LP duality similar to Section 6 reveals that a partial function f_n defined over domain S has ε-approximate degree at least d if and only if there exists a dual polynomial $\phi\colon \{-1,1\}^n \to \mathbb{R}$ of pure high degree at least d and ℓ_1-norm equal to 1, such that $\langle \phi, f_n \rangle \geq \varepsilon$. Here, for a partial function f_n, we define

$$\langle \phi, f_n \rangle \coloneqq \sum_{x \in S} f(x)\phi(x) - \sum_{x \in \{-1,1\}^n \setminus S} |\phi(x)|.$$

That is, any mass that ϕ places on inputs $x \notin S$ automatically count *against* the correlation $\langle \phi, f_n \rangle$ of ϕ with f_n.

Let PrOR_n be the partial function obtained by restricting the domain of OR_n to inputs of Hamming weight zero or one. It is known that the $\Omega(\sqrt{n})$ lower bound on the approximate degree of OR_n holds even for PrOR_n. One way to see this is by inspecting the proof of Theorem 23 via Minsky-Papert-symmetrization in Section 5.1.[21] It can also be easily checked that the dual witness

[21] The requirement that the approximating polynomial $p(x)$ be bounded in magnitude even at inputs outside of the promise is essential for this lower bound to hold. Indeed, $x \mapsto 1 - 2|x|$ is a degree-one polynomial that exactly computes OR_n on inputs of Hamming weight 0 and 1, but can take values of magnitude $\Omega(n)$ at inputs of larger Hamming weight.

ψ for $\widetilde{\deg}(\mathsf{OR}_n) \geq \Omega(\sqrt{n})$ constructed in Section 6.1 is a dual witness for this fact (i.e., ψ places at least an ε fraction of its ℓ_1-mass on inputs of Hamming 0 and 1, where ε can be a constant arbitrarily close to 1).

Recall that Theorem 44 asserts that for any Boolean functions $f : \{-1,1\}^m \to \{-1,1\}$ and g, if $d = \widetilde{\deg}(f)$ and $D = \widetilde{\deg}_{1-d/(16m)}(g)$, then

$$\widetilde{\deg}(f \circ g) \geq \Omega\left(d \cdot D\right).$$

The theorem was stated for total functions f, g, but the proof applies directly even when g is partial. Specifically, for a total function f and partial function g, let $f \circ g$ denotes the partial function defined over the domain of all $(x_1, \ldots, x_m) \in \left(\{-1,1\}^b\right)^m$ such that each x_i is in the domain of g. Then the dual witness constructed in the proof of Theorem 44 establishes that $\widetilde{\deg}(f \circ g) \geq \Omega(\widetilde{\deg}(f) \cdot \widetilde{\deg}_{1-d/(16m)}(g))$.[22] The claimed lower bound $\widetilde{\deg}(\mathsf{THR}^t_n) \geq \Omega(\sqrt{nt})$ then follows from Theorem 44 applied with $f = \mathsf{MAJ}_{2t}$ and $g = \mathsf{PrOR}_{n/(2t)}$, together with the fact that $(\mathsf{MAJ}_{2t} \circ \mathsf{PrOR}_{n/(2t)})(x) = \mathsf{THR}^t(x)$ for all x in the domain of $f \circ g$.

In fact, by exploiting slightly more specific properties of the dual witness for the fact that $\widetilde{\deg}(\mathsf{MAJ}_{2t}) = \Omega(t)$, one can show that the dual polynomial constructed above implies an $\Omega(\sqrt{nt})$ approximate degree lower bound for *any* symmetric function with a "jump" between Hamming weights $t-1$ and t for $t \leq n/2$. Specifically, the dual witness for MAJ_{2t} places almost all of its mass on inputs of Hamming weight $t-1$ and t, i.e., it is in fact a lower bound for MAJ_{2t} under the promise that the Hamming weight is either $t-1$ or t.

8 Beyond Block-Composed Functions

Section 7 showed that dual block composition can yield tight lower bounds for the approximate degree of a variety of block-composed functions. In Section 7.3 we, also saw a few examples of how dual block composition can help us understand the approximate degree of functions that are not (obviously) block composed. Many functions of great interest in quantum computing and complexity theory are not block compositions. Can we systematically apply dual block composition to understand the approximate degree of these functions?

This turns out to be possible. For many *non*-block-composed functions f_n on n-bit inputs, the approximate degree of f_n is *equivalent* to the approximate degree of a related block-composed function F_m defined over inputs of size $m \gg n$, but under the promise that the input to F has Hamming weight at most n.[23] That is, approximating f to error ε by a degree d polynomial is equivalent to constructing a degree d polynomial p over domain $\{-1,1\}^m$ such that

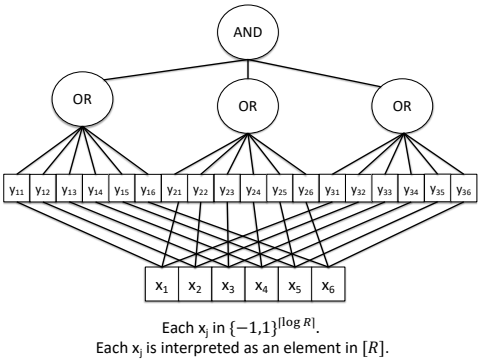
$$|p(x) - F(x)| \leq \varepsilon \text{ for all } |x| \leq n. \tag{32}$$

Note, crucially, that p is allowed to behave arbitrarily on inputs of Hamming weight larger than n.

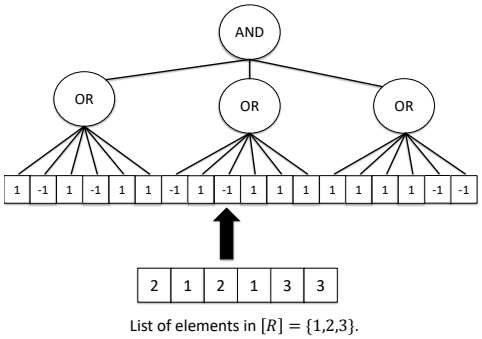
Let us denote by $F^{\leq n}$ the partial function obtained by restricting the domain of F to inputs of Hamming weight at most n, and by $\widetilde{\deg}_\varepsilon(F^{\leq n})$ the least degree of a polynomial p satisfying Condition (32). As we will see, if $F = f \circ g$ is a block composition of two functions whose

[22] Within the proof of the theorem, if g is a partial function, then the "error set" $\mathcal{E}_\phi$ for ϕ should be defined as $\{y\colon \operatorname{sgn}(\phi(y)) \neq g(y)\} \cup \{y\colon y \notin S\}$, where S is the domain of g.

[23] This contrasts with the constructions in Section 7.3, where we identified related block-composed functions over the same domain as the original function.



(a) Depiction of the right hand side of Equation (33) for SURJ with domain size $N = 6$ and range size $R = 3$.



(b) Specific example of the right hand size of Equation (33).

approximate degree is understood, then dual block composition can sometimes prove tight lower bounds on $\widetilde{\deg}_\varepsilon(F^{\leq n})$.

8.1 Surjectivity: A Case Study

The above connection between a non-block-composed function f and a block composed function F is best demonstrated with an example. Recall from Section 2.2 that the Surjectivity function (SURJ) takes as input a vector in $x \in \{-1,1\}^n$ with $n = N \log_2 R$. It interprets the vector as a list of (the binary representations of) N numbers $(k_1, \ldots, k_N)$ from range $[R] = \{1, \ldots, R\}$, and it outputs -1 if and only if for every $i \in [R]$, there is at least one index j such that $k_j = i$.

8.1.1 Approximate degree upper bound

We now relate SURJ to the block composition $\mathsf{AND}_R \circ \mathsf{OR}_N$. A natural way to do this is to consider representing the list $(k_1, \ldots, k_N) \in [R]^N$ via a set of $N \cdot R$ variables $y(x) = \{y_{i,j} : i \in [R], j \in [N]\}$ in which $y_{i,j} = -1$ if $k_j = i$ and $y_{i,j} = 1$ otherwise. Observe that each variable $y_{i,j}$ depends on only $\log_2 R$ bits of x, and moreover

$$\mathsf{SURJ}(x) = (\mathsf{AND}_R \circ \mathsf{OR}_N)(y(x)). \tag{33}$$

One can think of the input x to SURJ as a compressed representation of the input $y(x)$ to $\mathsf{AND}_R \circ \mathsf{OR}_N$, in that $y(x)$ consists of $N \cdot R$ bits while x consists of just $N \log_2 R$ bits. See Figures 4a and 4b for a depiction and example.

A key observation is that for *any* input x to SURJ, the Hamming weight of the corresponding vector $y(x)$ is exactly N. This means that if p approximates $(\mathsf{AND}_R \circ \mathsf{OR}_N)^{\leq N}$ to error ε then $p(y(x))$ approximates SURJ to error ε, and has degree at most $\deg(p) \cdot \log_2 R$. Crucially, this holds regardless of how p behaves on inputs in $\{-1,1\}^{R \cdot N}$ of Hamming weight more than N.

Observation 52. $\widetilde{\deg}_\varepsilon(\mathsf{SURJ}) \leq \widetilde{\deg}_\varepsilon\left((\mathsf{AND}_R \circ \mathsf{OR}_N)^{\leq N}\right) \cdot \log_2(R)$.

We've already seen that $\widetilde{\deg}(\mathsf{AND}_R \circ \mathsf{OR}_N) = \Theta(\sqrt{RN})$ (see Theorems 11 and 42). It turns out that $\mathsf{AND}_R \circ \mathsf{OR}_N$ is substantially easier to approximate when the approximation only needs to be

accurate on inputs of Hamming weight at most N. Multiple proofs of this upper bound are known [She18a, BKT18]. Here we describe the approximating polynomial from [She18a].

Theorem 53. $\widetilde{\deg}\left((\mathsf{AND}_R \circ \mathsf{OR}_N)^{\leq N}\right) \leq O(R^{1/4} \cdot N^{1/2})$.

Proof. Let q be a polynomial over domain $\{-1,1\}^R$ of degree $O(\sqrt{R})$ that approximates AND_R to error $1/4$. A change of basis argument allows us to express q as a linear combination of *disjunctions*, i.e, terms of the form $\mathsf{OR}_S(x) = \vee_{i \in S} x_i$ for some subset $S \subseteq [R]$.[24] Moreover, the sum of the magnitudes of the coefficients in the linear combination is at most $2^{O(\sqrt{R})}$.

Clearly $|q \circ \mathsf{OR}_N - \mathsf{AND}_R \circ \mathsf{OR}_N| \leq 1/4$. Because the composition of any two disjunctions is itself a disjunction, $q \circ \mathsf{OR}_N$ is itself a linear combination of disjunctions over domain $\{-1,1\}^{RN}$ in which the sum of the magnitudes of the coefficients is at most $W \leq 2^{O(\sqrt{R})}$. Let us write this linear combination as

$$(q \circ \mathsf{OR}_N)(y) = \sum_{S \subseteq \{-1,1\}^{RN}} c_S \cdot \mathsf{OR}_S(y). \tag{34}$$

Here is where we exploit the fact that we only require our final approximation to accurately approximate $\mathsf{AND}_R \circ \mathsf{OR}_N$ on inputs of Hamming weight at most N. A generalization of the construction in Theorem 14 shows that $\widetilde{\deg}_\varepsilon(\mathsf{OR}_{R \cdot N}^{\leq N}) \leq O\left(\sqrt{N \log(1/\varepsilon)}\right)$ for any $\varepsilon > 0$ regardless of $R \cdot N$. Note that the approximating polynomial may take values that are exponentially large in its degree when evaluated at inputs x of Hamming weight more than N.

Now set $\varepsilon = 1/(12W)$, and let us replace each disjunction OR_S on the right hand side of Equation (34) with an ε-approximation to $\mathsf{OR}_S^{\leq N}$. The resulting polynomial p has degree $O(\sqrt{N \log W}) = O(R^{1/4} N^{1/2})$. On any input y of Hamming weight at most N, we have $|\,(q \circ \mathsf{OR}_N)\,(y) - p(y)| \leq 1/12$ and hence $|\,(\mathsf{AND}_R \circ \mathsf{OR}_N)\,(y) - p(y)| \leq 1/12 + 1/4 = 1/3$. □

8.1.2 Approximate degree lower bound

One might suspect that the approximation for SURJ constructed above is unnecessarily tying its own hands by ignoring all structure in the vector $y(x)$ besides the fact that $y(x)$ has Hamming weight at most N. For example, it is ignoring the fact that for each $j \in [N]$, $y_{i,j} = -1$ for *exactly* one index $i \in [R]$. It turns out that this additional structure in the vector $y(x)$ cannot be leveraged by low-degree polynomials. That is, the approximate degree of SURJ is not just upper bounded by that of $(\mathsf{AND}_R \circ \mathsf{OR}_N)^{\leq N}$, but in fact is equivalent to it.

Lemma 54. $\widetilde{\deg}_\varepsilon(\mathsf{SURJ}) \geq \tilde{\Omega}\left(\widetilde{\deg}_\varepsilon((\mathsf{AND}_R \circ \mathsf{OR}_N)^{\leq N})\right)$.

Lemma 54 was shown in [BT19b] using a symmetrization argument due to Ambainis [Amb05]. We defer a proof until Section 8.4. A tight lower bound on the approximate degree of SURJ now follows from one for $\widetilde{\deg}((\mathsf{AND}_R \circ \mathsf{OR}_N)^{\leq N})$, which can be proved by dual block composition.

Theorem 55 ([BKT18])**.** $\widetilde{\deg}((\mathsf{AND}_R \circ \mathsf{OR}_N)^{\leq N}) \geq \tilde{\Omega}(R^{1/4} \cdot N^{1/2})$.

[24] That is, the set of disjunctions over $\{-1,1\}^n$ form a basis for the 2^n-dimensional vector space of functions $f : \{-1,1\}^n \to \mathbb{R}$. This is because there are 2^n disjunctions which we can see to be linearly independent, inductively, as follows. Suppose instead that we could write some $\mathsf{OR}_S = \sum_{T \in \mathcal{T}} a_T \mathsf{OR}_T$ where $T \setminus S \neq \emptyset$ for every $T \in \mathcal{T}$. Then on input $y \in \{-1,1\}^n$ where $y_i = -1 \iff i \notin S$, we have $\mathsf{OR}_S(y) = 1$ while $\mathsf{OR}_T(y) = -1$ for every $T \in \mathcal{T}$, proving that $\sum_{T \in \mathcal{T}} a_T < 0$. Meanwhile, $\mathsf{OR}_S(\mathbf{1}_n) = 1 = \mathsf{OR}_T(\mathbf{1}_n)$ for every T, so $\sum_{T \in \mathcal{T}} a_T < 0$, a contradiction.

Proof sketch. Let ψ be any dual polynomial for the fact that $\widetilde{\deg}_{7/8}(\mathsf{AND}_R) \geq \Omega(R^{1/2})$, and let $N' \coloneqq N/R^{1/2}$. It turns out to be useful to focus on the function $(\mathsf{AND}_R \circ \mathsf{OR}_{N'})^{\leq N}$ rather than $(\mathsf{AND}_R \circ \mathsf{OR}_N)^{\leq N}$ (the former is a subfunction of the latter, so a lower bound for the former will imply our desired lower bound for the latter).

Let ϕ be the dual polynomial for $\deg_{7/8}(\mathsf{OR}_{N'}) \geq \Omega(\sqrt{N'})$ constructed in Section 6.1. Lemmas 39-41 show that $\psi \star \phi$ is a dual polynomial for the fact that $\widetilde{\deg}(\mathsf{AND}_R \circ \mathsf{OR}_{N'}) \geq \Omega(\sqrt{R \cdot N'}) \geq \Omega(R^{1/4} \cdot N^{1/2})$. Unfortunately, this is not enough, as we need our degree lower bound to hold against polynomials that can behave arbitrarily on inputs of Hamming weight larger than N, i.e., we must lower bound $\widetilde{\deg}((\mathsf{AND}_R \circ \mathsf{OR}_{N'})^{\leq N})$.

The property making this possible is that $|\psi \star \phi|$ places *very little* mass on inputs of Hamming weight larger than N. Quantitatively,

$$\sum_{y \in \{-1,1\}^{R \cdot N'} \colon |y| > N} |(\psi \star \phi)(y)| \leq 2^{-\Omega(N/\sqrt{N'})} = 2^{-\Omega(R^{1/4} N^{1/2})}. \tag{35}$$

At a high level, this bound arises as follows. Theorem 34 shows that $|\phi|$ places most of its mass on inputs of very low Hamming weight. In particular, an exponentially small fraction of its mass lies on inputs of Hamming weight more than $\sqrt{N'}$. Recall that the probability distribution $|\psi \star \phi|$ can be thought of as first choosing z according to the distribution $|\psi|$, and then choosing $y = (y_1, \ldots, y_R) \in \left(\{-1,1\}^{N'}\right)^R$ from the product distribution $\otimes_{i=1}^R \phi_{z_i}$. Because $|\phi|$ (and hence also ϕ_{+1} and ϕ_{-1}) places such little mass on inputs of Hamming weight more $\sqrt{N'}$, it turns out that for $y = (y_1, \ldots, y_R) \sim \otimes_{i=1}^R \phi_{z_i}$, the probability that y has Hamming weight greater than N is dominated by the probability of the following event: there are at least $\ell \coloneqq N/\sqrt{N'}$ values of i for which $|y_i| \approx \sqrt{N'}$. And this probability is exponentially small in ℓ.

We now explain how Condition (35) implies that $\widetilde{\deg}((\mathsf{AND}_R \circ \mathsf{OR}_{N'})^{\leq N}) \geq d$ for

$$d = R^{1/4} \cdot N^{1/2} / \log N.$$

Suppose p approximates $\mathsf{AND}_R \circ \mathsf{OR}_{N'}$ for all inputs of Hamming weight at most N. Then in particular, $|p(y)| \leq 4/3$ for all $|y| \leq d < N$. An interpolation argument of Razborov and Sherstov shows that this implies p is bounded in magnitude by $\exp(\tilde{O}(d))$ for *all* inputs, even those of very large Hamming weight.

Lemma 56 ([RS10])**.** Let $p \colon \{-1,1\}^{R \cdot N} \to \mathbb{R}$ be a polynomial of degree at most d. If $|p(y)| \leq O(1)$ for all $|y| \leq N$, then $|p(y)| \leq (RN)^{O(d)}$ for all $y \in \{-1,1\}^{RN}$.

Hence, we conclude that $|p(y)| \leq (RN)^{O(d)} = 2^{O(R^{1/4} N^{1/4})}$ for all $y \in \{-1,1\}^{RN}$, where we have used that $d = R^{1/4} \cdot N^{1/2} / \log N$. Now recall that, as captured in Claim 37, if a dual polynomial for a function F places mass at most δ on a set S, then it in fact lower bounds the degree of polynomial approximations p to F that are permitted to be as large as roughly $1/\delta$ at inputs in S. Taking S to be the set of all inputs of Hamming weight greater than N and $\delta = 2^{-\Omega(R^{1/4} N^{1/2})}$, Condition (35) thus implies that p requires degree at least d. This completes the proof. □

8.1.3 Threshold degree of SURJ

The facts that Observation 52 and Lemma 54 hold for every $\varepsilon > 0$ imply that, up to logarithmic factors, $\deg_{\pm}(\mathsf{SURJ})$ is equivalent to $\deg_{\pm}(\mathsf{AND}_R \circ \mathsf{OR}_N)^{\leq N})$. As explained next, this latter quantity is $\tilde{\Theta}(n^{1/2})$, where recall that $n = N \cdot \log_2 R$.

For the upper bound, observe that, even without the promise that the input has Hamming weight at most N, the threshold degree upper bound for CNFs given in Lemma 8 guarantees that $\deg_{\pm}(\mathsf{AND}_R \circ \mathsf{OR}_N)$ is at most $O(\sqrt{N \log R}) = \tilde{O}(\sqrt{N})$.

To prove a matching lower bound, observe that if $R = N^{1/2}$, then $\mathsf{AND}_R \circ \mathsf{OR}_N$ is simply the Minsky-Papert CNF defined over $N^{3/2}$ input bits. This function has threshold degree $\Omega\left((N^{3/2})^{1/3}\right) = \Omega\left(N^{1/2}\right)$ (Theorem 26).

By itself, this does not yield the lower bound we require: we need to show that $\mathsf{AND}_{N^{1/2}} \circ \mathsf{OR}_N$ has threshold degree $\Omega(N^{1/2})$ *even under the promise* that that the input has Hamming weight at most N. Note that, here, N is much less than the maximal possible Hamming weight, of $N^{3/2}$.

Fortunately, more recent and general proofs of Minsky and Papert's lower bound, which are based on dual block composition (see Theorems 48 or the proof of Theorem 92), *can* be extended to prove that $\deg_{\pm}\left((\mathsf{AND}_{N^{1/2}} \circ \mathsf{OR}_N)^{\leq N}\right) \geq \Omega(N^{1/2})$. One simply combines the known constructions of dual witnesses for the high threshold degree of the Minsky-Papert CNF with with the analysis used to prove Theorem 55. The interested reader is directed to [BT19a] for details.

Theorem 57 ([BT19a])**.** The threshold degree of SURJ is $\tilde{\Theta}(n^{1/2})$.

8.2 Other Functions and Applications to Quantum Query Complexity

A number of other problems that arise in quantum query complexity can be related to block-composed functions under a Hamming weight promise. Recall from Section 2.2 that the k-distinctness function k-ED interprets its input as a list of N numbers from a range of size R and outputs 1 if and only if there is some range item that appears at least k times in the list. It is easy to see that $k\text{-}\mathsf{ED}(x) = (\mathsf{NOR}_R \circ \mathsf{THR}_N^k)(y(x))$ where THR^k denotes the symmetric k-threshold function that outputs -1 iff its input has Hamming weight at least k. Analogously to Theorem 55, we have:

Lemma 58. For $k \geq 2$, $\widetilde{\deg}(k\text{-}\mathsf{ED}) = \tilde{\Theta}\left(\widetilde{\deg}\left(\left(\mathsf{NOR} \circ \mathsf{THR}_N^k\right)^{\leq N}\right)\right)$.

Dual block composition can be used to show that $\widetilde{\deg}(\left(\mathsf{NOR} \circ \mathsf{THR}_N^k\right)^{\leq N}) \geq \Omega(N^{3/4-1/(4k)})$ for any constant $k \geq 2$ [BKT18, MTZ20]. For large k, this nearly matches a known upper bound of $O\left(n^{3/4-\frac{1}{2^{k+2}-4}}\right)$ on the quantum query complexity, and hence also approximate degree, of k-ED [Bel12]. Similar connections give tight lower bounds (up to logarithmic factors) on both the approximate degree and quantum query complexity of various property testing problems, including junta testing, statistical distance estimation, entropy approximation, and image size testing [BKT18].

Converses to the polynomial method in quantum query complexity. Recall that while approximate degree lower bounds imply quantum query lower bounds (Theorem 70), the converse is not true [Amb06, She18a]. For example, we have seen that $\widetilde{\deg}(\mathsf{SURJ}) = \tilde{\Theta}(n^{3/4})$, but it is known that its quantum query complexity is $\Theta(n)$ [BM12, She18d]. However, partial converses are possible. This means that if one proves an approximate degree upper bound for a function, and the approximating polynomial satisfies additional properties, then in fact an efficient quantum query algorithm may be implied. For example, Arunachalam, Briët, and Palazuelos [ABP19] showed that quantum query complexity is *characterized* by one of these variants, called approximation by completely-bounded forms. The characterization has so far been used primarily to study fine-grained relationships between constant-query quantum algorithms and constant-degree polynomials (see [BG22] and the references therein)—no one has yet been able to use this characterization to

give a new quantum algorithm for any natural problem. Can known constructions of approximating polynomials be modified to yield completely-bounded forms? If so, this has the potential to offer a new paradigm in quantum algorithm design.

8.3 Approximate Degree of AC^0

One of our favorite open questions in the study of approximate degree is to ascertain whether there are AC^0 circuits of approximate degree $\Omega(n)$. The Parity and Majority functions have linear approximate degree, but they are not in AC^0. For a long time, the best known lower bound on the approximate degree of an AC^0 function was $\Omega(n^{2/3})$, proved by Aaronson and Shi [AS04]. Analyzing non-block-composed functions, as described above, brings us a lot closer to answering this question. In particular, SURJ is in AC^0 and has approximate degree $\tilde{\Theta}(n^{3/4})$. In fact, the key to the SURJ lower bound (Theorem 55) can be seen as another hardness amplification theorem, showing that the function $(\mathsf{AND}_R \circ \mathsf{OR}_N)^{\leq N}$ requires higher degree to approximate than does AND_R itself. The main property of AND_R used in Theorem 55 is that it has approximate degree $\Omega(\sqrt{R})$. Simplifying the actual construction slightly, replacing AND_R with SURJ_R yields a function $(\mathsf{SURJ}_R \circ \mathsf{OR}_N)^{\leq N}$ that has even larger approximate degree $\tilde{\Omega}(n^{7/8})$.[25]

By iteratively applying this hardness amplification technique, for any $\delta > 0$, one can obtain a family of AC^0 circuits with approximate degree $\Omega(n^{1-\delta})$ [BT19b, BKT18]. This was further improved by the authors from (1/3)-approximate degree to $(1 - 2^{-n^{1-\delta}})$-approximate degree [BT19a], and finally by Sherstov and Wu [SW19] to a $\Omega(n^{1-\delta})$ lower bound on the threshold degree of AC^0. Recent work of Sherstov [She22] establishes that the $\Omega(n^{1-\delta})$ approximate degree lower bound holds even for DNFs and CNFs of constant width.

Open problems. As indicated above, several problems in this research direction remain open. One is to ascertain whether the $\Omega(n^{1-\delta})$ threshold degree lower bound holds for depth-three AC^0 circuits, as current lower bound constructions require the circuit depth to grow with $1/\delta$. Another is to close the gap between the lower bounds above, which are all of the form $\Omega(n^{1-\delta})$, and the known approximate degree upper bounds for AC^0, which are all trivial (i.e., $O(n)$).

This gap may appear inconsequential—is there really a major difference between approximate degree $\Theta(n^{0.999})$ and $\Theta(n)$? However, we will see (Section 11.3) that even "barely sublinear" approximate degree upper bounds have important implications in circuit complexity. Hence, in our view, the gap between the known (sublinear) approximate degree lower bounds for AC^0 and the (trivial) linear upper bound is significant.

Open Problem 59. Exhibit a function in AC^0 with approximate degree $\Omega(n)$ or $\Omega(n/\log n)$, or prove that no such function exists.

8.4 Proof of Lemma 54

Let $\widetilde{\deg}_\varepsilon\left((\mathsf{AND}_R \circ \mathsf{OR}_N)^{=N}\right)$ denote the least degree of a real polynomial $p\colon \{-1,1\}^{R\cdot N} \to \{-1,1\}$ such that $|p(x) - (\mathsf{AND}_R \circ \mathsf{OR}_N)(x)| \leq \varepsilon$ for all x of Hamming weight exactly N. Note that p

[25] The function in the actual construction is $\mathsf{SURJ}_R \circ \mathsf{AND}_{O(\log R)} \circ \mathsf{OR}_N$. The "middle gadget" $\mathsf{AND}_{O(\log R)}$ is included in the function definition because $\widetilde{\deg}_\varepsilon(\mathsf{AND}_{O(\log R)} \circ \mathsf{OR}_N)$ is large even for $\varepsilon \geq 1 - 1/(3R)$ (see Theorem 47). This enables the use of dual block composition (see Theorem 44) to prove a lower bound on the approximate degree of $\mathsf{SURJ}_R \circ \mathsf{AND}_{O(\log R)} \circ \mathsf{OR}_N$.

may behave arbitrarily on inputs of Hamming weight strictly less than or strictly greater than N. To begin, rather than proving Lemma 54 itself, we prove the weaker result that $\widetilde{\deg}_\varepsilon(\mathsf{SURJ}) \cdot \log R \geq \widetilde{\deg}_\varepsilon\left((\mathsf{AND}_R \circ \mathsf{OR}_N)^{=N}\right)$, as this contains the main ideas. At the very end of the section (Section 8.4.1), we sketch how to prove Lemma 54 in full.

Recall that the Surjectivity function (SURJ) takes as input a vector in $x \in \{-1,1\}^n$ with $n = N \log_2 R$ and interprets the vector as a list of (the binary representations of) N numbers $(k_1, \ldots, k_N)$ from range $[R] = \{1, \ldots, R\}$. Our approximate degree upper bound for SURJ (Theorem 53) introduced a different representation of the list $(k_1, \ldots, k_N)$ via $N \cdot R$ variables $y(x) = \{y_{i,j} \colon i \in [R], j \in [N]\}$ in which $y_{i,j} = -1$ if $k_j = i$ and $y_{i,j} = 1$ otherwise.

Up to a $\log_2 R$ factor, these two representations of the list, namely x and y, are equivalent from the perspective of low-degree polynomial approximations, as formalized by the following proposition.

Claim 60. Let $F \colon \{-1,1\}^n \to \{-1,1\}$ be any function. For any $\varepsilon > 0$, let $p \colon \{-1,1\}^n \to \mathbb{R}$ be the lowest degree polynomial such that $|p(x) - F(x)| \leq \varepsilon$ for all $x \in \{-1,1\}^n$, and $q \colon \{-1,1\}^{N \cdot R} \to \mathbb{R}$ be the lowest degree polynomial such that $|q(y(x)) - F(x)| \leq \varepsilon$ for all $x \in \{-1,1\}^n$. Then

$$\deg(p) \leq \deg(q) \cdot \log R \tag{36}$$

and

$$\deg(q) \leq \deg(p). \tag{37}$$

Proof. Equation (36) holds because each bit $y_{i,j}$ of y depends on only $\log R$ bits of x; hence, given any q such that $|q(y(x)) - F(x)| \leq \varepsilon$, $p(x) \coloneqq q(y(x))$ is a polynomial of degree at most $\deg(q) \cdot \log(R)$ that approximates F to error ε. The second claim holds because each bit of x is a degree-1 function in entries $(y_{1,1}, \ldots, y_{R,N})$ of $y(x)$. That is, if we express a string $x \in \{-1,1\}^n$ as $x = (x_1, \ldots, x_N)$ where each $x_j \in \{-1,1\}^{\log R}$, then using y as shorthand for $y(x)$, it holds that

$$x_{j,k} = 1 - \sum_{i \colon \mathsf{bin}(i)_k = -1} (1 - y_{i,j}). \tag{38}$$

Here, $\mathsf{bin}(i)$ denotes the $\log R$-bit binary representation of range element $i \in [R]$.

Hence, if $|p(x) - F(x)| \leq \varepsilon$ for all $x \in \{-1,1\}^n$, let $q(y) \colon \{-1,1\}^{N \cdot R} \to \mathbb{R}$ be the polynomial that replaces each input $x_{i,j}$ to p with the right hand side of Equation (38). Since the right hand side of Equation (38) is a degree-1 polynomial in y, $\deg(q) \leq \deg(p)$ as claimed. □

Claim 60 shows that constructing a low-degree approximating polynomial for the *total* function

$$F(x) = \mathsf{SURJ}(x) \colon \{-1,1\}^n \to \{-1,1\}$$

is *equivalent* (up to a logarithmic factor in the degree) to approximating the *partial* function defined over the $y_{i,j}$ variables (in which the approximating polynomial is allowed to behave arbitrarily on inputs in $\{-1,1\}^{N \cdot R}$ that do not equal $y(x)$ for some $x \in \{-1,1\}^n$). That is, from the perspective of low-degree polynomials, the "$x \in \{-1,1\}^n$ representation" of the input is *equivalent* to the "$y \in \{-1,1\}^{N \cdot R}$ representation" of the same input. Henceforth, we refer to these respective representations of the input to SURJ simply as the "x-representation" and the "y-representation".

The crux of Lemma 54 is to identify a *third* equivalent representation of the input. This representation is *not* a bit-vector like the x-representation or y-representation, but rather a "frequency vector", meaning a vector of non-negative integers summing to N. Specifically, given an

$x \in \{-1,1\}^n$ interpreted as a list of N numbers $(k_1, \ldots, k_N)$ from range $[R] = \{1, \ldots, R\}$, define $z(x) = (z_1, \ldots, z_R) \in ([N]^*)^R$ where z_i is the number of times range item $i \in [R]$ appears in the list specified by x. For instance, in the example of Figure 4b, the associated frequency vector $z = (2, 2, 2)$, because each of the three range items appears twice in the input list.

Claim 61 below shows that for purposes of constructing an approximating polynomial for $\mathsf{SURJ}(x)$, it is without loss of generality to represent the input list via its frequency vector z. The only property of SURJ used in the claim is that $\mathsf{SURJ}(x)$ depends only on the frequency vector $z(x)$.

In conclusion, we can summarize Claim 60 and Claim 61 as follows: Let $n = N \log R$ and $F\colon \{-1,1\}^n \to \{-1,1\}$ be any function that interprets its input x as a list of N numbers from a range of size R, such that $F(x)$ depends only on the frequency vector $z(x)$ of the input $x \in \{-1,1\}^n$. Then approximating F in its "x-representation" and its "y-representation" are equivalent up to a factor of $\log R$ in degree, while approximating F in its "y-representation" and "z-representation" are *perfectly* equivalent (with no change whatsoever in the degree).

Claim 61 (Ambainis [Amb05])**.** Let $F\colon \{-1,1\}^n \to \{-1,1\}$ be any function such that $F(x)$ depends only on the frequency vector $z(x)$. For any $\varepsilon > 0$, let $q\colon \{-1,1\}^{N \cdot R} \to \mathbb{R}$ be the lowest degree polynomial such that

$$|q(y(x)) - F(x)| \leq \varepsilon \text{ for all } x \in \{-1,1\}^n, \tag{39}$$

and let $P\colon ([N]^*)^R \to \mathbb{R}$ be the lowest degree polynomial such that

$$|P(z(x)) - F(x)| \leq \varepsilon \text{ for all } x \in \{-1,1\}^n. \tag{40}$$

Then $\deg(q) = \deg(P)$.

We clarify that, in Claim 61, while it is without loss of generality to assume that q is multilinear owing to its domain being $\{-1,1\}^{R \cdot N}$, the polynomial P may not be multilinear.

Before proving the claim, we explain why it implies the desired result that $\widetilde{\deg}_\varepsilon(\mathsf{SURJ}) \cdot \log R \geq \widetilde{\deg}_\varepsilon\left((\mathsf{AND}_R \circ \mathsf{OR}_N)^{=N}\right)$. By Claims 60 and 61, if $\widetilde{\deg}_\varepsilon(\mathsf{SURJ}) \leq d$, then there is a polynomial P of degree at most $d \cdot \log R$ such that

$$|P(z(x)) - \mathsf{SURJ}(x)| \leq \varepsilon \text{ for all } x \in \{-1,1\}^n. \tag{41}$$

We now use P to construct an approximation Q to $(\mathsf{AND}_R \circ \mathsf{OR}_N)^{=N}$. Let $w = (w_1, \ldots, w_R) \in \{-1,1\}^{R \cdot N}$, and define $Q(w) = P(|w_1|, \ldots, |w_R|)$. Since P has total degree at most d, and $|w_i|$ is a linear function of w_i, Q has total degree d as well. If $|w| = N$, then $|w_1|, \ldots, |w_R|$ are non-negative integers summing to N and hence $w = y(x)$ for some $x \in \{-1,1\}^n$. By Equation (41), this implies that $|Q(w) - (\mathsf{AND}_R \circ \mathsf{OR}_N)^{=N}(w)| \leq \varepsilon$.

Proof of Claim 61. The fact that $\deg(q) \leq \deg(P)$ holds because if $y = y(x)$ and $z = z(x)$, then

$$z_i = \sum_{j \in [N]} \frac{1 - y_{i,j}}{2}. \tag{42}$$

That is, each entry of z_i is a degree-1 function of y. Hence, if $P(z)$ satisfies Equation (40), then replacing each input z_i to P with the right hand size of Equation (42) yields a polynomial q satisfying Equation (39) with $\deg(q) \leq \deg(p)$.

The fact that $\deg(P) \leq \deg(q)$ is far less straightforward. Let $q(y)$ be a multilinear polynomial satisfying Equation (39) for a vector $z \in ([N]^*)^R$ of non-negative integers summing to N, and define $P(z)$ to be the expected value of $q(y(x))$ over all inputs x such that $z = z(x)$. Clearly, since q satisfies Equation (39), P satisfies Equation (40). All that remains is to show that $P(z)$ can be written as a polynomial of degree at most $d \coloneqq \deg(q)$.

We begin by mimicking the analysis of Minsky-Papert symmetrization (Lemma 22): by linearity of expectation, it is enough to assume that $q(y(x))$ consists of a single monomial. Applying the variable transformation $y_{i,j} \mapsto (1 - y_{i,j})$ does not alter the degree of q, and hence it is also without loss of generality to assume that q has the form:

$$q(y) = 2^{-d} \cdot (1 - y_{i_1,j_1}) \cdot (1 - y_{i_2,j_2}) \cdots (1 - y_{i_d,j_d}), \tag{43}$$

where each $i_k \in [R]$ and $j_k \in [N]$, and $i_1 \leq i_2 \leq \cdots \leq i_d$. Note that in this case $q(y)$ equals 1 if $y_{i_1,j_1} = y_{i_2,j_2} = \cdots = y_{i_d,j_d} = -1$ and otherwise $q(y) = 0$.

Some intuition. Before completing the calculation, let us give some intuition as to why $\deg(P(z)) \leq d$. Fix a $z \in ([N]^*)^R$ of non-negative integers summing to N, and let us define X as the set of all x such that $z(x) = z$. Note that if $x \sim X$, then x is a random input with frequency vector z. Let us define

$$Y = \{y(x) \colon x \in X\}. \tag{44}$$

Recall that $P(z)$ is defined to equal the expected value of $q(y)$, where $y \sim Y$. For each $i \in [R]$ and $j \in [N]$, $\Pr_{y \sim Y}[y_{i,j} = -1] = z_i/N$. This is because $y_{i,j} = -1$ if and only if that j'th entry of the input list equals range item i, and every $y \sim Y$ represents an input with exactly z_i occurrences of range item i. This immediately implies that if $q(y)$ has degree 1 then $P(z)$ does as well.

Imagine (counterfactually) that the entries of y were all independent. Then the probability that $y_{i_1,j_1} = y_{i_2,j_2} = \cdots = y_{i_d,j_d} = -1$ would equal $\frac{1}{N^d} z_{i_1} \cdot z_{i_2} \cdots z_{i_d}$, and hence the claim that $\deg(P(z)) \leq \deg(q(y))$ would be clear. This calculation is analogous to t-biased symmetrization (Lemma 21), in which an n-variate polynomial $q(y)$ was transformed into a univariate polynomial $P(t)$ by taking the expected value of $q(y)$ under a distribution in which each coordinate of y was chosen *independently* to have expected value t.

However, when $y \sim Y$, the entries of y are not independent. For example, if $z_1 = 1$, then conditioned upon $y_{1,1} = -1$, we know with certainty that $y_{1,j} = 1$ for all $j \neq 1$. This is analogous to how, in Minsky-Papert symmetrization (Lemma 22), an n-variate polynomial $q(y)$ was transformed into a univariate polynomial $P(t)$ by taking the expected value of $q(y)$ under a distribution in which y was chosen to have Hamming weight exactly t, and hence the coordinates of y were *not* independent. Fortunately, just as with Lemma 22, the dependencies introduced turn out not to increase the degree of P relative to the independent case.

Completing the calculation. For notational convenience, let us express $q(y)$ as

$$2^{-d} \prod_{i \in [R]} \prod_{j \in A_i} (1 - y_{i,j})$$

where $A_i \subseteq [d]$ is the set of all j such that variable $y_{i,j}$ appears in the right hand side of Equation (43). If the A_i sets are not pairwise disjoint, then $q(y(x)) = 0$ for all $x \in \{-1, 1\}^N$, meaning that P has degree $0 \leq \deg(q)$ as desired. This is because for each $j \in [N]$, $y(x)_{i,j}$ can only equal -1 for

exactly one value of i, as the jth item of the input list can only equal one range item. For example, if $q(y) = (1/4) \cdot (1 - y_{1,1}) \cdot (1 - y_{2,1})$, then $q(y(x))$ will equal 0 for all $x \in \{-1, 1\}^n$.

So henceforth let us assume that the A_i's are pairwise disjoint. Fix a vector $z = (z_1, \ldots, z_N)$ of nonnegative integers summing to N. Then:

$$\begin{aligned} \mathbb{E}_{y \sim Y}[q(y)] &= \Pr[y_{i_k,j_k} = -1 \text{ for all } k = 1, \ldots, d] \\ &= \Pr[y_{i_1,j_1} = -1] \cdot \Pr[y_{i_2,j_2} = -1 | y_{i_1,j_1} = -1] \cdot \cdots \cdot \Pr[y_{i_d,j_d} = -1], \end{aligned} \tag{45}$$

where the expectation is taken over $y \sim Y$ (see Equation (44)).

The calculation of the right hand side of Equation (45) in terms of the entries of z is best illustrated with an example. Suppose that $q(y) = \frac{1}{8}(1-y_{1,1})(1-y_{1,2})(1-y_{2,3})$. Then $\Pr[y_{1,1} = -1]$ is the probability that, out of the z_1 occurrences of range item 1, one of them is at index 1 of the input list. This probability is exactly z_1/N. Then, conditioned on $y_{1,1} = -1$, the probability that $y_{1,2} = -1$ is the probability that, out of the remaining $z_1 - 1$ occurrences of range item 1 in the input list, one of them is at index 2 of the input list. This probability is exactly $(z_1 - 1)/(N - 1)$. This is because, once we condition on the first item of the list equalling range item 1, there are $z_1 - 1$ remaining occurrences of range item 1 elsewhere in the list, and $N - 1$ indices at which those $z_1 - 1$ occurrences may reside, namely indices $\{2, 3, \ldots, N\}$.

Then, conditioned on both $y_{1,1} = -1$ and $y_{1,2} = -1$, the probability that $y_{2,3} = -1$ is the probability that, out of the z_2 occurrences of range item 2 in the input list, one of them is at index 3 of the input list. This probability is exactly $z_2/(N - 2)$. Hence, the right hand side of Equation (45) equals $\frac{1}{N(N-1)(N-2)} \cdot z_1(z_1 - 1)z_2$, which is polynomial in $z = (z_1, \ldots, z_R)$ of total degree $3 = \deg(q)$.

In general, letting $A_{<i} = \cup_{k=1}^{i-1} A_i$, the right hand size of Equation (45) equals:

$$\prod_{i \in [R]} C_i \cdot z_i \cdot (z_i - 1) \ldots (z_i - |A_i| + 1)$$

where

$$C_i = (N - |A_{<i}|)(N - |A_{<i}| - 1) \cdot \cdots \cdot (N - |A_{<i}| - |A_i| + 1).$$

Each factor in this expression is of the form $z_i \cdot (z_i - 1) \ldots (z_i - |A_i| + 1)$ times some factor C_i that is independent of z. Hence, this is a polynomial in $z = (z_1, \ldots, z_R)$ of total degree at most $\sum_{i \in [R]} |A_i| = d$.

□

8.4.1 Obtaining the full lemma

Above, we proved that

$$\widetilde{\deg}_\varepsilon(\mathsf{SURJ}) \cdot \log R \geq \widetilde{\deg}_\varepsilon \left((\mathsf{AND}_R \circ \mathsf{OR}_N)^{=N} \right),$$

while Lemma 54 had $\leq N$ rather than $=N$ as the superscript on the right hand side of the inequality. To prove the full lemma, we need to introduce a slight variant of SURJ, that we call dSURJ. This variant extends the range of SURJ by one, by adding a "dummy range element" whose presence or absence in the input list does not affect the output. That is, dSURJ tests whether each range element *other than the designated dummy range element* appears at least once in the input list.

One then proceeds in a two-step analysis. First, one shows that approximating dSURJ is no harder than approximating SURJ itself. That is, any degree-d ε-approximating polynomial for SURJ can be transformed into one for dSURJ. Essentially, the transformation "hard-codes" in one copy of the dummy range element into the input list to SURJ, so that the presence or lack of the dummy range element in the "real" input list to SURJ no longer affects its output.

Second, one shows that, up to a logarithmic factor, any approximating polynomial for dSURJ yields an approximating polynomial for $(\mathsf{AND}_R \circ \mathsf{OR}_N)^{\leq N}$. The two steps together then imply as desired that

$$\widetilde{\deg}_\varepsilon(\mathsf{SURJ}) \geq \widetilde{\deg}_\varepsilon(\mathsf{dSURJ}) \geq \tilde{\Omega}\left(\widetilde{\deg}_\varepsilon\left((\mathsf{AND}_R \circ \mathsf{OR}_N)^{\leq N}\right)\right).$$

To prove the second step, one applies the argument we already covered to conclude that any approximating polynomial for dSURJ (with R non-dummy range elements) implies an approximation for the following slight modification of $(\mathsf{AND}_R \circ \mathsf{OR}_N)^{=N}$, which we denote by $F^{=N}$. In $F^{=N}$, there are $R+1$ rather than R copies of OR_N, but the final copy of OR_N is simply ignored by the function. This corresponds to how the dummy range element is ignored by dSURJ.

The final piece of the argument is to show that $\widetilde{\deg}_\varepsilon(F^{=N}) \geq \widetilde{\deg}_\varepsilon\left((\mathsf{AND}_R \circ \mathsf{OR}_N)^{\leq N}\right)$. To see this, let $p\colon \{-1,1\}^{N\cdot(R+1)} \to \mathbb{R}$ be a polynomial of total degree d that ε-approximates $F^{=N}$. We transform p into an ε-approximation of the same degree for $(\mathsf{AND}_R \circ \mathsf{OR}_N)^{\leq N}$.

To accomplish this, consider the block-wise Minsky-Papert symmetrization of p. By this, we mean the polynomial $q(z_1,\dots,z_{R+1})\colon ([N]^*)^R \to \mathbb{R}$ of total degree at most d that, on input $z_1,\dots,z_{R+1} \in ([N]^*)^R$, outputs the average value of p across all inputs in which the ith copy of OR_N is fed an input of Hamming weight i (see Lemma 27).

Let $x = (x_1,\dots,x_R) \in \left(\{-1,1\}^N\right)^R$ be an input to $(\mathsf{AND}_R \circ \mathsf{OR}_N)^{\leq N}$. Let $z_i = |x_i|$ and $z_{R+1} = N - \sum_{i=1}^R z_i$ be the "unused" Hamming weight of x, i.e., the amount by which the Hamming weight of x is below the maximum allowable quantity N. Consider the polynomial $q(z_1,\dots,z_{R+1})$.

Observe that each z_i is a degree-1 polynomial in x, and hence $q(z_1,\dots,z_{R+1})$ is a polynomial in x total degree at most d. Next, observe that $z_{R+1} \geq 0$ and that, by construction, $\sum_{i=1}^{R+1} z_i = N$. Combined with the fact that p ε-approximates $F^{=N}$, this means that $q(z_1,\dots,z_{R+1})$ approximates $\mathsf{AND}_R \circ \mathsf{OR}_N$ at all inputs of Hamming weight *at most* N, as desired.

8.5 Collision and PTP Lower Bound

Let $n = N \log R$ where N is even and R is a power of 2. Recall from Section 2.2 that the Collision problem and PTP take as input (the binary specification of) a list of N numbers from a range of size R. The Collision problem is to distinguish 1-to-1 lists from 2-to-1 lists. Thus, the approximate degree of the Collision problem[26] is the least degree of a polynomial $p\colon \{-1,1\}^n \to \mathbb{R}$ such that:

- $p(x) \in [2/3, 4/3]$ for inputs x that are 1-to-1.
- $p(x) \in [-4/3, -2/3]$ for inputs x that are 2-to-1.
- $|p(x)| \leq 4/3$ otherwise.

[26]Here, Collision and PTP are partial functions, and we are using the notion of approximate degree of partial functions relevant to Section 7.3.2, whereby the approximating polynomial is required to be bounded even at inputs outside of the promise.

It is often helpful to think of Collision and polynomial approximations to it as working directly with the "frequency vector" representation of the input. Recall from Section 8.4 that, given an $x \in \{-1,1\}^n$ interpreted as a list of N numbers $(k_1, \ldots, k_N)$ from range $[R] = \{1, \ldots, R\}$, we define $z(x) = (z_1, \ldots, z_R) \in ([N]^*)^R$ where z_i is the number of times range item $i \in [R]$ appears in the list specified by x. Claims 60 and 61 together show that, up to a factor of $\log R$, the approximate degree of the Collision problem is exactly the least degree of a polynomial $P : ([N]^*)^R \to \mathbb{R}$ such that:

- $P(z) \in [2/3, 4/3]$ when there exist distinct indices $i_1, \ldots, i_N$ such that $z_{i_1} = \cdots = z_{i_N} = 1$, and $z_i = 0$ otherwise.
- $P(z) \in [-4/3, -2/3]$ when there exist distinct $i_1, \ldots, i_{N/2}$ such that $z_{i_1} = \cdots = z_{i_{N/2}} = 2$, and $z_i = 0$ otherwise.
- $|P(z)| \leq 4/3$ whenever $z_1 + \cdots + z_R = N$.

Define a list of numbers $(k_1, \ldots, k_N) \in [R]^N$ to be "far" from all 1-to-1 lists if its Hamming distance from any 1-to-1 list is at least $N/10$, i.e., if at least $N/10$ of the k_i's would have to be changed to yield a 1-to-1 input. Note that any 2-to-1 input is far from any 1-to-1 input.

An equivalent definition of what it means for a list in $[R]^N$ to be "far" from all 1-to-1 lists is that at most $N - N/10$ range elements appear one or more times in the list (this is in fact the definition we gave in Section 2.2). The following fact gives yet another equivalent definition.

Fact 62. A list $(k_1, \ldots, k_N) \in [R]^N$ is far from all 1-to-1 lists if and only if the number of k_i that "collide" with another k_j, i.e., $k_i = k_j$ for some $j \neq i$, is at least $N/10$.

Recall that the Permutation Testing problem, PTP, asks to distinguish 1-to-1 inputs from those that are far from any 1-to-1 input.[27] In the frequency-vector formulation, the approximate degree of PTP is the least degree of a polynomial $P : ([N]^*)^R \to \mathbb{R}$ such that:

- $P(z) \in [2/3, 4/3]$ when z is the frequency vector for a 1-to-1 list.
- $P(z) \in [-4/3, -2/3]$ when z is the frequency vector for a list that is far from any 1-to-1 input.
- $|P(z)| \leq 4/3$ whenever $z_1 + \cdots + z_R = N$.

The study of the approximate degree and quantum query complexity of the Collision problem and PTP were originally motivated by connections to collision-resistant hashing, a central primitive in cryptography. However, as we have already seen (Section 7.2.2), this study has led to unexpected results such as oracle separations for statistical zero-knowledge.

Observe that both the Collision problem and PTP have approximate degree 0 if the range size R is strictly less than the domain size N, because there are no 1-to-1 inputs in this degenerate case. In the non-degenerate case that $R \geq N$, Section 4.4.1 gives a degree upper bound of $O(N^{1/3})$. It turns out that this upper bound is tight. The key to proving this is the following lemma showing how to symmetrize the frequency-vector representation of a polynomial. While our presentation of its proof is somewhat novel, the lemma is due to Aaronson [Aar02].

[27] Typically, the PTP problem is only defined when the "range size" R equals the "domain size" N, so that 1-to-1 lists represent permutations. For convenience, in this section we define PTP for arbitrary positive integers N and R. In this more general setting, the problem really should be called *injectivity testing* rather than permutation testing.

Lemma 63. Let $P(z)\colon ([N]^*)^R \to \mathbb{R}$ be any polynomial of total degree at most d. For any positive integer ℓ that divides N, let μ_ℓ denote the distribution over $z(x)$ where x is a uniformly random ℓ-to-1 input. Then there is a univariate polynomial Q of degree at most d such that, for all positive integers ℓ that divide N, we have $Q(\ell) = \mathbb{E}_{z\sim\mu_\ell}[P(z)]$.

Proof. By linearity of expectation, it suffices to consider a polynomial P consisting of a single monomial $\prod_{i=1}^R z_i^{d_i}$ where $\sum_i d_i = d$ and z_i^0 is interpreted as the constant 1. If a frequency vector $z \in ([N]^*)^R$ represents an ℓ-to-1 input, then it has has N/ℓ entries equal to ℓ, and $R - N/\ell$ entries equal to 0.

For intuition, suppose for a moment (counterfactually) that z instead had all R entries equal to ℓ. Then clearly $\mathbb{E}_{z\sim\mu_\ell}[P(z)] = \prod_{i=1}^R \ell^{d_i} = \ell^d$ is a degree-d polynomial in ℓ as desired. This is analogous to the t-biased symmetrization for symmetric functions (Lemma 21).

The actual calculation of Q is as follows. For a random ℓ-to-1 input x, its frequency vector $z(x)$ is distributed according to the following random process: first, choose a set $\mathcal{R}$ of N/ℓ range items at random. Second, set $z_i = \ell$ for all $i \in \mathcal{R}$ and set $z_i = 0$ for all $i \notin \mathcal{R}$. Consequently, we can express:

$$\mathbb{E}_{z\sim\mu_\ell}[P(z)] = \Pr[\{i\colon d_i > 0\} \subseteq \mathcal{R}] \cdot \ell^d, \tag{46}$$

where the probability is over the random choice of $\mathcal{R}$. Letting $D = |\{i\colon d_i > 0\}|$,

$$\begin{aligned}
\Pr[\{i\colon d_i > 0\} \subseteq \mathcal{R}] &= \binom{R-D}{N/\ell - D} \Big/ \binom{R}{N/\ell} \\
&= \frac{(R-D)!}{(N/\ell - D)!(R - N/\ell)!} \cdot \frac{(N/\ell)!(R-N/\ell)!}{R!} \\
&= \frac{(N/\ell)(N/\ell - 1)\cdot \ldots \cdot (N/\ell - D + 1)}{R \cdot (R-1) \cdot \ldots \cdot (R-D+1)} \\
&= \frac{1}{R \cdot (R-1) \cdot \ldots \cdot (R-D+1)} \cdot \frac{1}{\ell^D} \cdot (N \cdot (N-\ell) \cdot (N - 2\ell) \cdot \ldots \cdot (N - (D-1)\ell))
\end{aligned}$$

Hence, Equation (46) equals

$$\frac{1}{R \cdot (R-1) \cdot \ldots \cdot (R-D+1)} \ell^{d-D} \cdot (N \cdot (N-\ell) \cdot (N - 2\ell) \cdot \ldots \cdot (N - (D-1)\ell)),$$

which is a polynomial in ℓ of degree at most d. □

Completing the $\Omega(N^{1/3})$ lower bound for Collision and PTP when $R \geq N$. If N were somehow divisible by every integer between 1 and $N^{2/3}$, we could prove an $\Omega(N^{1/3})$ approximate degree lower bound for the Collision problem as follows. Let P be an approximating polynomial for the frequency-vector representation of this problem. By Lemma 63, there exists a univariate polynomial Q of degree at most $\deg(P)$ such that $Q(1) \in [2/3, 4/3]$, $Q(2) \in [-2/3, -4/3]$, and $|Q(\ell)| \leq 4/3$ for all integers $\ell \in [N^{2/3}]$. The reasoning used in the Minsky-Papert-symmetrization-based lower bound for OR (Section 5.1) then implies that $\deg(Q) \geq \Omega(N^{1/3})$, and hence $\deg(P) \geq \Omega(N^{1/3})$ as well.

Unfortunately, it is not the case that N is divisible by all integers in $[N^{2/3}]$. Aaronson and Shi [AS04] side-stepped this complication with a yet-more-sophisticated symmetrization calculation, transforming a multivariate approximation to the Collision problem into a certain *trivariate*

polynomial Q of degree at most $\deg(p)$, and such that they could prove $\deg(Q) \geq \Omega(N^{1/3})$. We do not cover their more sophisticated symmetrization in this survey. However, we do give a detailed proof sketch of an $\Omega(N^{1/3})$ lower bound for the closely related PTP problem, for which (as we now show) Lemma 63 is sufficient.

Proof outline. The idea is to first invoke Lemma 63 with the domain size set to some enormous number $\tilde{N}$, so big that it is divisible by all integers in $[N]$ (say, $\tilde{N} = N!$). This allows us to directly prove the desired $\Omega(N^{1/3})$ lower bound on the degree required to distinguish 1-to-1 inputs from $(\tilde{N}/N)$-to-1 inputs when the domain and range size size are huge (namely $\tilde{N}$). We then show how one could take any degree-d approximating polynomial for $\mathsf{PTP}_{N,\tilde{N}}$ and use it to solve the aforementioned problem with the same degree. Note that here, the domain size equals N which is small, but the range size remains $\tilde{N}$ which is very large. In the final step, we reduce the range size by showing how an approximating polynomial for $\mathsf{PTP}_{N,N}$ can be used to approximate $\mathsf{PTP}_{N,\tilde{N}}$. Together, these steps imply that $d \geq \Omega(N^{1/3})$ as claimed.

Details of the first step. Suppose $\tilde{P}(z)\colon \left([\tilde{N}]^*\right)^{\tilde{N}} \to \mathbb{R}$ is a polynomial satisfying:

(a) $\tilde{P}(z) \in [2/3, 4/3]$ when z is the frequency vector of a 1-to-1 input.

(b) $\tilde{P}(z) \in [-4/3, -2/3]$ when z is the frequency vector of an $(\tilde{N}/N)$-to-1 input.

(c) $|\tilde{P}(z)| \leq 4/3$ for all other inputs z that represent an $(\tilde{N}/r)$-to-1 input for some $r \in [N]$.

Conditions (b) and (c) above are well-defined because we have ensured that $\tilde{N}$ is divisible by all integers in $[N]$, i.e., $\tilde{N}/r$ is an integer for all $r \in [N]$. Clearly, $\tilde{N}/r$ divides $\tilde{N}$. Hence, Lemma 63 allows us to conclude that there is a univariate polynomial $\tilde{Q}$ of degree at most $\deg(\tilde{P})$ such that $\tilde{Q}(1) \in [2/3, 4/3]$, $\tilde{Q}(\tilde{N}/N) \in [-4/3, -2/3]$, and $|\tilde{Q}(\tilde{N}/r)| \leq 4/3$ for all $r \in [N]$.

Let $Q(t) = \tilde{Q}(\frac{\tilde{N}}{N} \cdot t)$. Then $\deg(Q) = \deg(\tilde{Q}) \leq \deg(\tilde{P})$. Reformulating the conditions of the previous paragraph, we know that $Q(N/\tilde{N}) \in [2/3, 4/3]$, $Q(1) \in [-4/3, -2/3]$, and $|Q(N/r)| \leq 4/3$ for all integers $r \in [N]$ (note that this last condition applies even to integers $r \in [N]$ that do *not* divide N). We chose $\tilde{N}$ to be so much larger than N that the first condition is essentially equivalent to requiring that $Q(0) \in [2/3, 4/3]$, so from now on let us replace the $N/\tilde{N}$ appearing in that condition with 0.

The question then becomes: what is the least degree of a univariate polynomial Q with these properties? Before answering this question, observe that it has a similar flavor to the question arising in the Minsky-Papert-symmetrization lower bound for OR_N (Section 5.1). There, we needed to lower bound the degree of any univariate polynomial q satisfying $q(0) \in [2/3, 4/3]$, $q(1) \in [-4/3, -2/3]$, and $|q(t)| \leq 4/3$ for all $t \in [N]^*$. Both the polynomial Q considered in this section and the polynomial q from Section 5.1 have a "jump" between input 0 and input 1. The key difference is that q is bounded at all integers between 0 and N while here Q is bounded at inputs of the form N/r where r is an integer.

It turns out that the minimum degree Q satisfying the above conditions is $\Theta(N^{1/3})$ (contrast this with the minimum degree of the polynomial q arising the analysis of OR_N, which was $\Theta(N^{1/2})$). This result was first proved by Zhandry [Zha15]. For brevity, we will only give a *very* sketchy outline of how to prove this, following a dual polynomial construction from [AKKT20]. Specifically, the

idea is to modify the dual polynomial for OR given in Section 6.1. That dual (see Equation (22)) relied on a univariate function q_S that evaluated to zero everywhere except at integers in $S = \{0,1\} \cup \{ci^2 \colon i = 1, 2, \ldots, \lfloor\sqrt{N/c}\rfloor\}$ for a large enough constant c. We saw that the resulting univariate function q_S was uncorrelated with any polynomial of degree less than $|S|$, and was well-correlated with the appropriate univariate function (that which maps 0 to 1 and all integers between 1 and N to -1). This was enough to conclude that there is no univariate polynomial q of degree at most $|S| \approx \sqrt{N/c} = \Theta(\sqrt{N})$ satisfying the three conditions of the previous paragraph.

In order to lower bound the degree of Q, it turns out to be necessary and sufficient to tweak q_S so that the inputs at which it is non-zero are not integers, but rather of the form N/r for positive integer r. The natural way to do this is to take each integer ci^2 in S, and *round it* to the nearest quantity of the form N/r for an integer r.

One can show that, so long as $ci^2 \leq N^{2/3}$, the rounding does not significantly affect the correlation of the dual witness with the relevant target function. Intuitively, this is because every integer less than $N^{2/3}$ is "pretty close" (additive distance at most $O(N^{1/3})$) to some point of the form N/r for an integer r. That is, to lower bound the degree of Q, we set S to be

$$\{0,1\} \cup \left\{ \text{round}(ci^2) \colon 1 \leq i \leq \sqrt{N^{2/3}/c} \right\},$$

where round(j) denotes the closest rational number to j of the form N/r for integer r and $c > 0$ is a sufficiently large constant.

In summary, the above construction yields a univariate function that is uncorrelated with polynomials of degree at most $\sqrt{N^{2/3}/c} = \Theta(N^{1/3})$, is non-zero only at inputs of the form N/r for a positive integer r, and is well-correlated with the relevant target function. This turns out to be a dual witness to the fact that any polynomial of $\Omega(N^{1/3})$ lower bound on the degree of Q. Details of the dual construction and calculation can be found in [AKKT20].

From a lower bound on the degree of $\tilde{P}$ to a lower bound for PTP. Now let P approximate $\mathsf{PTP}_{N,\tilde{N}}$. That is, let $P \colon ([N]^*)^{\tilde{N}} \to \mathbb{R}$ be such that

(d) $P(z) \in [2/3, 4/3]$ for frequency vectors z representing 1-to-1 inputs.

(e) $P(z) \in [-4/3, -2/3]$ when z represents an input that is far from 1-to-1.

(f) $|P(z)| \leq 4/3$ when z represents any other input.

We will show how to use P to construct a polynomial $\tilde{P}$ of degree at most $\deg(P)$ with properties (a)-(c) defined earlier in Step 1. That analysis showed that $\tilde{P}$ requires degree $\Omega(N^{1/3})$, and hence P does as well, proving the claimed lower bound for $\mathsf{PTP}_{N,\tilde{N}}$.

We construct $\tilde{P} : ([\tilde{N}]^*)^{\tilde{N}}$ as follows. Given an input $\tilde{z} = (\tilde{z}_1, \ldots, \tilde{z}_{\tilde{N}})$ to $\tilde{P}$ where $\tilde{z}_1 + \cdots + \tilde{z}_{\tilde{N}} = \tilde{N}$, consider the following random process. Sample a list S of N numbers $(k_1, \ldots, k_N)$, where for each $i \in [\tilde{N}]$, we set $k_j = i$ with probability $\tilde{z}_i/\tilde{N}$. Then define $z = (z_1, \ldots, z_{\tilde{N}})$ where each z_i is the count of the number of samples k_j that are equal to i. For random S, the vector z exactly follows the *multinomial distribution* $\mu(\tilde{z}) := \text{Mult}(N; \tilde{z}_1/\tilde{N}, \ldots, \tilde{z}_{\tilde{N}}/\tilde{N})$: We take N independent samples, each of which lands in one of $\tilde{N}$ categories with probability $\tilde{z}_i/\tilde{N}$. Each entry of the vector z then counts the number of samples in that category.

Now define $\tilde{P}(\tilde{z}) = \mathbb{E}_{z\sim\mu(\tilde{z})}[P(z)]$. If $\tilde{z}$ represents a 1-to-1 input (i.e., all frequencies are either 0 or 1), then so does z with probability at least $1 - N^2/\tilde{N}$. Hence,

$$\tilde{P}(\tilde{z}) \in [2/3 - o(1), 4/3].$$

If $\tilde{z}$ represents a $(\tilde{N}/N)$-to-1 input, then let $\mathcal{R}$ be the N range elements in $[\tilde{N}]$ with non-zero frequency and observe that each element of the sampled list S is a random member of $\mathcal{R}$. A simple probabilistic calculation shows that with overwhelming probability, there are at least $N/10$ elements of S that collide, and hence z represents an input list that is far from 1-to-1—see Fact 62. Accordingly,

$$\tilde{P}(\tilde{z}) \in [-4/3, -2/3 + o(1)].$$

Finally, it is easy to check that $|\tilde{P}(\tilde{z})| \leq 4/3$ as it outputs an average of P's evaluations, all of which are assumed to have magnitude at most $4/3$.

Finally, we now argue that $\deg(\tilde{P}) \leq \deg(P)$. It is possible to do this by a direct calculation as in the proof of Claim 61, but for the sake of variety, let us give a different argument using properties of the multinomial distribution. By linearity of expectation, it suffices to show this when $P(z)$ is a single monomial $\prod_{i=1}^{\tilde{N}} z_i^{d_i}$. When this is the case, we have that $\tilde{P}(\tilde{z}) = \mathbb{E}_{z\sim\mu(\tilde{z})}\left[\prod_{i=1}^{\tilde{N}} z_i^{d_i}\right]$ is a moment of the multinomial distribution $\mu(\tilde{z})$. Using the fact that a draw from the distribution $\mathrm{Mult}(N; p_1, \ldots, p_{\tilde{N}})$ is distributed as the sum of N independent draws from the distribution $\mathrm{Mult}(1; p_1, \ldots, p_{\tilde{N}})$, we have that the moment generating function (MGF) of z is given by

$$M_z(t_1, \ldots, t_{\tilde{N}}) := \mathbb{E}_{z\sim\mu(\tilde{z})}\left[e^{\langle t,z\rangle}\right] = \left(\sum_{i=1}^{\tilde{N}} \frac{\tilde{z}_i}{\tilde{N}} \cdot e^{t_i}\right)^N.$$

Each moment of z is obtained by taking an appropriate derivative of this MGF evaluated at zero:

$$\tilde{P}(\tilde{z}) = \left.\frac{\partial^{d_1+\cdots+d_{\tilde{N}}}}{\partial t_1^{d_1} \ldots \partial t_{\tilde{N}}^{d_{\tilde{N}}}}\right|_{t_1=0,\ldots,t_{\tilde{N}}=0} \left(\sum_{i=1}^{\tilde{N}} \frac{\tilde{z}_i}{\tilde{N}} \cdot e^{t_i}\right)^N.$$

Every time this MGF is differentiated with respect to some variable t_i, it results in an additional factor of $\tilde{z}_i e^{t_i}/\tilde{N}$ being "brought out" of the product. For example, if $d_1 = 2, d_2 = 1$, and $d_i = 0$ for all other i, we would get

$$\frac{\partial^3}{\partial t_1^2 \partial t_2^1}\left(\sum_{i=1}^{\tilde{N}} \frac{\tilde{z}_i}{\tilde{N}} \cdot e^{t_i}\right)^N = \left(\frac{N(N-1)(N-2)}{\tilde{N}^3}\tilde{z}_1^2 e^{2t_1}\tilde{z}_2 e^{t_2} + \frac{N(N-1)}{\tilde{N}^2}\tilde{z}_1 e^{t_1}\tilde{z}_2 e^{t_2}\right) \cdot \left(\sum_{i=1}^{\tilde{N}} \frac{\tilde{z}_i}{\tilde{N}} \cdot e^{t_i}\right)^N$$

Evaluating this at $t_1 = \cdots = t_{\tilde{N}} = 0$ makes the first factor a degree-3 polynomial in $\tilde{z}$. Meanwhile, the fact that $\tilde{z}_1 + \cdots + \tilde{z}_{\tilde{N}} = \tilde{N}$ implies that the second factor always evaluates to 1. So the product is a degree-3 polynomial in $\tilde{z}$ overall. An inductive argument reveals that, in general, $\tilde{P}(\tilde{z})$ is a polynomial of total degree at most $d_1 + \cdots + d_{\tilde{N}} = \deg(P)$.

In summary, we have shown that, given a polynomial P with properties (d)-(f) above, we can obtain a polynomial $\tilde{P}$ of degree at most $\deg(P)$ with properties (a)-(c) above. We previously argued that such a polynomial $\tilde{P}$ requires degree $\Omega(N^{1/3})$ and hence $\deg(P)$ must also be $\Omega(N^{1/3})$. This yields a lower bound for $\mathsf{PTP}_{N,\tilde{N}}$ when the domain size is N and the range size is $\tilde{N}$.

From a large-range lower bound to a small-range lower bound. Finally, we explain why the approximate degree of $\mathsf{PTP}_{N,R}$ is the same for any range size $R \geq N$ [Amb05]. Hence, the lower bound established above for $\mathsf{PTP}_{N,\tilde{N}}$ implies the same lower bound for $\mathsf{PTP}_{N,N}$.

Let $Q\colon ([N]^*)^N \to \mathbb{R}$ approximate $\mathsf{PTP}_{N,N}$ in the frequency-vector representation when the domain size is N and the range size is N. We may assume without loss of generality that Q is invariant under permutations of its input, as if not we may replace Q with its expectation under a random permutation. Since $\mathsf{PTP}_{N,N}$ itself is invariant under permutations of the frequency vector z, the resulting symmetric polynomial also approximates PTP and has the same degree as Q. We will show how to use Q to construct a polynomial $P\colon ([N]^*)^R \to \mathbb{R}$ approximating $\mathsf{PTP}_{N,R}$ for any range size $R \geq N$ (in particular, $R = \tilde{N}$) without increasing its degree.

Since Q is invariant under permutations of its inputs, it is a linear combination of elementary symmetric polynomials, each of the form

$$\sum_{i_1,\ldots,i_k \in [N]} \prod_{j=1}^{k} z_{i_j}^{d_j} \tag{47}$$

for some non-negative integers $d_1, \ldots, d_k$ summing to at most $\deg(Q)$, where z_i^0 is interpreted as the constant 1. Let us obtain a polynomial P by replacing each such elementary symmetric polynomial from Q with:

$$\sum_{i_1,\ldots,i_k \in [R]} \prod_{j=1}^{k} z_{i_j}^{d_j}. \tag{48}$$

To clarify, the sums in Equations (47) and (48) are over *distinct* elements $i_1, \ldots, i_k$ in $[N]$ and $[R]$ respectively.

Clearly P has degree at most $\deg(Q)$. We claim that P approximates $\mathsf{PTP}_{N,R}$ with domain size N and range size R. To see this, observe that if $z \in ([N]^*)^R$ is the frequency vector of some input list, then it has at most N nonzero entries. Suppose these nonzero entries are $z_{i_1}, z_{i_2}, \ldots, z_{i_k} > 0$ for some $k \leq N$. Then define the new frequency vector $\overline{z}$ by setting $\overline{z}_j = z_{i_j}$ for every $j = 1, \ldots, k$ and $\overline{z}_j = 0$ for $j = k+1, \ldots, N$. Intuitively, $\overline{z}$ removes the 0-entries from z until at most N entries remain and shifts the nonzero entries to the "front" of the vector. It is immediate from the definition that $P(z) = Q(\overline{z})$.

This means that P approximates $\mathsf{PTP}_{N,R}$ with domain size N and range size R. This is because if z is the frequency vector for a 1-to-1 input, then $\overline{z}$ is the all-ones vector. On the other hand, if z represents a list that is far from any 1-to-1 input, then $\overline{z}$ is also the frequency vector of a list that is far from any 1-to-1 input.

The proof sketched in this section shows that any polynomial approximating $\mathsf{PTP}_{N,N}$ in the frequency vector or "z-representation" has degree $\Omega(N^{1/3})$. Recall that Claim 61 shows that approximability in the "z-representation" is equivalent to approximability in the "y-representation". Thus, we also have that every polynomial $p\colon \{-1,1\}^{N \cdot N} \to \mathbb{R}$ such that:

- $p(y(x)) \in [2/3, 4/3]$ for inputs x that are 1-to-1,
- $p(y(x)) \in [-4/3, -2/3]$ for inputs x that are far from any 1-to-1 input, and
- $|p(y(x))| \leq 4/3$ otherwise

requires degree $\Omega(N^{1/3})$.

8.6 Element Distinctness Lower Bound

Recall from Section 8.2 that the k-distinctness function k-ED (for constant k) interprets its input as a list of N numbers from a range of size R and outputs 1 if and only if there is some range item that appears at least k times in the list. 2-ED is a particularly natural special case that is referred to as Element Distinctness, or ED for short. In Section 8.2, we sketched a lower bound of $\Omega(N^{3/4-1/(4k)})$ on the approximate degree of k-ED. This lower bound is proved using variants of dual block composition. This bound is close to tight for large constants k, as there is an upper bound known that is (strictly better than) $O(N^{3/4})$ for all constants k. However, it is not tight for $k = 2$. As we now explain, the $\Omega(N^{1/3})$ approximate degree lower bound for PTP in fact implies an $\Omega(N^{2/3})$ approximate degree lower bound for ED with domain size and range size equal to N. This lower bound is optimal (we give a matching upper bound in Section 4.4.2).

Given a polynomial P of degree d for approximating ED with domain size N' and range size $(N'/100)^2$, we show below that one can obtain a polynomial Q of the same degree approximating PTP over domain size $N := (N'/100)^2$ and range size $R = N$. Since the approximate degree of PTP is $\Omega(N^{1/3})$, it follows that the approximate degree of ED with this domain and range size is $\Omega((N')^{2/3})$. The lower bound can be extended to the "small range" case (range size equal to domain size) using the same reasoning as for PTP (Section 8.5).

Here is how to use P to construct Q. $Q(x)$ outputs the expected value of the following random process: select a set S of $N' = 100\sqrt{N}$ list elements from x at random, and feed them into P. That is, Q equals:

$$\frac{1}{\binom{N}{100\sqrt{N}}} \sum_{S \subseteq [N]:\ |S| = 100\sqrt{N}} P(x|_S).$$

Clearly, $\deg(Q) \leq \deg(P)$. We now explain why Q approximates PTP. If x is 1-to-1 then with probability 1 over the choice of S, $x|_S$ is as well, and hence $\mathsf{ED}(x|_S) = -1$. Hence, $Q(x) \in [-4/3, -2/3]$. Meanwhile, if x is far from any 1-to-1 function, then by the birthday paradox, with probability at least $9/10$ over the random choice of S, $x|_S$ will not be 1-to-1, i.e., $\mathsf{ED}(x|_S) = 1$. Hence, $Q(x) \in [1/2, 4/3]$. It follows that Q approximates PTP to error at most $1/2$.

9 Spectral Sensitivity

A breakthrough result of Huang [Hua19] resolved the so-called *sensitivity conjecture* in the analysis of Boolean functions. The sensitivity conjecture itself is not directly relevant to our discussion of approximate degree. However, follow-on work by Aaronson et al. [ABDK+21] builds on Huang's analysis to derive a variety of important consequences for approximate degree. This includes a powerful new technique for proving degree and approximate degree lower bounds. The technique is based on a quantity called *spectral sensitivity*.

In order to define spectral sensitivity, we remind the reader of the definition of the spectral norm. Any symmetric $N \times N$ matrix $M \in \mathbb{R}^{N \times N}$ has N real eigenvalues, say, $\lambda_1, \ldots, \lambda_N$. Let us sort the magnitudes of the eigenvalues such that $|\lambda_1| \geq |\lambda_2| \cdots \geq |\lambda_N|$.

Definition 64. The spectral norm of a symmetric matrix M, denoted $\|M\|$, is its largest absolute eigenvalue

$$\|M\| := |\lambda_1|.$$

The spectral sensitivity of a Boolean function is the spectral norm of a certain sensitivity matrix associated to it.

Definition 65. Let $f : \{-1,1\}^n \to \{-1,1\}$ be a Boolean function. Define the matrix $A^f = [A^f_{x,y}]_{x\in\{-1,1\}^n, y\in\{-1,1\}^n} \in \{0,1\}^{2^n\times 2^n}$ by $A^f_{x,y} = 1$ if x and y differ in exactly one coordinate and $f(x) \neq f(y)$, and $A^f_{x,y} = 0$ otherwise. The *spectral sensitivity* of f is $\lambda(f) = \|A^f\|$, the spectral norm of A^f.

One can think of A^f as the adjacency matrix of the graph on vertex set $V = \{-1,1\}^n$, where two vertices x, y are connected by an edge iff they differ in one index and induce different values of f. Since this graph is bipartite, whenever λ is an eigenvalue of A^f, we have that $-\lambda$ is an eigenvalue as well. Hence $\lambda(f)$ is simply the maximum eigenvalue of A^f. Another basic fact from matrix analysis is that this eigenvalue equals the maximum of $v^T A^f v$ where v ranges over all all vectors in $\mathbb{R}^{2^n}$ of Euclidean norm 1.[28]

The main step in Huang's 2-page proof of the sensitivity conjecture was to show that the degree of a Boolean function is always at most its spectral sensitivity.

Theorem 66 ([Hua19, ABDK+21])**.** For every Boolean function $f : \{-1,1\}^n \to \{-1,1\}$, we have $\deg(f) \leq \lambda(f)^2$.

Aaronson et al. proved a converse relationship, not only between spectral sensitivity and exact degree, but to approximate degree. Specifically, they showed that, up to a constant factor, spectral sensitivity lower bounds approximate degree.

Theorem 67 ([ABDK+21])**.** For every Boolean function $f : \{-1,1\}^n \to \{-1,1\}$, $\lambda(f) \leq O(\widetilde{\deg}(f))$.

Aaronson et al., in fact, gave two proofs of Theorem 67. The first, which we sketch below, argues that if f is approximated by a low-degree polynomial, then A_f is similar to a matrix with an approximate sparsity property that ensures low spectral norm. The second proof bounds $\lambda(f)$ by the spectral norm of a different matrix derived from an approximating polynomial, in turn controlling this using an approximate factorization norm. This second proof is more self-contained and obtains a better constant, but in our opinion is less intuitive to describe.

Proof sketch of Theorem 67. We first sketch the weaker bound $\lambda(f) \leq O(\widetilde{\deg}(f)\log n)$. Then we explain how to use a "tensor power trick" to automatically improve this to the stated bound.

The proof of the weaker bound uses several intermediate $2^n \times 2^n$ matrices with rows and columns each indexed by $\{-1,1\}^n$.

- Let H be the normalized Walsh-Hadamard matrix defined by $H_{x,y} = 2^{-n/2}\prod_{i=1}^n (x_i \wedge y_i)$. This is a symmetric, orthogonal matrix, and hence an involution ($HH = I$).

  A useful interpretation of H is that it implements the Fourier (and inverse-Fourier) transform of a function $g : \{-1,1\}^n \to \mathbb{R}$. Specifically, if we let $v_g \in \mathbb{R}^{2^n}$ be the vector with $(v_g)_x = g(x)$, then $(Hv_g)_y = 2^{n/2}\widehat{g}(S)$ where $S = \{i \in [n] \mid y_i = -1\}$ is the set of indices indicated by y.

- Let W be the diagonal matrix with $W_{x,x} = |x|$, the Hamming weight of x.

[28] The spectral sensitivity of f is a distinct notion from the "sensitivity" $s(f)$ often studied in the analysis of Boolean functions, which is the maximum over all x of the number of inputs y such that y differs from x in exactly one coordinate and $f(x) \neq f(y)$. The relationship $\lambda(f) \leq s(f)$ follows from the fact that the spectral norm of an adjacency matrix is at most the maximum vertex degree of the associated graph, which for A^f is exactly $s(f)$.

- For any function $g : \{-1,1\}^n \to \mathbb{R}$, let $\mathrm{diag}(g)$ be the diagonal matrix with $\mathrm{diag}(g)_{x,x} = g(x)$.
- Let R^g be the symmetric, orthogonal matrix obtained by conjugating $\mathrm{diag}(g)$ by H, i.e., $R^g = H\,\mathrm{diag}(g)H$. Explicitly, $R^g_{x,y} = \widehat{g}(S)$ where $S = \{i \in [n] \mid x_i \oplus y_i = -1\}$. In particular, this means that if $\deg(g) \leq d$, then $R^g_{x,y} = 0$ whenever $|x \oplus y| > d$.

  This transformation is particularly nice (and useful) when applied to the function $w(x) = |x|$. Note that $w(x) = \frac{n}{2} - \frac{1}{2}(x_1 + \cdots + x_n)$ is a degree-1 polynomial, with Fourier coefficients $\widehat{w}(\emptyset) = n/2$, $\widehat{w}(\{i\}) = -1/2$ for all $i \in [n]$, and $\widehat{w}(S) = 0$ otherwise. Thus, $R^w = HWH$ is the matrix where

$$R^w_{x,y} = \begin{cases} n/2 & \text{if } x = y \\ -1/2 & \text{if } x, y \text{ differ in exactly one coordinate} \\ 0 & \text{otherwise.} \end{cases}$$

With these matrices in hand, the key idea is to use H to perform the following change of basis:

$$\lambda(f) = \max_{v:\|v\|=1} v^\top A^f v = \max_{v:\|v\|=1} v^\top H A^f H v = \max_{v:\|v\|=1} v^\top (R^f W R^f - W) v. \tag{49}$$

To see why the identity on the right of Equation 49 is true, note that by using the fact $HH = I$, it is equivalent to the statement that $A^f = \mathrm{diag}(f)HWH\,\mathrm{diag}(f) - HWH = \mathrm{diag}(f)R^w\,\mathrm{diag}(f) - R^w$. If $x = y \in \{-1,1\}^n$, then the (x,y)'th entry of the matrix on the right evaluates to $f(x)\cdot\frac{n}{2}\cdot f(x) - \frac{n}{2} = 0$. If $x, y \in \{-1,1\}^n$ differ in exactly one coordinate, it evaluates to $f(x)\cdot\left(-\frac{1}{2}\right)\cdot f(y) - 1/2$, which is 1 if $f(x) \neq f(y)$ and 0 otherwise. Finally, if x, y differ in more than one coordinate, then the relevant matrix entry is zero. So indeed we see that these matrices are equivalent entrywise.

Now let $p : \{-1,1\}^n \to \mathbb{R}$ be a degree-d polynomial that ε-approximates f for some ε to be chosen later. By taking an affine transformation of p and increasing ε by a factor of at most 2, we can assume that $p : \{-1,1\}^n \to [-1,1]$. This implies that $\|\mathrm{diag}(p)\| \leq 1$ and $\|\mathrm{diag}(f-p)\| \leq \varepsilon$. Since $\|H\| = 1$, it follows that $\|R^p - R^f\| \leq \varepsilon$. And since $\|W\| = n$, we have from (49) that

$$\lambda(f) \leq \max_{v:\|v\|=1} v^\top (R^p W R^p - W) v + 3\varepsilon n.$$

As a consequence of the facts that $\|R^p\| \leq 1$ and that R^p is supported only on entries (x,y) for which $|x \oplus y| \leq d$, one can show ([ABDK+21, Lemma 14]) that $v^\top(R^pWR^p - W)v \leq d$ for all unit vectors v. Therefore, $\lambda(f) \leq d + 3\varepsilon n$. Now taking $\varepsilon = 1/n$ and $d = O(\widetilde{\deg}(f)\log(1/\varepsilon)) = O(\widetilde{\deg}(f)\log n)$ (Theorem 10) reveals that $\lambda(f) \leq O(\widetilde{\deg}(f)\log n)$.

With this weaker bound in place, we now explain how to improve it to the stated bound of $\lambda(f) \leq O(\widetilde{\deg}(f))$. We have seen that there exists a constant C such that $\lambda(f) \leq C\widetilde{\deg}(f)\log n$ for every $f : \{-1,1\}^n \to \{-1,1\}$. Applying this to the k-fold composition of f with itself, denoted by $f^k : \{-1,1\}^{n^k} \to \{-1,1\}$, yields

$$\lambda(f^k) \leq C\widetilde{\deg}(f^k)\log(n^k)$$

for every $k \geq 1$.

Now on the left-hand side, spectral sensitivity obeys a perfect composition theorem: $\lambda(f \circ g) = \lambda(f)\lambda(g)$ for all functions f, g [ABDK$^+$21, Theorem 29], so $\lambda(f^k) = \lambda(f)^k$. Meanwhile on the right-hand side, Sherstov's robust composition theorem (Theorem 11) shows that there is a constant r for which $\widetilde{\deg}(f^k) \leq (r\widetilde{\deg}(f))^k$. Putting these together shows that

$$\lambda(f) \leq r(Ck \log n)^{1/k} \widetilde{\deg}(f).$$

Since this holds for arbitrarily large k, it follows that $\lambda(f) \leq r\widetilde{\deg}(f)$. □

The following paragraphs detail important implications of Theorems 66 and 67.

Symmetric functions. To gain familiarity with spectral sensitivity, we begin by explaining that Theorem 67 implies a particularly simple proof of a tight lower bound on the approximate degree of any symmetric Boolean function. Recall (Sections 5.2 and 7.3.2) that if $f : \{-1,1\}^n \to \{-1,1\}$ is a symmetric function with $f(x) = 1$ for $|x| = t-1$ and $f(x) = -1$ for $|x| = t$ and $t \leq n/2$, then $\widetilde{\deg}(f) = \Omega(\sqrt{nt})$. (Recall also the matching upper bound in Section 4.3.) Define the unit vector $u \in \mathbb{R}^{2^n}$, indexed by strings $x \in \{-1,1\}^n$, by

$$u_x = \begin{cases} \frac{1}{\sqrt{2\binom{n}{t-1}}} & \text{if } |x| = t-1 \\ \frac{1}{\sqrt{2\binom{n}{t}}} & \text{if } |x| = t \\ 0 & \text{otherwise,} \end{cases}$$

Let $x \sim y$ if they differ in exactly one coordinate. Then we have

$$\begin{aligned} \lambda(f) &= \|A^f\| \\ &\geq u^\top A^f u \\ &= \sum_{x,y \in \{-1,1\}^n} A_{x,y} u_x u_y \\ &\geq \sum_{(x,y):|x|=t-1,|y|=t,x\sim y} \frac{1}{\sqrt{\binom{n}{t-1} \cdot \binom{n}{t}}} \\ &\geq \frac{\binom{n}{t-1} \cdot (n-t+1)}{\sqrt{\binom{n}{t-1} \cdot \binom{n}{t}}} \\ &= \sqrt{\frac{t}{n-t+1}} \cdot (n-t+1) \\ &= \sqrt{t(n-t+1)}. \end{aligned}$$

Hence, Theorem 67 implies that $\widetilde{\deg}(f) \geq \lambda(f) \geq \Omega(\sqrt{nt})$.

Readers familiar with the quantum adversary method in quantum query complexity may recognize the similarity between this argument and the (positive-weights) adversary lower bound on the query complexity of symmetric functions. This is not a coincidence. Aaronson et al. [ABDK$^+$21] showed that the spectral sensitivity of a function f exactly matches the best lower bounds that can be proved by "single-bit" adversary methods, wherein the adversary matrix is restricted to be supported on pairs (x, y) for which x and y differ in exactly one index. The classic adversary bound for symmetric functions, in particular, meets this single-bit restriction.

Maximum separation between exact and approximate degree. Theorems 66 and 67 together imply that for every Boolean function f, the exact and approximate degrees are quadratically related. That is, up to constant factors, the following two inequalities hold:

$$\sqrt{\deg(f)} \leq \widetilde{\deg}(f) \leq \deg(f).$$

This is a tight result, as the OR function exhibits a quadratic separation between these two quantities (its exact degree is n, while its approximate degree is $O(\sqrt{n})$ by Lemma 7).

Read-once formulas. A consequence of the above is a tight lower bound on the approximate degree of any read-once Boolean formulas. Let $f : \{-1,1\}^n \to \{-1,1\}$ be a De Morgan formula (over the basis $\mathsf{AND}, \mathsf{OR}, \mathsf{NOT}$) in which every variable appears exactly once.[29] It is not hard to show via induction over the depth of the formula that f has (exact) degree exactly n, i.e., $\deg(f) = n$. Theorems 66 and 67 together then imply that

$$\widetilde{\deg}(f) \geq \Omega(\lambda(f)) \geq \Omega\left(\sqrt{\deg(f)}\right) = \Omega(\sqrt{n}).$$

Meanwhile, a deep result of Reichardt [Rei11] (closing a long line of work) establishes that the quantum query complexity of read-once De Morgan formulas over n inputs is $O(\sqrt{n})$. Since quantum query upper bounds imply approximate degree upper bounds (Theorem 70), these results characterize the approximate degree of read-once formulas.

Theorem 68. If $f : \{-1,1\}^n \to \{-1,1\}$ is any read-once De Morgan formula, then $\widetilde{\deg}(f) = \Theta(\sqrt{n})$.

Bipartite perfect matching. Toward understanding the ability of combinatorial algorithms to quickly solve graph problems, Beniamini and Nisan [BN21, Ben20] studied the exact and approximate degree of the bipartite perfect matching problem, BPM. The Boolean function $\mathrm{BPM}_{n^2}(x_{1,1}, \ldots, x_{n,n})$ evaluates to -1 iff the bipartite graph whose adjacency matrix is encoded by $(x_{1,1}, \ldots, x_{n,n})$ has a perfect matching.

Theorem 69. Bipartite perfect matching satisfies $\widetilde{\deg}(\mathrm{BPM}_{n^2}) = \widetilde{\Theta}(n^{3/2})$.

To obtain the upper bound, Beniamini and Nisan showed that the Boolean dual of BPM, i.e., the function that outputs -1 iff the complement of its input graph does not contain a perfect matching, can be expressed as a low-weight linear combination of $2^{O(n \log n)}$ conjunctions. The approximate degree upper bound then follows from the same technique described for symmetric functions (Section 4.3), Element Distinctness (Section 4.4.2), and Surjectivity (Section 8.1), wherein the target function is represented as a linear combination of more easily analyzed functions.

As for the lower bound, Beniamini [Ben20] determined the spectral sensitivity of bipartite perfect matching to be $\lambda(\mathrm{BPM}_{n^2}) = \Theta(n^{3/2})$. This implies the lower bound of Theorem 69 by Theorem 67.

[29] A De Morgan formula is essentially a Boolean circuit in which each gate is required to have fan-out only one. In more detail, a De Morgan formula over input variables $x_1, \ldots, x_n \in \{-1,1\}^n$ is a binary tree in which each leaf is labeled with a variable x_i or its negation, and each internal node computes either than AND or OR of its two children. The size of a De Morgan formula is the number of leaves of the tree.

Additional discussion. Recall that Aaronson et al. [ABDK+21] showed that $\lambda(f)$ also lower bounds a quantity called the positive weights adversary method, a popular technique for lower bounding quantum query complexity. Accordingly, proving approximate degree lower bounds via spectral sensitivity is unlikely to yield new quantum query lower bounds: any quantum query lower bound that could be established via spectral sensitivity would likely already have been established via the easier-to-apply positive weights adversary technique. This also means that spectral sensitivity is subject to various limitations, such as the so-called certificate complexity barrier, that imply that the method cannot yield tight approximate degree lower bounds for many functions.

Nonetheless, the results described above on read-once formulas and BPM provide examples whereby spectral sensitivity was used to "strengthen" a known quantum query lower bound to an approximate degree lower bound. That is, prior to the results described above, it was already known that the quantum query complexity of read-once De Morgan formulas and BPM were $\Omega(n^{1/2})$ and $\Omega(n^{3/2})$ respectively. The new works used spectral sensitivity to show that these lower bounds hold for approximate degree as well.

10 Approximate Rank Lower Bounds from Approximate Degree

Section 7.2 proved a variety of hardness amplification theorems for approximate degree under block composition. Specifically, it showed that for many pairs of functions f, g, the composed function $F \coloneqq f \circ g$ is harder to approximate by low-degree polynomials than are f and g individually. The main technical tool used to prove these theorems was dual block composition, a powerful technique for combining dual witnesses ψ for f and ϕ for g to obtain a dual witness $\psi \star \phi$ for $f \circ g$.

This section considers a different type of hardness amplification theorem for approximate degree. We view the composed function $F = f \circ g$ as a *matrix* M_F in a natural way, and consider a matrix-analytic analog of ε-approximate degree that is known as ε-approximate rank. Analogously to how ε-approximate degree considers pointwise ε-approximations to real-valued functions via low-degree polynomials, ε-approximate rank studies pointwise ε-approximations to real-valued matrices via low-rank matrices.

Roughly speaking, the key technical theorems (Theorems 77 and 89) in this section show that if f has ε-approximate degree at least d and g is "sufficiently complicated", then M_F has ε-approximate rank at least $2^{\Omega(d)}$. The way we prove these results is largely analogous to the approach in Section 7.2: we take a dual witness ψ to the fact that $\widetilde{\deg}_\varepsilon(f)$ is large, and a dual witness ϕ for the fact that g is "sufficiently complicated", and show that their dual block composition $\psi \star \phi$ is a dual witness to the high ε-approximate rank of the matrix M_F.

Query-to-communication-lifting perspective. As we explain shortly, (the logarithm of) ε-rank is a lower bound on (and in some cases actually *characterizes*) three important communication models, known as $\mathbf{BQP}^{\mathrm{cc}}$, $\mathbf{PP}^{\mathrm{cc}}$, and $\mathbf{UPP}^{\mathrm{cc}}$. Similarly, ε-approximate degree lower bounds or characterizes the *query* analogs of these models. Accordingly, an alternative perspective on the results in this section is that they translate *query* lower bounds for **BQP**, **PP**, and **UPP** into *communication* lower bounds. These communication lower bounds are a principle motivation for the results in this section. With this perspective in mind, we begin this section by introducing the relevant notions in query complexity, followed by communication complexity.

10.1 A Query Complexity Zoo

Recall from Section 4.1 that in query complexity, an algorithm wishes to evaluate a (known) function f at an (unknown) input x while querying as few bits of x as possible. Just as complexity theorists study many different models of computation (deterministic, randomized, quantum, nondeterministic, space-bounded, etc.), query complexity comes in many variants.

In deterministic query complexity, denoted $\mathbf{P}^{\mathrm{dt}}$, a query algorithm is a deterministic procedure that must output $f(x)$ on every input x, and the *query cost* of the algorithm is the maximum over all inputs x of the number of bits queried by the algorithm before outputting $f(x)$.[30] We refer to the algorithm outputting -1 as *accepting* input x and outputting $+1$ as *rejecting* x.

In randomized query complexity, denoted $\mathbf{BPP}^{\mathrm{dt}}$, the query algorithm begins by tossing a sequence r of random coins, after which it executes a deterministic communication protocol (which may depend on r); the algorithm is only required to output the correct answer $f(x)$ with probability at least $2/3$ over the choice of r. The quantum analog of $\mathbf{BPP}^{\mathrm{dt}}$, denoted $\mathbf{BQP}^{\mathrm{dt}}$, was briefly discussed in Sections 4.1 and 8.2.[31]

We are also interested in two powerful variants of randomized query complexity, denoted $\mathbf{PP}^{\mathrm{dt}}$ and $\mathbf{UPP}^{\mathrm{dt}}$, which both capture randomized protocols that do *only slightly better than random guessing*. By random guessing, we mean the algorithm that on any input x, ignores x and outputs a random bit. This has query cost 0 and success probability exactly $1/2$. $\mathbf{PP}^{\mathrm{dt}}$ and $\mathbf{UPP}^{\mathrm{dt}}$ are so-named because they are query analogs of the classical complexity class $\mathbf{PP}$ that captures decision problems solvable by efficient randomized algorithms that output the correct answer with probability strictly greater than $1/2$.[32]

Specifically, the $\mathbf{PP}^{\mathrm{dt}}$ complexity of a Boolean function f is the least d for which there is a randomized algorithm querying at most d bits that, on any input x, outputs $f(x)$ with probability at least $1/2 + 2^{-d}$. The $\mathbf{UPP}^{\mathrm{dt}}$ cost is defined identically, except the advantage over random guessing is only required to be strictly positive, rather than at least 2^{-d}.

Relationship between query complexity and approximate degree. We have already seen (Section 4.1) that approximate degree *lower bounds* quantum query complexity.

Theorem 70 ([BBC+01])**.** For any function $f\colon \{-1,1\}^n \to \{-1,1\}$, $\mathbf{BQP}^{\mathrm{dt}}(f) \geq \Omega(\widetilde{\deg}(f))$.

We now show that $\mathbf{PP}^{\mathrm{dt}}$ and $\mathbf{UPP}^{\mathrm{dt}}$ are *characterized* by ε-approximate degree, for ε very close to 1 (these results appear to be folklore).

Fact 71. Let $f\colon \{-1,1\}^n \to \{-1,1\}$ be any function. Then $\mathbf{UPP}^{\mathrm{dt}}(f) = \deg_{\pm}(f)$.

Proof. As shown by Beals et al. [BBC+01] when they proved Theorem 70, the acceptance probability of any T-query quantum algorithm is computed by a polynomial of degree at most $2T$. An even more basic fact is that for any classical (randomized) algorithm $\mathcal{A}$ making at most T queries to x, there is a degree T polynomial p such that $p(x) = \Pr[\mathcal{A}(x) = -1]$. If $\mathbf{UPP}^{\mathbf{dt}}(f) \leq d$, then

[30] The dt in $\mathbf{P}^{\mathrm{dt}}$ stands for *decision tree*, which is a synonym for a query algorithm.

[31] **BPP** stands for Bounded-Error Probabilistic Polynomial Time while **BQP** stands for Bounded-Error Quantum Polynomial Time. See Footnote 32 below for additional details.

[32] **PP** and **BPP** were both introduced in a paper by Gill [Gil77]. **PP** stands for Probabilistic Polynomial Time, while **BPP** stands for Bounded-Error Probabilistic Polynomial-Time. Despite its longer name, **BPP** proved to be a more-commonly studied class than **PP**, as practical algorithms should output the correct answer with high probability rather than with probability slightly better than 1/2.

there is a d-query algorithm $\mathcal{A}$ such that, for some $\delta > 0$, $f(x) = 1 \implies \Pr[\mathcal{A}(x) = 1] \geq 1/2 + \delta$ and $f(x) = -1 \implies \Pr[\mathcal{A}(x) = 1] \leq 1/2 - \delta$. Hence, there is a degree d polynomial p such that: $f(x) = 1 \implies p(x) - 1/2 \in [\delta, 1]$, and $f(x) = -1 \implies p(x) - 1/2 \in [-1, -\delta]$. Hence, $p - 1/2$ is the desired approximation to f.

For the converse, suppose there is a degree-d polynomial $p(x) = \sum_{|S| \leq d} c_S \chi_S(x)$ satisfying $|f(x) - p(x)| \leq 1 - \delta$ for some $\delta > 0$, where recall that χ_S is the parity function over the variables in S. Consider the query algorithm that randomly selects a parity S with probability proportional to $|c_S|$ and accepts if $\operatorname{sgn}(c_S)\chi_S(x) < 0$. The acceptance probability of this algorithm is $\frac{1}{2}(1 - \frac{p(x)}{\sum_{|S| \leq d} |c_S|})$. Since $p(x) \cdot f(x) > 0$ for all $x \in \{-1, 1\}^n$, this acceptance probability is greater than $1/2$ if $f(x) = -1$ and less than $1/2$ if $f(x) = 1$. □

Fact 72. Let $f \colon \{-1, 1\}^n \to \{-1, 1\}$ be any function. If $\mathbf{PP}^{\mathbf{dt}}(f) \leq d$ then $\widetilde{\deg}_\varepsilon(f) \leq d$ for some $\varepsilon = 1 - 2^{-d}$. Conversely, if $\widetilde{\deg}_{1-2^{-d}}(f) \leq d$, then $\mathbf{PP}^{\mathbf{dt}}(f) \leq O(d \log n)$.

Proof. The first claim follows by the same reasoning as in Fact 71, using that if $\mathcal{A}$ is a $\mathbf{PP}^{\mathrm{dt}}$ algorithm rather than a $\mathbf{UPP}^{\mathrm{dt}}$ algorithm, then δ is not merely positive, but is in fact at least 2^{-d}.

For the converse, suppose there is a degree-d polynomial $p(x) = \sum_{|S| \leq d} c_S \chi_S(x)$ satisfying $|f(x) - p(x)| \leq 1 - 2^{-d}$. Since $|p(x)| \leq 2$ for all $x \in \{-1, 1\}^n$, Fact 1 implies that $\sum_{|S| \leq d} |c_S| \leq 2\binom{n}{d}$. As in Fact 71, consider the query algorithm that randomly selects a parity S with probability proportional to $|c_S|$ and accepts if $\operatorname{sgn}(c_S)\chi_S(x) < 0$. The acceptance probability of this algorithm is $\frac{1}{2}(1 - \frac{p(x)}{\sum_{|S| \leq d} |c_S|})$. Since $p(x) \cdot f(x) \geq 2^{-d}$ and $\sum_{|S| \leq d} |c_S| \leq 2\binom{n}{\leq d} \leq n^{O(d)}$, this acceptance probability is greater than $1/2 + n^{-O(d)}$ if $f(x) = -1$ and less than $1/2 - n^{-O(d)}$ if $f(x) = 1$. Hence, $\mathbf{PP}^{\mathrm{dt}}(f) \leq O(d \log n)$. □

10.2 Communication Complexity

In communication complexity, there are two parties, Alice and Bob, who wish to work together to compute a function of their inputs. Specifically, Alice has input x, Bob has input y, and their goal is to compute some function $f(x, y)$ of their inputs, while exchanging as few bits as possible. Here, both Alice and Bob know the function f, but Bob does not know x and Alice does not know y. The *cost* of a communication protocol is the maximum number of bits exchanged over any pair of inputs, and the communication complexity of f is the least cost of any communication protocol computing f.

As with query complexity, there are many variant models of interest. In the most basic setting, deterministic communication complexity (denoted $\mathbf{P}^{\mathrm{cc}}$), Alice and Bob must always output $f(x, y)$. In ε-error randomized communication complexity, Alice and Bob begin the protocol by tossing a sequence r of random coins, after which they execute a deterministic communication protocol (which may depend on r), and they are only required to output the correct answer $f(x, y)$ with probability at least $1 - \varepsilon$ over the choice of r. If $\varepsilon = 1/3$, the communication model is denoted $\mathbf{BPP}^{\mathrm{cc}}$.

In this survey, we will only consider *private random coins*, meaning that Alice and Bob both have access to their own random coins, but Bob can't see Alice's coins unless she sends them to him (which counts toward the communication cost) or vice versa.

Randomized vs. deterministic communication, and the Equality function. Randomized communication protocols can be far more efficient than deterministic ones for some functions. The prototypical example is the Equality function EQ, which takes as input two n-bit strings x and y, and evaluates to -1 if and only if $x = y$. It is known that the deterministic communication complexity of this problem is n (or $n+1$ if Alice and Bob are both required to know the output). But its ε-error randomized communication complexity is just $O(\log(1/\varepsilon) + \log n)$, and in particular its $\mathbf{BPP}^{\mathrm{cc}}$ complexity is just *logarithmic* in n.

The idea is that the random string can be used to select a hash function h, and then Alice can send $h(x)$ to Bob, who will output -1 if and only if $h(x) = h(y)$. Suppose h is a random function with domain $\{-1,1\}^n$ and range $\{1, 2, \ldots, 1/\varepsilon\}$. If $x = y$, then the probability (over the random choice of h) of this event is 1. On the other hand, if $x \neq y$, then the probability $h(x) = h(y)$ is ε.[33] In the private coin setting, the communication cost is what is required for Alice to send both $h(x)$ and a description of the hash function h to Bob, as this lets Bob compute $h(y)$ and compare it to $h(x)$. Clearly, specifying $h(x)$ requires just $\lceil \log_2(1/\varepsilon) \rceil$ bits. Unfortunately, h does not have a short description if it is a random hash function. However, it turns out that h need not be a uniform random function; there are ways to choose h such that the above protocol works and Alice only needs to send $O(\log(1/\varepsilon) + \log n)$ bits to Bob to specify h.

Even more powerful communication models. The above example demonstrates that randomized communication protocols can be quite powerful, at least relative to deterministic ones. This means that we might expect proving *lower bounds* against randomized communication protocols to be difficult. Yet in this section, we are interested in proving lower bounds against *even more powerful* communication models. We are particularly interested in the following settings. First is quantum communication complexity (denoted $\mathbf{BQP}^{\mathrm{cc}}$). Roughly speaking, in a $\mathbf{BQP}^{\mathrm{cc}}$ protocol, Alice and Bob are allowed to exchange *quantum bits* rather than just classical bits, and must output $f(x, y)$ with probability at least $2/3$. As with $\mathbf{BQP}^{\mathrm{dt}}$, the precise details of the $\mathbf{BQP}^{\mathrm{cc}}$ model will not be relevant to this survey.

Second are the communication analogs of $\mathbf{PP}^{\mathrm{dt}}$ and $\mathbf{UPP}^{\mathrm{dt}}$. $\mathbf{PP}^{\mathrm{cc}}(f)$ is the least d for which a randomized communication protocol in which Alice and Bob exchange at most d bits, and output $f(x)$ with probability at least $1/2 + 2^{-d}$. $\mathbf{UPP}^{\mathrm{cc}}(f)$ is defined identically, except the advantage over random guessing is only required to be strictly positive.

10.3 Lifting Theorems: Communication Lower Bounds from Query Lower Bounds

A powerful approach to constructing hard communication problems is called "query-to-communication lifting". In this approach, one starts with a function $f\colon \{-1,1\}^n \to \{-1,1\}$ that is hard in the corresponding query model, and transforms f into a communication problem by *composing* it with a simple communication problem $g\colon \{-1,1\}^m \times \{-1,1\}^m \to \{-1,1\}$ (g is typically called a "gadget" function). Specifically, in the communication problem $F = f \circ g$, Alice and Bob are respectively given inputs $x = (x_1, \ldots, x_n) \in (\{-1,1\}^m)^n$ and $y = (y_1, \ldots, y_n) \in (\{-1,1\}^m)^n$, and their goal is to compute $F(x, y) = f(g(x_1, y_1), \ldots, g(x_n, y_n))$. That is, Alice's and Bob's inputs define n independent copies of the communication problem g, and the outputs of these copies of g are fed into f to determine the function value $F(x, y)$.

[33] To see this, observe that since h is a random function with range $R = \{1, 2, \ldots, 1/\varepsilon\}$, then $h(x)$ and $h(y)$ are random range elements. Hence, $h(x) = h(y)$ with probability $1/|R| = 1/(1/\epsilon) = \epsilon$.

The communication complexity of the lifted function F is always *at most* the query complexity of f times the deterministic communication complexity of g [BCW98]. This is because Alice and Bob can "work together" to run a query algorithm $\mathcal{A}$ for f, and each time $\mathcal{A}$ queries an input i to f, Alice and Bob can "answer" the query with $g(x_i, y_i)$. The communication cost of this protocol is the number of queries made by $\mathcal{A}$, times the deterministic communication cost of g [34]

Lifting theorems show that for many communication models, if g is a "sufficiently complicated" gadget, then this is essentially the *best* that Alice and Bob can do. That is, the communication complexity of F is *at least* the query complexity of f. The term lifting refers to the fact that the theorem takes a weak lower bound (i.e., one that applies only in a query complexity setting) and lifts it up, or strengthens it, into a stronger lower bound (i.e., one that applies to a richer model, namely the corresponding communication model).

In this chapter, we prove lifting theorems for $\mathbf{PP}^{\mathrm{cc}}$ and $\mathbf{UPP}^{\mathrm{cc}}$. We will also show that an approximate degree lower bound for f "lifts" to a quantum communication lower bound for F. All of these results go through a matrix-analytic notion called *approximate rank*.

Making sure the gadget is not "too simple". If the gadget function g is too simple, then there will often be much better communication protocols for $f \circ g$ than the one described above in which Alice and Bob simulate a query protocol for f. That is, lifting theorems will *fail to hold* for $f \circ g$.

For example, if $f = \oplus_n$ is the parity function on n bits, and $g = \oplus_2$ is the parity function on two bits, then $f \circ g = \oplus_n \circ \oplus_2 = \oplus_{2n}$ is simply the parity function on $2n$ bits. There is a deterministic communication protocol for this function of constant cost: Alice sends Bob a single bit indicating whether the parity of x is even or odd. If the parity of x is even, then Bob outputs the parity of his input y; otherwise, Bob outputs the negation of the parity of y.

In contrast, a consequence of the fact that $\deg_{\pm}(\oplus_n) = n$ is that any query protocol for $f = \oplus_n$ requires $\Omega(n)$ queries (this holds even for $\mathbf{UPP}$ query protocols, see Fact 71).

Similar examples hold for other two-bit gadgets. For example, if $f = \mathsf{AND}_n$ and $g = \mathsf{AND}_2$, then $f \circ g = \mathsf{AND}_{2n}$. It is easily seen that there is a constant-cost deterministic communication protocol for AND_{2n}. In contrast, since $\widetilde{\deg}(f) = \Omega(\sqrt{n})$ (Theorem 23), there is no bounded-error query protocol protocol for f (not even a quantum one) of cost less than $\Omega(\sqrt{n})$ (see Theorem 70).

Because of these examples, the gadget g appearing in the lifting theorems we establish in this chapter are necessarily defined over inputs consisting of more than 2 bits, but the input size to g is still constant (specifically, our g takes *six* bits as input).

10.4 Communication Lower Bounds via Approximate Rank

Given a matrix $M \in \{-1, 1\}^{m \times n}$, the ε-approximate rank of M, denoted $\widetilde{\mathrm{rank}}_{\varepsilon}(M)$ is the smallest r such that there exists a rank-r matrix R satisfying $|M_{i,j} - R_{i,j}| \leq \varepsilon$ for all $(i, j) \in [m] \times [n]$. Conceptually, ε-approximate rank can be thought of as a matrix-analytic analog of approximate degree: It is asking for the "lowest-complexity" matrix (as measured by rank) that approximates M entry-wise up to error ε, just as the ε-approximate degree of f asks for the "lowest-complexity polynomials" (as measured by degree) that approximates f point-wise up to error ε.[35] As with

[34] If the query algorithm for f is randomized, and the communication model is private coin, then Alice will also have to choose the random coins to use within the query algorithm, and send them to Bob.

[35] A more apt analogy is that ε-approximate rank is a matrix analog of ε-approximate *sparsity*, meaning the least number of monomials in an ε-approximation for the target function. This is because a rank-1 matrix can be thought

approximate degree, if ε is not specified, its value is interpreted as $1/3$ by convention.

Similarly, the *sign-rank* of M, denoted $\mathrm{rank}_{\pm}(M)$ is the least rank r of a matrix R that agrees in sign with M entry-wise. Just as the threshold degree of f equals the limit of its ε-approximate degree as ε approaches 1 from below, so too does the sign-rank of M equal the limit of its approximate rank as ε approaches 1 from below.

Communication problems as matrices. Any communication problem $f(x, y)$ can be viewed as a matrix. Specifically, if $x \in \{-1, 1\}^n$ and $y \in \{-1, 1\}^m$, consider the $2^n \times 2^m$ matrix M_f whose (x, y)'th entry is $f(x, y)$ (here, we are using x to index the rows of M and y to index the columns).

Communication lower bounds via matrix analysis. It turns out that the complexity of $f(x, y)$ in various communication models is closely related to the matrix-analytic notion of approximate rank. To illustrate why variants of matrix rank and communication complexity are often related, the following fact spells out the details of such a relationship between *deterministic* communication complexity and *exact* matrix rank.

Fact 73. The deterministic communication complexity of $f\colon \{-1, 1\}^n \times \{-1, 1\}^n \to \{-1, 1\}$ is lower bounded by the logarithm of the rank (over the real numbers) of the communication matrix M_f.

Proof. In communication complexity, the term *rectangle* refers to any subset of the inputs with a product structure, i.e. any set of the form $A \times B$ for $A, B \subseteq \{-1, 1\}^n$.

A basic fact regarding deterministic communication protocols is the so-called *rectangle property*. This states that any protocol with communication cost c for a function f partitions the input set $\{-1, 1\}^n \times \{-1, 1\}^n$ into at most 2^c rectangles $R_1, \ldots, R_{2^c}$ such that f is constant on each rectangle in the partition. That is, if (x, y) and (x', y') lie in the same rectangle R_i then $f(x, y) = f(x', y')$. Each rectangle R_i is referred to as *monochromatic* for f.

To see why the rectangle property holds, let us refer to a record of all bits exchanged by Alice and Bob during the execution of the communication protocol as a *transcript*. Say that an input (x, y) to f is *consistent* with a transcript if Alice and Bob produce that transcript when they run the communication protocol on (x, y). It is easily shown by induction on the number of rounds of the communication protocol that if (x, y) and (x', y') are both consistent with a given transcript, then so is (x, y') and (x', y). That is, the set of inputs consistent with any particular transcript form a rectangle. Intuitively, this means that if Alice has input x or x', then the protocol does not permit her to tell the difference between Bob holding input y vs. holding input y'. And similarly Bob cannot tell the difference between Alice holding x vs. holding x'. This is compatible with the protocol computing the function f only if $f(x, y) = f(x, y') = f(x', y') = f(x', y')$.

The rectangle property then follows from the observation that there are at most 2^c distinct transcripts that can be generated by a communication protocol of cost c.

With the rectangle property in hand, Fact 73 is derived via the following analysis. For a rectangle $R = A \times B$, define the matrix M_R via:

$$(M_R)_{x,y} = \begin{cases} 1 \text{ if } (x, y) \in A \times B \\ 0 \text{ otherwise.} \end{cases}$$

of as a matrix-analog of a monomial: just as an s-sparse polynomial can be written as a linear combination of s monomials, a rank-r matrix can be written as a linear combination of r rank-1 matrices. This analogy is described in detail in Section 10.4.2.

It is easy to see that M_R has rank 1, because its row space is spanned by the indicator vector of B (and similarly its column space is spanned by the indicator vector of A). That is, every row x of M_R is either all-0s (if x is not in A) or equal to the indicator vector of B (if x is in A). Similarly, every column of M_R is either the all-0s vector, or equal to the indicator vector of A.

If f has deterministic communication complexity at most c, let $R_1, \ldots, R_{2^c}$ be the partition of $\{-1,1\}^n \times \{-1,1\}^n$ into at most 2^c monochromatic rectangles as guaranteed by the rectangle property above. Then $M_f = \sum_{i=1}^{2^c} M_{R_i}$ expresses M_f as a sum of at most 2^c matrices of rank one. By sub-additivity of matrix rank, this implies that the rank of M_f is at most 2^c. □

Fact 73 is a communication analog of the fact (see Section 10.1) that Fourier degree lower bounds deterministic query complexity. In turn, the following facts are communication analogs of Theorem 70, Fact 71, and Fact 72. Specifically, just as these earlier results show that ε-approximate *degree* lower bounds quantum, **UPP**, and **PP** query complexity, the following results show that (the logarithm of) ε-approximate *rank* lower bounds quantum, **UPP**, and **PP** *communication* complexity. The more powerful the communication model, the closer the relevant value of ε is to 1.

Fact 74 ([Kre95, Yao93, BdW01, LS09a])**.** For any Boolean function $f(x,y)$,

$$\mathbf{BQP}^{\mathrm{cc}}(f) \geq \Omega(\log(\widetilde{\mathrm{rank}}(M_f))).$$

We do not prove Fact 74, as this would require us to formally define $\mathbf{BQP}^{\mathrm{cc}}$, which we prefer to avoid in this survey.[36] As indicated above, it can be interpreted as the "communication analog" of the query-complexity result that $\mathbf{BQP}^{\mathrm{dt}}(f) \geq \Omega(\widetilde{\deg}(f))$ (Theorem 70).

The next two facts show that $\mathbf{UPP}^{\mathrm{cc}}$ and $\mathbf{PP}^{\mathrm{cc}}$ are *characterized* by $\log(\widetilde{\mathrm{rank}}_\varepsilon(M_f))$ for appropriate values of ε. These results are communication analogs of Facts 71 and 72.

Fact 75 (Paturi and Simon [PS86])**.** For any Boolean function $f(x,y)\colon \{-1,1\}^n \times \{-1,1\}^n \to \{-1,1\}$, $\mathbf{UPP}^{\mathrm{cc}}(f) = \log(\mathrm{rank}_\pm(M_f)) \pm \Theta(1)$.

Fact 76. For any Boolean function $f(x,y)$,

1. If $\mathbf{PP}^{\mathrm{cc}}(f) \leq d$, then $\log(\widetilde{\mathrm{rank}}_{1-2^{-d}}(M_f)) \leq O(d + \log n)$.
2. If $\log(\widetilde{\mathrm{rank}}_{1-2^{-d}}(M_f)) \leq O(d)$, then $\mathbf{PP}^{\mathrm{cc}}(f) \leq O(d + \log n)$.

Proof. Lee and Shraibman [LS09a, Theorem 1] showed the following relationship between approximate rank and another matrix analytic quantity, an "approximate factorization norm" denoted γ_2^α for a parameter $\alpha \in [1,\infty]$:

$$\delta^2 \cdot \gamma_2^{1/\delta}(M)^2 \leq \widetilde{\mathrm{rank}}_{1-\delta}(M) \leq O\left(\left(\frac{1}{1-\delta}\right)^6 \cdot n^3 \cdot \gamma_2^{1/\delta}(M)^6\right). \tag{50}$$

When M is the communication matrix of a function f, the quantity $\gamma_2^\alpha(M)$ provides a lower bound on its randomized communication complexity. Specifically [LS09d, Theorem 10] shows that if there

[36] Fact 74 holds even if $\mathbf{BQP}^{\mathrm{cc}}$ is defined to allow prior entanglement, i.e., Alice and Bob may share an unlimited number of entangled qubits prior to learning their inputs x and y.

is a randomized communication protocol for f with error probability $\varepsilon = \frac{1-\delta}{2}$ and using c bits of communication, then $c \geq 2\log \gamma_2^{1/\delta}(M_f) - 2\log(1/\delta)$. Thus, if $\mathbf{PP}^{\mathrm{cc}}(f) \leq d$, we have $\log \gamma_2^{2^{-d}}(M_f) \leq O(d)$, which by the right-hand inequality of (50) implies that $\log(\widetilde{\mathrm{rank}}_{1-2^{-d}}(M_f)) \leq O(d + \log n)$.

Conversely, the quantity γ_2^{α} is related to a combinatorial measure of a sign matrix called its *discrepancy*, via $\gamma_2^{\alpha}(M) \geq \gamma_2^{\infty}(M) \geq \frac{1}{8\,\mathrm{disc}(M)}$ [LS09c, Theorem 3.1]. The discrepancy of a communication matrix is known to tightly capture its $\mathbf{PP}^{\mathrm{cc}}$ communication complexity [Kla03]: $\log(1/\mathrm{disc}(M_f)) \leq \mathbf{PP}^{\mathrm{cc}}(f) \leq O(\log(n/\mathrm{disc}(M_f)))$. Thus, if $\log(\widetilde{\mathrm{rank}}_{1-2^{-d}}(M_f)) \leq O(d)$, then the left-hand inequality of (50) implies $\gamma_2^{2^{-d}}(M_f) \leq 2^{O(d)}$, which implies that $\mathrm{disc}(M_f) \geq 2^{-O(d)}$, and hence $\mathbf{PP}^{\mathrm{cc}}(f) \leq O(d + \log n)$. □

With Facts 74-76 in hand, we turn to the following task. Given a function $f\colon \{-1,1\}^n \to \{-1,1\}$ of ε-approximate degree at least $\widetilde{d}$, identify a communication problem $F \coloneqq f \circ g$ for a function $g(x,y)$ on $O(1)$ bits such that $\widetilde{\mathrm{rank}}_{\varepsilon}(M_F)$ is at least 2^d. In this manner, we "lift" approximate degree lower bounds for f (which on their own suffice to establish query lower bounds) to $\mathbf{BQP}^{\mathrm{cc}}$, $\mathbf{UPP}^{\mathrm{cc}}$, and $\mathbf{PP}^{\mathrm{cc}}$ communication lower bounds for F.

Approximate rank lower bounds for composed functions. In Section 7.2, we used dual block composition to establish hardness-amplification results via block-composition. That is, we showed that $f \circ g$ is harder to approximate by low-degree polynomials than f or g alone. In this section we will use dual block composition to show that $f \circ g$ is harder in a different sense: whereas f is hard for query algorithms and ε-approximation via low-degree polynomials, $f \circ g$ will be hard for communication protocols and ε-approximation via low-rank matrices.

Theorem 77. Let $f\colon \{-1,1\}^n$ to $\{-1,1\}$ satisfy $\widetilde{\deg}_{\varepsilon}(f) \geq d$. There is a function $g\colon \{-1,1\}^6 \to \{-1,1\}$ such that the following holds. For every $\delta \in (0, \varepsilon - 2^{-d/2}]$, $F \coloneqq f \circ g$ satisfies $\widetilde{\mathrm{rank}}_{\delta}(M_F) \geq (2^{d/2} + \varepsilon - 2^{-d/2} - \delta)^2/(1+\delta)^2$. In particular, setting $\delta = \varepsilon - 2^{-d/2}$ yields: $\widetilde{\mathrm{rank}}_{\varepsilon - 2^{-d/2}}(M_F) \geq 2^{\Omega(d)}$.

We will prove this result in two steps.

- First, we will show that composing f with a certain 3-bit gadget $\mathsf{idx}\colon \{-1,1\}^3 \to \{-1,1\}$, called the *indexing* gadget, yields a function $f \circ \mathsf{idx}$ with large ε-*approximate weight*, specifically weight at least 2^d. By approximate weight, we mean the minimum-weight polynomial that ε-approximates f, where the weight of a polynomial is the sum of its coefficients over the parity basis (see Section 2.1).
- Second, we will show that composing $f \circ \mathsf{idx}$ with the two-bit XOR gadget, $\oplus_2$, yields a function F with the claimed approximate rank lower bound.

Hence, the 6-bit gadget g in Theorem 77 is simply the block-composition 3-bit indexing gadget and the two-bit XOR gadget. Theorem 77 is highly similar to [She11a, Theorem 8.1] in Sherstov's paper introducing the so-called *pattern matrix method*, and indeed many of the calculations and notions that we use to prove Theorem 77 closely follow Sherstov's analysis. Still, we believe that our two-step presentation to proving Theorem 77 is subtantially simpler and more intuitive than prior treatments in the literature [She11a, SZ09, Lok09, LS09b].[37]

[37] A minor downside of our two-step approach is that we obtain a 6-bit gadget g, which is slightly larger than the 4-bit gadget in Sherstov's pattern-matrix method [She11a]. Still, any constant-sized gadget suffices for all applications we cover in this survey. Other works have attempted to characterize the class of gadgets g for which Theorem 77 holds [LZ10], which is not a focus in this survey.

Both Step 1 and Step 2 above can be understood as proving optimality of simple and natural techniques for ε-approximating $h := f \circ \mathsf{idx}$ by a low-weight or sparse polynomial, and for ε-approximating $F = h \circ \oplus_2$ by a low-rank matrix. For both steps, we present details of the optimal approximation technique before delving into the proof of optimality.

10.4.1 Step 1: From high approximate degree to high approximate weight

In this section, define $\mathsf{idx}\colon \{-1,1\}^2 \times \{-1,1\} \to \{-1,1\}$ via $\mathsf{idx}(x,y,z) = (z \wedge x) \vee (\overline{z} \wedge y)$. That is, if $z = -1$, then idx outputs x, while if $z = +1$, then idx outputs y. idx is referred to as the *3-bit indexing gadget*.

Recall (Section 2.1) that for any real-valued function $p\colon \{-1,1\}^n \to \mathbb{R}$, $\mathrm{weight}(p)$ is the ℓ_1-norm of the Fourier coefficients of p.

Lemma 78. Let $f\colon \{-1,1\}^n \to \{-1,1\}$ be a Boolean function and suppose that f satisfies $\widetilde{\deg}_\varepsilon(f) \geq d$. Then $\widetilde{\mathrm{weight}}_{\varepsilon - 2^{-d/2}}(f \circ \mathsf{idx}) \geq 2^{d/2}$.

The optimal approximation technique. Lemma 78 shows that the following approach to constructing a low-weight ε-approximation to $f \circ \mathsf{idx}$ is essentially optimal: take the lowest *degree* ε-approximation p to f, and let q be the multilinear polynomial that exactly computes idx, and consider the polynomial obtained by composing p and q. Clearly, $p \circ q$ ε-approximates $f \circ \mathsf{idx}$, and since idx is a Boolean function on only 3 bits, $\mathrm{weight}(q) \leq 3$. Thus, $\mathrm{weight}(p \circ q) \leq \mathrm{weight}(p) \cdot 3^{\deg(p)}$. If $\mathrm{weight}(p) \leq 2^{O(d)}$, Lemma 78 proves a matching lower bound up to a constant factor in the exponent, and with an additive $2^{-d/2}$ loss in the error for which the lower bound holds.

Before proving Lemma 78, we introduce a dual formulation of approximate weight. Recall (Section 6) that a dual polynomial $\psi\colon \{-1,1\}^n \to \mathbb{R}$ showing that $\widetilde{\deg}_\varepsilon(f) \geq d$ must satisfy: (1) ε-correlated with f, i.e., $\langle f, \psi \rangle > \varepsilon \cdot \|\psi\|_1$ and (2) uncorrelated with polynomials of degree at most d. The dual formulation below shows that, in order to witness the lower bound $\widetilde{\mathrm{weight}}_{\varepsilon - 2^{-d/2}}(f) \geq 2^{d/2}$, it is enough for the dual witness to satisfy (1), while (2) becomes: the correlation of ψ with *all* parities (regardless of degree) is at most 2^{-d}.

Dual formulation of approximate weight. Fix a function $f\colon \{-1,1\}^n \to \{-1,1\}$ of interest and a weight bound W. What is the smallest error to which any polynomial (regardless of its degree) of weight at most W over the parity basis can approximate f? The answer to this question is the value of the following linear program [BT15b, Theorem 17]. It has $2^n + 1$ variables, one for each coefficient of p over the parity basis and one for the error parameter ε, and $2 \cdot 2^n$ linear constraints that force p to approximate f to error at most ε at each input $x \in \{-1,1\}^n$.

$$\begin{aligned} \min_{p,\varepsilon} \quad & \varepsilon \\ \text{s.t.} \quad & |p(x) - f(x)| \leq \varepsilon && \text{for all } x \in \{-1,1\}^n \\ & \mathrm{weight}(p) < W \end{aligned}$$

Taking the dual yields the following.

$$\begin{aligned}
\max_{\psi\colon \{-1,1\}^n\to\mathbb{R},\gamma\in\mathbb{R}} \quad & -W\cdot\gamma + \sum_{x\in\{-1,1\}^n} \psi(x)f(x) \\
\text{s.t.} \quad & \sum_{x\in\{-1,1\}^n} |\psi(x)| = 1 \\
& \left|\sum_{x\in\{-1,1\}^n} \psi(x)\chi_S(x)\right| \le \gamma \qquad \text{for all parity functions } \chi_S \\
& \gamma \ge 0
\end{aligned}$$

Suppose $\langle \psi, f\rangle = \varepsilon$ and let $M = \max_{S\subseteq[n]} \left|\sum_{x\in\{-1,1\}^n} \psi(x)\chi_S(x)\right|$. Then by setting $\gamma = M$, we obtain a feasible solution to the above dual program with objective value $\varepsilon - W\cdot M$, thereby witnessing that $\widetilde{\deg}_{\varepsilon - W\cdot M}(f) \ge W$. Setting $W = 2^{d/2}$ in the linear program above, weak LP duality implies the following. In order to prove that $\widetilde{\mathrm{weight}}_{\varepsilon - 2^{-d/2}}(f) \ge 2^{d/2}$, it suffices to identify a function $\psi\colon \{-1,1\}^n \to \mathbb{R}$ satisfying the following three conditions.

$$\sum_{x\in\{-1,1\}^n} |\psi(x)| = 1 \tag{51}$$

$$\sum_{x\in\{-1,1\}^n} \psi(x)f(x) > \varepsilon, \tag{52}$$

$$\left|\sum_{x\in\{-1,1\}^n} \psi(x)\chi_S(x)\right| \le 2^{-d/2}/W = 2^{-d} \text{ for all parities } \chi_S. \tag{53}$$

With this dual formulation in hand, we now prove Lemma 78.

Proof of Lemma 78. Let ψ be a dual witness for $\widetilde{\deg}_\varepsilon(f) \ge d$. As we typically do for composed functions, we can build a dual witness ν for $f\circ \mathsf{idx}$ as the dual block composition (Definition 38 in Section 7) of ψ with a suitable dual witness ϕ for idx. Specifically, let $\phi\colon \{-1,1\}^3 \to \mathbb{R}$ be the natural dual witness for the fact that $\deg_\pm(\mathsf{idx}) \ge 1$. Note that ϕ is just h itself, suitably scaled to have ℓ_1-norm 1, i.e.,

$$\phi(x_i,y_i,z_i) = (1/8)\cdot \mathsf{idx}(x_i,y_i,z_i) = (1/8)\left(\frac{x_i(1-z_i)}{2} + \frac{y_i(1+z_i)}{2}\right) = \frac{1}{16}(x_i - x_iz_i + y_i + y_iz_i). \tag{54}$$

It is easy to see that $\|\phi\| = 1$, ϕ is perfectly correlated with idx, and ϕ has pure high degree 1, as required by a dual witness for the fact that $\deg_\pm(\mathsf{idx}) \ge 1$. Then we define $\nu \coloneqq \psi \star \phi$. Concretely, if we write the input to $f\circ\mathsf{idx}$ as $(x,y,z) \in (\{-1,1\}^n)^3$, then $\nu(x,y,z) = 4^{-n}\cdot(\psi\circ\mathsf{idx})(x,y,z)$.

Lemma 40 implies that $\|\nu\|_1 = 1$, while Theorem 43 shows that $\langle \nu, f\circ\mathsf{idx}\rangle = \langle f,\psi\rangle > \varepsilon$. It remains to show that for all $S\subseteq[3n]$,

$$\left|\sum_{(x,y,z)\in\{-1,1\}^{3n}} \nu(x,y,z)\chi_S(x,y,z)\right| \le 2^{-d}. \tag{55}$$

Recall that $\hat{\nu}(S)$ denotes the S'th Fourier coefficient of ν, i.e., $\nu = \sum_{S \subseteq [3n]} \hat{\nu}(S) \cdot \chi_S(x)$, and note that the left hand side of Equation (55) is

$$8^n \cdot \hat{\nu}(S). \tag{56}$$

Hence, we establish Equation (55) by explicitly writing out the Fourier coefficients of ν in terms of those of ψ. Writing out ψ in terms of its Fourier coefficients yields:

$$\psi(w) = \sum_{T \subseteq [n]} \hat{\psi}(T) \cdot \chi_T(w). \tag{57}$$

Since $\|\psi\|_1 = 1$, $|\hat{\psi}(T)| \leq 2^{-n}$ for all T. Moreover, since ψ has pure high degree d, we know that $\hat{\psi}(T) = 0$ for all $|T| \leq d$. Because $\nu = 4^{-n}\,(\psi \circ \mathsf{idx})$ and $\mathsf{idx}(x_i, y_i, z_i) = \frac{1}{2}\,(x_i - x_i z_i + y_i + y_i z_i)$, each non-zero term in Equation (57), when composed with idx and expanded out via the distributive law, turns into a sum of $4^{|T|}$ parities, each with coefficient $4^{-n} \cdot 2^{-|T|} \leq 4^{-n} \cdot 2^{-d}$. Moreover, for any two distinct sets $S, S' \subseteq [n]$, the set of parities appearing in the expansion of $\chi_S \circ \mathsf{idx}$ is disjoint from the set of parities appearing in the expansion of $\chi_{S'} \circ \mathsf{idx}$. This is because each term in Equation (54) involves x_i or y_i, and hence if $i \in S \setminus S'$, then each parity in the expansion of $(\chi_S \circ \mathsf{idx})(x, y, z)$ involves (exactly one of) x_i or y_i while no parity in the expansion of $(\chi_{S'} \circ \mathsf{idx})(x, y, z)$ involves x_i or y_i. We conclude that the maximum magnitude of any Fourier coefficient of ν is at most $\max_{T \subseteq [n]} |\hat{\psi}(T)| \cdot 4^{-n} \cdot 2^{-|T|} \leq 2^{-n} \cdot 4^{-n} \cdot 2^{-d} = 8^{-n} \cdot 2^{-d}$. Equation (55) follows by combining this with Equation (56).

□

Discussion. It is worth reflecting upon what properties of the 3-bit indexing gadget idx we exploited in the proof of Lemma 78, to ensure that $f \circ \mathsf{idx}$ had large ε-approximate weight. We use just two properties of $\mathsf{idx}(x_i, y_i, z_i) = \frac{1}{2}\,(x_i - x_i z_i + y_i + y_i z_i)$. First, that it is balanced, i.e., its degree-0 Fourier coefficient is 0, and second, that all of its Fourier coefficients have magnitude at most $1/2$.

The balanced property ensured that idx (suitably scaled to have ℓ_1-norm 1) is self-dual, in the sense of witnessing that its own threshold degree is at least 1. This meant that the dual block composition $\psi \star \phi$ of ψ with $\phi = \frac{1}{8}\mathsf{idx}$ is just a scaled version of the literal block composition $\psi \circ \phi$. This in turn enabled us to compute the Fourier coefficients of $\psi \star \phi$. We exploited the second property (that all of idx's Fourier coefficients are $1/2$) to conclude that the Fourier coefficients of $\psi \star \phi$ itself are exponentially small as required for any approximate-weight dual witness.

Some simpler gadgets and why they fail. There are two natural gadgets that are even simpler than idx, namely the two-bit $\oplus$ and AND gadgets. As discussed at the end of Section 10.3, our lifting theorems such as Theorem 77 do *not* hold for these gadgets, and indeed they both fail to satisfy one of the two properties above. In the former case, $x \oplus y = x \cdot y$ is balanced, but has a Fourier coefficient of magnitude 1. In the latter case,

$$\mathsf{AND}(x, y) = \frac{1}{2}(1 + x + y - xy)$$

is not balanced.

It is not just that the proof of Lemma 78 breaks down for these gadgets: the lemma is false in general if we replace idx with $\oplus_2$ or AND_2. To see this for $\oplus_2$, consider letting $f = \oplus_n$, which has

threshold degree n. As discussed in Section 10.3, $f \circ \oplus_2 = \oplus_{2n}$, which is computed exactly by a polynomial of weight just 1. Similarly, as in Section 10.3, for AND_2, consider letting $f = \mathsf{AND}_n$, which has $\widetilde{\deg}(f) \geq \Omega(\sqrt{n})$. Then $f \circ \mathsf{AND}_2 = \mathsf{AND}_{2n}$, which is computed exactly by a polynomial of weight $O(1)$, namely $1 - 2^{-n+1} + \sum_{S \subseteq [n]} (-1)^{|S|+1} \cdot 2^{-n+1} \cdot \chi_S(x)$.

10.4.2 Step 2: From high approximate weight to high approximate rank

Let $h\colon \{-1,1\}^m \to \{-1,1\}$ be any Boolean function and define $F\colon \{-1,1\}^m \times \{-1,1\}^m \to \{-1,1\}$ via $F = h \circ \oplus_2$, i.e., $F(x,y) = h(x \oplus y)$. Recall that M_F denotes the $2^m \times 2^m$ matrix whose (x,y)'th entry is $F(x,y)$ (here, we are indexing the 2^m rows and columns of M_F by bit-vectors in $\{-1,1\}^m$ in the natural way). In this section, our goal is to show that if h has large approximate weight, then M_F has large approximate rank.

Lemma 79. Suppose $\widetilde{\mathrm{weight}}_\varepsilon(h) \geq 2^d$ and let $F = h \circ \oplus_2$. Then for any $\delta \in [0, \varepsilon]$, $\widetilde{\mathrm{rank}}_\delta(M_F) \geq \big((2^d + \varepsilon - \delta)/(1+\delta)\big)^2$.

Theorem 77 will follow by combining this result with Lemma 78.

The optimal approximation technique. Lemma 78 shows that the following approach to constructing a low-rank ε-approximation to the matrix M_F is essentially optimal: take the *sparsest* polynomial p (over the parity basis) that ε-approximates h, let $P(x,y) = p(x \oplus y)$, and approximate M_F with the matrix M_P. Since p is an ε-approximation to h, clearly, $|(M_F)_{x,y} - (M_P)_{x,y}| \leq \varepsilon$ for all $x, y \in \{-1,1\}^m$.

Moreover, $\mathrm{rank}(M_P)$ is at most $\mathrm{sparsity}(p)$ (recall from Section 2.1 that the sparsity of p is the number of non-zero Fourier coefficients of p). To see this, for any subset $S \subseteq [m]$, let us abuse notation and view χ_S as a column vector of length 2^m whose y'th entry is $\chi_S(y)$. Then clearly $\chi_S \cdot \chi_S^T$ is a rank-1 matrix with 2^m rows and 2^m columns, and with (x,y)'th entry equal to $\chi_S(x) \cdot \chi_S(y) = \chi_S(x \oplus y)$. Hence, we can write

$$M_P = \sum_{S \subseteq [n]} \hat{p}(S) \cdot \big(\chi_S \cdot \chi_S^T\big)$$

as a sum of at most $\mathrm{sparsity}(p)$-many rank-1 matrices. Since matrix rank is subadditive, we conclude that the rank of M_P is at most the Fourier sparsity of p.

Discussion and intuition. In this section, we show that this technique for approximating $F(x,y) = h \circ \oplus_2$ by a low-rank matrix function P is essentially optimal. That is, we show that for M_F to be ε'-approximated by a rank-r matrix, h must be ε-approximated by a sparse polynomial p for some ε that is very close to ε'. Here, the sparsity of p is roughly proportional to r, the rank of the approximation to M_F.

To be more precise, our analysis directly shows that if M_F is ε'-approximated by a rank-r matrix where $r \approx 2^d$, then h can be $\varepsilon' + 2^{-\Theta(d)}$ approximated by a polynomial q whose Fourier weight $\mathrm{weight}(q)$ is not much larger than r. In turn, any weight-W polynomial q can be ε''-approximated by a polynomial p of sparsity at most $\ell := 100 \cdot Wm/(\varepsilon'')^2$. This follows from a probabilistic construction due to Bruck and Smolensky [BS92]: randomly sample ℓ Fourier basis functions $\chi_{S_1}, \ldots, \chi_{S_\ell}$ with replacement, where χ_S is chosen with probability proportional to $|\hat{q}(S)|$, and let $p(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \chi_{S_i}$. Chernoff bounds imply that, for each input $x \in \{-1,1\}^m$, with probability at

least $1 - 2^{-2m}$, $|p(x) - q(x)| \leq \varepsilon''$. Hence, a union bound over all $x \in \{-1,1\}^m$ guarantees that with non-zero probability, $|p(x) - q(x)| \leq \varepsilon''$ for all $x \in \{-1,1\}^m$. Putting everything together, we conclude as claimed that if M_F has (ε')-approximate rank at most $r \approx 2^d$, then h can be $(\varepsilon + 2^{-\Theta(d)})$-approximated by a polynomial of sparsity $2^{\Theta(d)}$.

Overview of the proof. Recall that our goal is to show that if $F(x,y) = h(x \oplus y)$ and h has large approximate weight, then M_F has large approximate rank. Let $\nu\colon \{-1,1\}^m \to \mathbb{R}$ be the dual witness for the fact that h has large approximate weight, as per Section 10.4.1. Then ν is well-correlated with h but has tiny Fourier coefficients, meaning it is poorly correlated with *any* parity function χ_S (regardless of the degree of χ_S). Letting $\eta(x,y) = \nu(x \oplus y)$, this means that η is poorly correlated with the rank-1 matrix $\chi_S \cdot \chi_S^T$.

The rest of the analysis essentially shows that η is effectively a dual witness for the high approximate rank of M_F. This proceeds as follows.

The first thing we do is show that M_ν has small spectral norm (i.e., all of its eigenvalues have small magnitude). We show this by establishing that the eigenvalues of M_ν are exactly given by the Fourier coefficients of ν (up to scaling), with the corresponding eigenspaces precisely given by the rank-1 matrices $\chi_S \cdot \chi_S^T$. Intuitively, this result means that not only is M_ν poorly correlated with the "special" rank-1 matrices of the form $\chi_S \cdot \chi_S^T$, but in fact M_ν is poorly correlated with *all* rank-1 matrices.

With the above spectral norm bound in hand, the remainder of the analysis mimics the following first-principles proof that a dual polynomial $\psi\colon \{-1,1\}^n \to \mathbb{R}$ indeed lower bounds the ε-approximate degree of f.

First-principles proof that a dual polynomial lower bounds approximate degree. If ψ has pure high degree d, ℓ_1-norm 1, and $\sum_{x \in \{-1,1\}^n} \psi(x) f(x) > \varepsilon$, then ψ rules out a degree-d ε-approximation p for f via the following calculation. On the one hand, since p has degree at most d and ψ has pure high degree d, we know that

$$\langle \psi, p \rangle = \sum_{x \in \{-1,1\}^n} \psi(x) p(x) = 0. \tag{58}$$

On the other hand, since p approximates f and ψ is strictly greater than ε-correlated with f,

$$\begin{aligned} \langle \psi, p \rangle &= \sum_{x \in \{-1,1\}^n} \psi(x) f(x) - \sum_{x \in \{-1,1\}^n} \psi(x)(f(x) - p(x)) \\ &> \varepsilon - \sum_{x \in \{-1,1\}^n} |\psi(x)| \cdot |f(x) - p(x)| \geq \varepsilon - \|\psi\|_1 \cdot \varepsilon = 0. \end{aligned} \tag{59}$$

This contradicts Equation (58).

Our proof in this section ports this argument to the matrix setting, with the matrix M_F in place of f and η in place of the dual polynomial ψ for f. The spectral norm bound of Equation (66) for M_η will mimic the effect of ψ being uncorrelated with low-degree polynomials, as it implies that M_η is only very weakly correlated with low-rank matrices. Meanwhile, since ν is well-correlated with h, we can show that M_η is well-correlated with M_F (our proof of this exploits the fact that η can be expressed as the dual block composition of ν with a dual polynomial for the fact that

$\deg_{\pm}(\oplus_2) \geq 2$). Together, these two facts imply that M_F cannot be approximated well by low-rank matrices, via a calculation analogous to Equation (59).

The formal analysis follows.

Bounding the spectral norm of M_η. Recall that the spectral norm of a symmetric matrix M (Definition 64) is its largest eigenvalue in absolute value. The following lemma bounds the spectral norm of M_η in terms of the Fourier spectrum of ν. This lemma is folklore; we present a proof given in Mande's thesis [Man18, Lemma 2.2.3].

Lemma 80. Let $\nu\colon \{-1,1\}^m \to \mathbb{R}$ be a real-valued function, let $\eta(x,y) = 2^{-m} \cdot \nu(x \oplus y)$, and recall that M_η is the $2^m \times 2^m$ matrix with (x,y)'th entry equal to $\eta(x,y)$. Then $\|M_\eta\| = \max_{S \subseteq [m]} \hat{\nu}(S)$, where recall that $\hat{\nu}(S)$ denotes the S'th Fourier coefficient of ν.

Proof. In fact, the proof will identify all 2^m eigenvalues of M_η, revealing them to be

$$\{\hat{\nu}(S)\colon S \subseteq [m]\}.$$

Recall from the discussion in the overview of this proof that for any subset $S \subseteq [m]$, we abuse notation and view χ_S as a vector of length 2^m whose y'th entry is $\chi_S(y)$. We show that χ_S is an eigenvector of M_η with eigenvalue $\hat{\nu}(S)$.

Note that

$$(M_\eta)_{x,y} = 2^{-m} \sum_{T \subseteq [m]} \hat{\nu}(T) \cdot \chi_T(x \oplus y) = 2^{-m} \sum_{T \subseteq [m]} \hat{\nu}(T) \cdot \chi_T(x) \cdot \chi_T(y).$$

Hence, for each $x \in \{-1,1\}^m$,

$$\begin{aligned}(M_\eta \cdot \chi_S)_x = \sum_{y \in \{-1,1\}^m} (M_\eta)_{x,y} \cdot \chi_S(y) &= 2^{-m} \sum_{y \in \{-1,1\}^m} \sum_{T \subseteq [m]} \hat{\nu}(T) \cdot \chi_T(x) \cdot \chi_T(y) \cdot \chi_S(y) \\ &= 2^{-m} \sum_{T \subseteq [m]} \hat{\nu}(T) \cdot \chi_T(x) \cdot \sum_{y \in \{-1,1\}^m} \chi_T(y) \cdot \chi_S(y).\end{aligned}$$

Let $S \Delta T$ denote the symmetric difference between sets S and T, the above equals:

$$2^{-m} \sum_{T \subseteq [m]} \hat{\nu}(T) \cdot \chi_T(x) \cdot \left(\sum_{y \in \{-1,1\}^m} \chi_{S \Delta T}(y) \right).$$

The sum in parenthesis is 0 if $S \neq T$, and otherwise is 2^m. Hence, the above equals:

$$\hat{\nu}(S) \cdot \chi_S(x).$$

This establishes that χ_S is an eigenvector of M_η with eigenvalue $\hat{\nu}(S)$ as claimed.

□

Before completing the proof of Theorem 77, we introduce some additional matrix-analytic notions.

Matrix norms. Let M be a symmetric $N \times N$ matrix $M \in \mathbb{R}^{N \times N}$ and recall that M has N real eigenvalues, say, $\lambda_1, \ldots, \lambda_{2^m}$, where we order the eigenvalues so that $|\lambda_1| \geq |\lambda_2| \cdots \geq |\lambda_{2^m}|$. The trace norm of M, denoted $\|M\|_\Sigma$, is

$$\|M\|_\Sigma = \sum_{i=1}^{N} |\lambda_i|,$$

while the Frobenius norm is

$$\|M\|_F \coloneqq \sqrt{\sum_{i=1}^{N} \lambda_i^2}. \tag{60}$$

That is, if λ is the vector of eigenvalues of M, then the spectral norm is the ℓ_∞-norm of λ, the trace norm is the ℓ_1-norm, and the Frobenius norm is the ℓ_2-norm.

A basic property of the Frobenius norm is that it also equals

$$\sqrt{\sum_{1 \leq i,j \leq N} M_{i,j}^2}. \tag{61}$$

This holds by the following reasoning. Thinking of the rows of the matrix M as a collection of N vectors each of length N, Equation (61) is the sum of the ℓ_2-norms of the vectors. Meanwhile, Equation (60) is the ℓ_2-norm of the matrix with λ along the diagonal, which is simply a representation of the same collection of N vectors in a different orthonormal basis (given by the eigenvectors of M). And an orthornormal change of basis preserves the ℓ_2-norm of each vector in the collection.

For two $N \times N$ matrices, A, B, we let $\langle A, B \rangle = \sum_{1 \leq i,j \leq N} A_{i,j} \cdot B_{i,j}$. That is, viewing the matrices as length-N^2 vectors, $\langle A, B \rangle$ denotes their correlation exactly as in Section 6. The final property of the above matrix norms that we will need is:

$$\langle A, B \rangle \leq \|A\| \cdot \|B\|_\Sigma. \tag{62}$$

This is a matrix-analog of the fact that for any two vectors a, b, their inner product $\langle a, b \rangle$ is at most $\|a\|_\infty \cdot \|b\|_1$.

Completing the proof of Lemma 79. Recall that Lemma 79 states that if $\widetilde{\text{weight}}_\varepsilon(h) \geq 2^d$, then $F = h \circ \oplus_2$ has large approximate rank. Specifically, we must show that $\widetilde{\text{rank}}_\delta(M_F) \geq \big((2^d + \varepsilon - \delta)/(1+\delta)\big)^2$.

Proof. Appealing to strong duality in the LP characterizing approximate weight in the proof of Lemma 78, we have that there exists a function ν and parameter $\gamma \in (0, 1]$ such that

$$\|\nu\|_1 = 1, \tag{63}$$

$$\langle \nu, h \rangle \geq \varepsilon + 2^d \gamma \tag{64}$$

and

$$\left| \sum_{x \in \{-1,1\}^m} \nu(x) \chi_S(x) \right| \leq \gamma \text{ for all } S \subseteq [m],$$

or equivalently that

$$\hat{\nu}(S) \leq 2^{-m} \cdot \gamma \text{ for all } S \subseteq [m]. \tag{65}$$

Define $\mu\colon \{-1,1\}^2 \to \mathbb{R}$ as $\mu(a,b) = \frac{1}{4}a \cdot b$. That is, μ is the natural dual polynomial for the fact that $\deg_{\pm}(\oplus_2) = 2$. Define $\eta\colon \{-1,1\}^m \times \{-1,1\}^m \to \mathbb{R}$ via $\eta = \nu \star \mu$. That is, for $x, y \in \{-1,1\}^m$,

$$\eta(x,y) = 2^{-m}\nu(x \oplus y).$$

Recall that $M_\eta \in \mathbb{R}^{2^m \times 2^m}$ is the matrix whose (x,y)'th entry equals $\eta(x,y)$. By Equation (65) and Lemma 80, we conclude that

$$\|M_\eta\| \leq 2^{-m} \cdot \gamma. \tag{66}$$

Meanwhile, by Equations (63) and (64), we have:

$$\sum_{x,y \in \{-1,1\}^m} (M_\eta)_{x,y} \cdot (M_F)_{x,y} = \langle \eta, F \rangle = \langle \nu, h \rangle \geq \varepsilon + 2^d\gamma. \tag{67}$$

Here, the second equality follows from Theorem 43 and the fact that the dual witness $\eta = \nu \star \mu$ for $F = h \circ \oplus_2$ is the dual block composition of ν and a dual witness μ for the fact that $\deg_{\pm}(\oplus_2) \geq 2$.

Suppose R is a $2^m \times 2^m$ matrix of rank-r that entry-wise approximates M_F to error at most δ.[38] Then the absolute value of each entry of R is at most $1 + \delta$, and hence by Equation (61), $\|R\|_F \leq (1+\delta) \cdot 2^m$. We have:

$$\begin{aligned}\langle M_\eta, R \rangle &\leq \|M_\eta\| \cdot \|R\|_\Sigma \leq \|M_\eta\| \cdot \|R\|_F \cdot \sqrt{r} \leq \|M_\eta\| \cdot (1+\delta)\sqrt{r} \cdot 2^m \\ &\leq 2^{-m}\gamma\sqrt{r}2^m(1+\delta) = \gamma\sqrt{r}(1+\delta). \end{aligned} \tag{68}$$

Here, the first inequality follows from Equation (62), while the second holds by the following reasoning. Since R has rank r, its vector of eigenvalues has at most r non-zero entries. By the Cauchy-Schwarz inequality the ℓ_1 norm of this vector (which equals the trace norm of R) is at most $\sqrt{r}$ times the ℓ_2-norm (which equals the Frobenius norm of R).

On the other hand,

$$\begin{aligned}\langle M_\eta, R \rangle &= \sum_{x,y \in \{-1,1\}^m \times \{-1,1\}^m} R_{x,y} \cdot (M_\eta)_{x,y} \\ \geq \left(\sum_{x,y \in \{-1,1\}^m \times \{-1,1\}^m} (M_F)_{x,y} \cdot (M_\eta)_{x,y} \right) &- \sum_{x,y \in \{-1,1\}^m \times \{-1,1\}^m} |R_{x,y} - (M_F)_{x,y}| \cdot |(M_\eta)_{x,y}| \\ &> \varepsilon + 2^d\gamma - \delta \cdot \|\eta\|_1 = \varepsilon + 2^d\gamma - \delta. \end{aligned} \tag{69}$$

Here, the final inequality uses Equation (67), and the final equality follows from $\|\eta\|_1 = 1$. We conclude that $\gamma\sqrt{r}(1+\delta) \geq \varepsilon + 2^d\gamma - \delta$. Hence, using the fact that $0 < \gamma \leq 1$, we get $r \geq (2^d + \varepsilon - \delta)^2/(1+\delta)^2$.

□

[38] Since $M_F = [F(x \oplus y)]_{x,y}$ is a symmetric matrix, we may assume R is as well without loss of generality. Indeed, if R is not symmetric then we can replace it with $\frac{1}{2}(R + R^T)$, which will still be an ε-approximation to M_F and at most double the rank of R.

Putting everything together: Proof of Theorem 77. Our goal now is to complete the proof of Theorem 77, which states that if $\widetilde{\deg}_\varepsilon(f) \geq d$, then $F = f \circ \mathsf{idx} \circ \oplus_2$ has large approximate rank, where idx is the 3-bit indexing gadget.

Proof. Lemma 78 tells us that the function $h = f \circ \mathsf{idx}$ has approximate weight $\widetilde{\mathrm{weight}}_{\varepsilon - 2^{-d/2}} \geq 2^{d/2}$. Letting $F = h \circ \oplus_2$, Lemma 79 then shows that $\widetilde{\mathrm{rank}}_\delta(M_F) \geq (2^{d/2} + \varepsilon - 2^{-d/2} - \delta)^2/(1+\delta)^2$. □

10.4.3 Communication applications

Let g be the 6-bit gadget from Theorem 77. Suppose that we know that $\widetilde{\deg}_\varepsilon(f) \geq d$. Theorem 77 gives a δ-approximate rank lower bound for $f \circ g$ of $2^{\Omega(d)}$ where $\delta = \varepsilon - 2^{-O(d)}$. This additive loss of $2^{-O(d)}$ in the error parameter is typically irrelevant when $\varepsilon \in (0, 1 - 2^{-d})$, which is the appropriate parameter regime for applications to $\mathbf{BQP}^{\mathrm{cc}}$ and $\mathbf{PP}^{\mathrm{cc}}$. Specifically, we can generically transform any (1/3)-approximate degree lower bound for f into a quantum communication lower bounds for $f \circ g$, and generically transform any **PP** *query* lower for f into a $\mathbf{PP}^{\mathrm{cc}}$ lower bound for $f \circ g$.

Corollary 81. If $\widetilde{\deg}_{1/3}(f) \geq d$, then $\mathbf{BQP}^{\mathrm{cc}}(f \circ g) \geq \Omega(d)$.

Proof. Combine Theorem 77 with Fact 74. □

Sherstov [She11a] showed how to use Corollary 81 to recover Razborov's celebrated result [Raz03] that the quantum communication complexity of the Disjointness function, DISJ, is $\Omega(\sqrt{n})$. Here, for $x, y \in \{-1,1\}^n$, $\mathsf{DISJ}(x,y) = \mathsf{NOR}(x \wedge y)$ is the "lift" (i.e., composition) of the NOR_n function with the two-bit AND gadget. One can interpret x and y as the indicator vectors of two sets X and Y over a universe of size n, with $x_i = -1$ ($y_i = -1$) interpreted as indicating that $i \in X$ ($i \in Y$), and DISJ evaluates to -1 if and only if the two sets are disjoint.

Corollary 82 ([She11a], original result due to [Raz03])**.** $\mathbf{BQP}^{\mathrm{cc}}(\mathsf{DISJ}_n) \geq \Omega(\sqrt{n})$.

Proof. Since $\widetilde{\deg}(\mathsf{NOR}_n) \geq \Omega(\sqrt{n})$, Corollary 81 implies that $\mathbf{BQP}^{\mathrm{cc}}(\mathsf{NOR}_n \circ g) \geq \Omega(\sqrt{n})$ where g is the 6-bit gadget from Theorem 77. We now explain that $\mathsf{NOR} \circ g$ can be embedded into an instance of DISJ on $O(n)$ bits. That is, Alice and Bob can reduce the task of computing $(\mathsf{NOR} \circ g)(x,y)$ to the task of evaluating $\mathsf{DISJ}_N(x', y')$ where $N = O(n)$, x' depends only on x, and y' depends only on y. Hence, any communication protocol for DISJ_N of cost $o(\sqrt{N})$ would imply a communication protocol for $\mathsf{NOR} \circ g$ of cost $o(\sqrt{n})$.

First, observe that for *any* gadget function g defined on a constant number of bits, $\mathsf{OR} \circ g$ is computed by a DNF formula of size $O(n)$. This is because any function g on $O(1)$ bits is computable by a DNF of size $O(1)$. Hence, $\mathsf{OR} \circ g$ is computed by a depth-three circuit of size $O(n)$, where the top two layers are OR gates. By collapsing the adjacent layers of OR gates into a single OR gate, we obtain a DNF of size $O(n)$.

Next, we explain that for any DNF $F(x,y)$ of size N, computing $F(x,y)$ can be reduced to an instance of $\mathsf{DISJ}_N(x', y')$ where x' depends only x and y' depends only on y. Write $F(x,y) = \mathsf{OR}(C_1(x,y), \ldots, C_N(x,y))$ where each C_i is a conjunction (i.e., an AND of literals). We can partition the inputs to C_i into two halves, say, A and B: those literals fed into C_i that depend on x, and those that depend on y. Define $x'_i = -1$ if and only all literals in A evaluate to true, and define $y'_i = -1$ if and only if all literals in B evaluate to true. Then $F(x,y) = 1$ if and only if $\mathsf{DISJ}(x', y') = -1$.

Hence, Alice and Bob can determine whether $(\mathsf{NOR}_n \circ g)(x, y) = 1$ by representing $\mathsf{NOR} \circ g$ as a DNF of size $N = O(n)$, transforming the DNF into an equivalent instance (x', y') of DISJ_N as per the previous paragraph, applying a $\mathbf{BQP}^{\mathrm{cc}}$ protocol for DISJ_N on input (x', y'), and negating the output. If the $\mathbf{BQP}^{\mathrm{cc}}$ protocol for DISJ_N has communication cost c, then so does the resulting protocol for $\mathsf{NOR}_n \circ g$. □

Our next corollaries focus on **PP** communication complexity rather than $\mathbf{BQP}^{\mathrm{cc}}$.

Corollary 83. If $\widetilde{\deg}_{1-2^{-d}}(f) \geq d$ for some $d = \omega(\log n)$, then $\mathbf{PP}^{\mathrm{cc}}(f \circ g) \geq \Omega(d)$. Hence, if $\mathbf{PP}^{\mathrm{dt}}(f) \geq d$, then $\mathbf{PP}^{\mathrm{cc}}(f \circ g) \geq \Omega(d/\log n)$.

Proof. Theorem 77 shows that if $\widetilde{\deg}_{1-2^{-d}}(f) \geq d$, then $\widetilde{\mathrm{rank}}_{1-2^{-d}-2^{-d/2}}(M_{f \circ g}) \geq 2^{\Omega(d)}$. Fact 76 then yields the lower bound $\mathbf{PP}^{\mathrm{cc}}(f \circ g) \geq \Omega(d)$. The statement relating $\mathbf{PP}^{\mathrm{dt}}(f)$ to $\mathbf{PP}^{\mathrm{cc}}(f \circ g)$ follows from Fact 72, which says that if $\mathbf{PP}^{\mathrm{dt}}(f) \geq d$, then $\widetilde{\deg}_{1-2^{-d'}}(f) \geq d'$ for some $d' = \Omega(d/\log n)$. □

One example application of Corollary 83 yields a tight $\mathbf{PP}^{\mathrm{cc}}$ lower bound for the well-known inner-product-mod-two function $\mathsf{IP2}_n \colon \{-1,1\}^n \times \{-1,1\}^n \to \{-1,1\}$, where

$$\mathsf{IP2}_n(x, y) = \oplus_n(x \wedge y).$$

Corollary 84. $\mathbf{PP}^{\mathrm{cc}}(\mathsf{IP2}) \geq \Omega(n)$.

Proof. Apply Corollary 83 with $f = \oplus_n$. Since $\deg_{\pm}(f) = n$, we conclude that $\mathbf{PP}^{\mathrm{cc}}(f \circ g) \geq \Omega(n)$. Now analogously to Corollary 82, one can show that the task of computing $(f \circ g)(x, y)$ reduces to the task of evaluating $\mathsf{IP2}_N(x', y')$ where $N = O(n)$. □

For reasons discussed shortly (Section 10.4.4), it is of interest to obtain large $\mathbf{PP}^{\mathrm{cc}}$ lower bounds for functions in AC^0, and unfortunately the $\mathsf{IP2}$ function from Corollary 84 is not. The first polynomial $\mathbf{PP}^{\mathrm{cc}}$ lower bounds for AC^0 functions were proved in independent papers of [BVdW07] and [She09]. As shown by Sherstov [She11a], the techniques developed in this section give a modular (and quantitatively stronger) proof of such a lower bound. Recall that Minsky and Papert proved an $\Omega(n^{1/3})$ threshold degree lower bound for the Minsky-Papert CNF $f = \mathsf{AND}_{n^{1/3}} \circ \mathsf{OR}_{n^{2/3}}$ (Theorem 26), and hence as per Fact 72, $\mathbf{PP}^{\mathrm{dt}}(f) \geq \Omega(n^{1/3})$. Observe that if f is in AC^0, then so is $f \circ g$ (this is because, as mentioned in the proof of Corollary 82, any function g defined on a constant number of bits is itself computed by a constant-sized DNF or CNF). Hence applying Corollary 83 to f yields an AC^0 function $F = f \circ g$ (in fact, a depth-three circuit) with $\mathbf{PP}^{\mathrm{cc}}(F) \geq \Omega(n^{1/3})$.

Corollary 85 ([She11a])**.** There is an AC^0 function F with $\mathbf{PP}^{\mathrm{cc}}(F) \geq \Omega(n^{1/3})$.

There are now AC^0 functions F known with $\mathbf{PP}^{\mathrm{cc}}(F) \geq \Omega(n^{1-\delta})$ [BT19a], for any constant $\delta > 0$ (in fact, even $\mathbf{UPP}^{\mathrm{cc}}(F) \geq \Omega(n^{1-\delta})$ [SW19]). See Section 8.3 for further details. We remark that we will prove results in Section 10.5 that strengthen the $\mathbf{PP}^{\mathrm{cc}}$ lower bounds of Corollaries 84 and 85 to $\mathbf{UPP}^{\mathrm{cc}}$ lower bounds.

Separating $\mathbf{PP}^{\mathrm{cc}}$ and $\mathbf{UPP}^{\mathrm{cc}}$. Another application of Corollary 83 is a separation between $\mathbf{PP}^{\mathrm{cc}}(F)$ and $\mathbf{UPP}^{\mathrm{cc}}(F)$ for an explicit function F. Specifically, Beigel [Bei94] separated these classes in the query complexity setting, identifying a function f with threshold degree just 1 (hence $\mathbf{UPP}^{\mathrm{dt}(f)} \leq O(1)$, see Fact 71) for which $\mathbf{PP}^{\mathrm{dt}}(f) \geq \Omega(n^{1/3})$.[39] Corollary 83 lifts the separation to the communication setting. That is, it implies that for $F = f \circ g$, $\mathbf{PP}^{\mathrm{cc}}(F) \geq \Omega(n^{1/3})$, while $\mathbf{UPP}^{\mathrm{cc}}(F) \leq O(\log n)$ by standard communication simulation of the $\mathbf{UPP}^{\mathrm{dt}}$ query protocol for f (Section 10.3). Such a separation between $\mathbf{UPP}^{\mathrm{cc}}$ and $\mathbf{PP}^{\mathrm{cc}}$ was first given by Buhrman et al. [BVdW07] and independently by Sherstov for a different function [She08]. Subsequent works [She13c, Tha16, BT18b, She21] have obtained improved separations, yielding the following result.

Corollary 86. There is an explicit function $F : \{-1,1\}^n \times \{-1,1\}^n \to \{-1,1\}$ such that $\mathbf{UPP}^{\mathrm{cc}}(F) \leq O(\log n)$ and $\mathbf{PP}^{\mathrm{cc}}(F) \geq \Omega(n)$.

Lower bounds for other communication classes. In characterizing the $\mathbf{PP}^{\mathrm{cc}}$ complexity of $f \circ g$ in terms of approximate degree of f, Fact 72 has applications to additional communication complexity classes, notably $\mathbf{QMA}^{\mathrm{cc}}$, where **QMA** is the quantum analog of **NP**. This is because it is known that any $\mathbf{QMA}^{\mathrm{cc}}$ protocol can be simulated by a $\mathbf{PP}^{\mathrm{cc}}$ protocol with at most a quadratic blowup in cost [Vya03, MW05]. Hence, any $\mathbf{PP}^{\mathrm{cc}}$ lower bound for a function also implies a $\mathbf{QMA}^{\mathrm{cc}}$ lower bound. For example, combined with Corollary 84, this yields an $\Omega(\sqrt{n})$ $\mathbf{QMA}^{\mathrm{cc}}$ lower bound for the inner product function, which is tight up to a logarithmic factor by a well-known result of Aaronson and Wigderson [AW09]. Other works have used approximate degree to prove tight or nearly tight **QMA** lower bounds for the Disjointness function [Kla11], the permutation testing problem [ST19], and a problem called approximate counting [AKKT20].

10.4.4 $\mathsf{MAJ} \circ \mathsf{LTF}$ circuit lower bounds

Recall that the Majority function MAJ, takes as input a bit-vector in $\{-1,1\}^n$ and outputs -1 if and only if at least $n/2$ of the bits are -1. Meanwhile, a *linear threshold function* (LTF) is any Boolean function f of threshold degree at most 1. That is, there are real weights $w_0, \ldots, w_n \in \mathbb{R}$ such that $f(x_1, \ldots, x_n) = \operatorname{sgn}(w_0 + \sum_{i=1}^n w_i x_i)$. An LTF is also called a *halfspace*.

A $\mathsf{MAJ} \circ \mathsf{LTF}$ circuit is any depth-two circuit of size $s+1$ that outputs the majority vote of s LTFs. In this section, we will see that a $\mathbf{PP}^{\mathrm{cc}}$ lower bound for a function F implies a lower bound on the size of $\mathsf{MAJ} \circ \mathsf{LTF}$ circuits computing F.

Here are two reasons to care about such lower bounds. First, a notorious open problem in circuit complexity is to identify an explicit Boolean function f that cannot be computed by $\mathsf{LTF} \circ \mathsf{LTF}$ circuits of polynomial size. $\mathsf{MAJ} \circ \mathsf{LTF}$ is a natural subclass of such circuits, so proving lower bounds against such circuits is a natural step toward resolving the $\mathsf{LTF} \circ \mathsf{LTF}$ question. Second, as discussed in Section 5.3, Allender [All89] showed that quasipolynomial-size depth-3 majority circuits can compute all of AC^0, and for a long time it was open whether the same is true of depth-2 majority circuits. The lower bounds in this section show that this is not the case, even for $\mathsf{MAJ} \circ \mathsf{LTF}$ circuits.

Theorem 87 ([Nis93])**.** Suppose the function $F(x,y)$ is computable by a $\mathsf{MAJ} \circ \mathsf{LTF}$ circuit $\mathcal{C}$ of size $s+1 \geq n$. Then $\mathbf{PP}^{\mathrm{cc}}(F) \leq O(\log^2 s)$.

[39] Beigel's function f, called $\mathsf{ODDMAXBIT}$, is in fact a very special kind of halfspace known as a *decision list*. Beigel showed that $\deg_\varepsilon(f) \geq d$ for $\varepsilon = 1 - 2^{-\Omega(n/d^2)}$. In particular, $\deg_{1-2^{-n^{1/3}}}(f) \geq \Omega(n^{1/3})$, implying that $\mathbf{PP}^{\mathrm{dt}}(f) \geq \Omega(n^{1/3})$ by Fact 72.

Combining Theorem 87 with Corollary 85 yields an AC^0 function that cannot be computed by $\mathsf{MAJ} \circ \mathsf{LTF}$ circuits of size smaller than $2^{\Omega(n^{1/3})}$.

Proof. Assume for simplicity that s is odd. In the $\mathbf{PP}^{\mathrm{cc}}$ protocol for F, Alice uses her private randomness to select a random LTF gate of $\mathcal{C}$ and send the identity of this gate to Bob. This costs $\log s$ bits of communication. Letting $\varepsilon = 1/s^2$, Alice and Bob then execute an ε-error randomized communication protocol (described below) to evaluate the LTF gate at (x, y) and output the result. As we will show, they can accomplish this with communication just $O(\log^2 s)$. The success probability of this protocol is at least $1/2 + 1/(2s) - 1/s^2 > 1/2 + 1/(3s)$. This is because on input (x, y), at least $(s+1)/2$ out of the s LTF gates in $\mathcal{C}$ output $F(x, y)$. Accordingly, the $\mathbf{PP}^{\mathrm{cc}}$ cost of this communication protocol is $O(\log^2 s)$.

Let $x, y \in \{-1, 1\}^n$ and let

$$G(x, y) = \operatorname{sgn}\left(w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{j=1}^{n} w_{n+j} y_j\right) \tag{70}$$

be any LTF; here is the randomized communication protocol for evaluating $G(x, y)$. A basic fact about LTFs is that it can be guaranteed that the w_i's are integers such that $\sum_{i=0}^{2n} |w_i| \leq n^{O(n)}$ [MTT61], so we will assume this for the remainder of the protocol description.

Clearly, to compute $G(x, y)$, it is enough to determine whether

$$w_0 + \sum_{i=1}^{n} w_i x_i \geq -\left(\sum_{j=1}^{n} w_{n+j} \cdot y_j\right). \tag{71}$$

Let X and Y denote the left hand and right hand size of Equation (71) respectively, and observe that X and Y are integers of magnitude $n^{O(n)}$, and hence can be represented in binary with $\ell \leq O(n \log n)$ bits. Let $X^*, Y^* \in \{-1, 1\}^\ell$ denote these binary representations. Since X^* is independent of y and Y^* is independent of x, Alice knows X^* and Bob knows Y^*.

We saw in Section 10.2 that there is an ε-error private-coin randomized communication protocol for the Equality function on n bits with cost $O(\log(n) + \log(1/\varepsilon))$. First, Alice and Bob can run the Equality protocol on input (X^*, Y^*) to determine whether $X^* = Y^*$. If so, they know whether or not Equation (71) holds. If not, the idea is to set $\varepsilon \leq 1/(s^2 \cdot n)$ and to use this subroutine to perform a binary search to determine the highest-order bit at which X^* and Y^* disagree. This is sufficient information for Bob to determine whether or not Equation (71) holds, and hence to determine $G(x, y)$.

That is, first Alice and Bob run the Equality protocol on input the first half of X^* and Y^* to determine if X^* and Y^* agree on their first $\ell/2$ bits; if no, they recurse on the first half of X^* and Y^*; if yes, they recurse on the second half. They continue this process until they have found a bit i such that $X^*_i \neq Y^*_i$, yet X^* and Y^* agree on their first $i - 1$ bits.

The total number of invocations of the Equality protocol during this binary search procedure is at most $O(\log \ell) = O(\log n)$. Since each invocation of the Equality protocol errs with probability at most $1/(s^2 \cdot n)$, the probability that *any* of the invocations fail is at most $O(\log n/(s^2 n))$. Hence Alice and Bob successfully output $G(x, y)$ with probability at least $1 - o(1/s^2)$ as desired. The total communication cost of the protocol is $O(\log(\ell) \cdot \log(1/\varepsilon)) = O(\log(n) \cdot \log(s^2 \cdot n)) = O(\log^2 s)$, where the final equality holds because we assumed $s + 1 \geq n$. □

10.5 Sign-Rank Lower Bounds

Unfortunately, the additive loss of $2^{-O(d)}$ in the error parameter of Theorem 77 is devastating if our goal is to obtain sign-rank lower bounds on $f \circ g$. This is because a sign-rank lower bound requires an ε-approximate rank lower bound for *any* $\varepsilon < 1$, including values of ε that might be (significantly) closer to 1 than $1 - 2^{-O(d)}$. Yet the loss of $2^{-\Theta(d)}$ in the error parameter prevents proving an ε-approximate rank lower bound for $\varepsilon > 1 - 2^{-\Theta(d)}$.

Here is our dream theorem in this context, which unfortunately remains an unproven conjecture:

Conjecture 88. Let g be the six-bit gadget from Theorem 77, and let $f\colon \{-1,1\}^n \to \{-1,1\}$ be any Boolean function with $\deg_\pm(f) \geq d$. Then $\mathrm{rank}_\pm(f \circ g) \geq 2^d$.

Unfortunately, we only know how to prove Conjecture 88 if $\deg_\pm(f) \geq d$ is witnessed by a dual polynomial ψ satisfying an additional *smoothness* condition. This condition requires that ψ is "reasonably large" on all inputs in $\{-1,1\}^n$. Specifically, $|\psi(x)|$ needs to be at least $2^{-d/2} \cdot 2^{-n}$ for all $x \in \{-1,1\}^n$. Fortunately, as we will see later (Section 10.5.1), threshold degree lower bounds for many important functions can be proven via such smooth dual witnesses.

Theorem 89. Let g be the six-bit gadget from Theorem 77, and let $f\colon \{-1,1\}^n \to \{-1,1\}$ be any Boolean function with $\deg_\pm(f) \geq d$ and such that there is a dual witness ψ for this fact (see Section 6) satisfying $|\psi(x)| \geq 2^{-d/2} \cdot 2^{-n}$ for all $x \in \{-1,1\}^n$. Let $F = f \circ g$. Then $\mathrm{rank}_\pm(M_F) \geq 2^{d/3}$.

This particular formulation of Theorem 89 is due to [SW19, Theorem A.4], but it appeared implicitly in work going back to Sherstov [She11b], and then Razborov and Sherstov [RS10].

Overview of the proof. Let $p\colon \{-1,1\}^n \to \mathbb{R}$ be a polynomial that agrees in sign with f and assume that $|p(x)| \leq 1$ for all $x \in \{-1,1\}^n$. Then $\min_{x \in \{-1,1\}^n} |p(x)|$ is called the *margin* of f: it is essentially the "closest" p gets to disagreeing in sign with f at any input.

There are known Boolean functions that have threshold degree d but any degree-d polynomial agreeing in sign with f has margin 2^{-n^d}, which is *doubly-exponentially* small (in d) [Pod08, Pod09, BT18a]. A key lemma of Forster [For02] shows that, nonetheless, any function f of threshold degree at most d is sign-represented by a polynomial p with large *average* margin (by large, we mean "only" singly-exponentially small in d, rather than doubly). That is, if $\deg_\pm(f) \leq d$, then[40] there exists a degree d polynomial p such that

$$\text{(sign-agreement) } p(x) \cdot f(x) > 0 \text{ for all } x \in \{-1,1\}^n \tag{72}$$

$$\text{(large average margin) } 2^{-n} \cdot \sum_x p(x) f(x) \geq 2^{-d}. \tag{73}$$

As we now explain, a threshold degree dual witness ψ (see Section 6) for f that is smooth rules out the existence not only of a degree-d sign-representing polynomial p for f, but also any polynomial p that satisfies Equation (73), even those that *disagree in sign* with f in a limited way. Specifically, suppose p has degree at most d, $|p(x)| \leq 1$ for all $x \in \{-1,1\}^n$, p satisfies Equation (73) and, in place of Equation (72), we have:

$$\text{(bounded sign errors) } |p(x) - f(x)| \leq 1 + 2^{-3d} \text{ for all } x \in \{-1,1\}^n. \tag{74}$$

[40] Equation (73) is not quite accurate; Forster's analysis only implies $2^{-n} \cdot \sum_x p(x) f(x) \geq n^{-d}$. We pretend the n^{-d} is 2^{-d} in this intuitive overview for expository reasons.

We show that this contradicts the existence of a dual polynomial ψ for $\deg_{\pm}(f) \geq d$ satisfying the smoothness condition

$$|\psi(x)| \geq 2^{-d/2} \cdot 2^{-n} \text{ for all } x \in \{-1,1\}^n. \tag{75}$$

This can be seen by the following first-principles analysis. On the one hand, since ψ has pure high degree at least d and p has degree at most d, $\langle \psi, p \rangle = 0$. On the other hand, since p satisfies Equations (73) and (74),

$$\begin{aligned}
\langle \psi, p \rangle &\geq \sum_{x\colon f(x)\cdot p(x) \geq 0} |\psi(x)||p(x)| - \sum_{x\colon f(x)\cdot p(x) < 0} |\psi(x)||p(x)| && (76)\\
&\geq 2^{-d/2} \cdot 2^{-n} \cdot \sum_{x\colon f(x)\cdot p(x) \geq 0} |p(x)| - \sum_{x\colon f(x)\cdot p(x) < 0} |\psi(x)| \cdot 2^{-3d} && (77)\\
&\geq 2^{-d/2} 2^{-d} - 2^{-3d} > 2^{-2d} > 0. && (78)
\end{aligned}$$

Here, Equation (76) follows from the fact that $\operatorname{sgn}(\psi(x)) \cdot f(x) > 0$ for all $x \in \{-1,1\}^n$, as ψ is a threshold degree dual witness for f. Equation (77) follows from Equations (74) and (75). Equation (78) follows from Equation (73) and the fact that $\|\psi\|_1 \leq 1$.

To prove Theorem 89, we essentially perform the above analysis in the matrix setting, whereby f is replaced by $F = f \circ g$, and ψ is replaced with η (the "approximate-rank dual witness" from the proof of Theorem 77).

Why this analysis overcomes the $2^{-\Theta(d)}$ loss in the error parameter of Theorem 77. Intuitively, if the dual witness for f is smooth, then the analysis above establishes a lower bound on the degree of ε-approximating polynomials p *even when* ε *can be as large as* $1 + 2^{-\Theta(d)}$ (see Equation (74)), so long as p has average margin at least 2^{-d}. And the assumption that p has this large an average margin is without loss of generality by Forster's lemma.

Put another way, the analysis above does not actually avoid the $2^{-\Theta(d)}$ loss in the error parameter ε incurred in Theorem 77, but it does render the loss innocuous. Because the smooth dual witness ψ lower bounds the degree of polynomials even if they make errors as large as $\varepsilon = 1 + 2^{-\Theta(d)}$ (so long as the average margin is large), a $2^{-\Theta(d)}$ degradation in the error parameter does *not* prevent the "degraded" error from being arbitrarily close to 1.

Digging deeper, we can pinpoint the reason that Theorem 77 experiences an additive loss of $2^{-\Theta(d)}$ in the error parameter, and see why our analysis here circumvents the issue if ψ is smooth. A threshold degree dual witness ψ for f rules out the existence of a degree-d sign-representation of f because ψ is uncorrelated with all degree d polynomials, but by virtue of agreeing in sign with f at all inputs, ψ is positively correlated with any sign-representation p of f. The reason for the additive 2^{-d} loss in the error parameter of Theorem 77 is that, when porting this style of analysis to the matrix-analytic setting, we cannot show that the "matrix dual witness" $\eta = \psi \star \phi \star \mu$ has *zero* correlation with rank-r matrices. Rather, we can only show that η has *very low* correlation with rank-r matrices (correlation roughly bounded by $2^{-d}/\sqrt{r}$). Fortunately, the smoothness condition on the dual witness ψ turns out to guarantee that η's correlation with any sign-representation of F with large average margin is not only positive, but is in fact noticeably so (roughly, at least $2^{-d/2}$ times the average margin, which in turn is at least $1/r$ by Forster's lemma). This is enough to conclude that any sign-representation for F must have large rank (at least $2^{\Omega(d)}$).

Proof of Theorem 89. The following lemma is implicit in work of Forster and was explicitly distilled by Razborov and Sherstov [RS10]. We omit the proof from this survey, directing the interested reader to [RS10, Appendix B].

Lemma 90. Let X, Y be finite sets and $M = [M_{x,y}]_{x \in X, y \in Y} \in \{-1,1\}^{|X| \times |Y|}$ a Boolean matrix. Let $r = \mathrm{rank}_{\pm}(M)$. Then there is a matrix R of rank r that sign-represents M, and moreover:

$$|R_{x,y}| \leq 1 \text{ for all } x \in X, y \in Y, \tag{79}$$

and

$$\|R\|_F = \sqrt{|X||Y|/r}. \tag{80}$$

Equations (79) and (80) guarantee that R not only sign-represents M, but does so with average margin at least $1/r$. Indeed, these equations guarantee that

$$\frac{1}{|X| \cdot |Y|} \sum_{x \in X, y \in Y} |R_{x,y}| \geq \frac{1}{|X| \cdot |Y|} \sum_{x \in X, y \in Y} R_{x,y}^2 = 1/r. \tag{81}$$

Let ψ be a dual witness for $\deg_{\pm}(f) \geq d$ satisfying $|\psi(x)| \geq 2^{d/2} \cdot 2^{-n}$ for all $x \in \{-1,1\}^n$. Let $\eta = \psi \star \phi \star \mu$ be the dual witness for $\deg_{\pm}(f \circ \mathsf{id} \circ \oplus_2) \geq d$ constructed in the proof of Theorem 77. Analogous to Equation (68), we have that:

$$\begin{aligned} \langle M_\eta, R \rangle &\leq \|M_\eta\| \cdot \|R\|_\Sigma \leq \|M_\eta\| \cdot \|R\|_F \cdot \sqrt{r} \leq \|M_\eta\| \cdot \sqrt{2^{6n}} \\ &\leq 2^{-3n} 2^{-d} 2^{3n} = 2^{-d}. \end{aligned} \tag{82}$$

On the other hand,

$$\langle M_\eta, R \rangle = \tag{83}$$

$$\sum_{x,y \in \{-1,1\}^{3n}} |(M_\eta)_{x,y}| \cdot |R_{x,y}| \tag{84}$$

$$\geq 2^{-d/2} 2^{-6n} \sum_{x,y \in \{-1,1\}^{3n}} |R_{x,y}| \tag{85}$$

$$\geq 2^{-d/2}/r. \tag{86}$$

Here, Equation (84) follows from the fact that $R_{x,y} \cdot (M_\eta)_{x,y} \geq 0$ for all $x, y \in \{-1,1\}^{3n}$. Equation (85) holds because the smoothness of ψ implies that $|\eta(x,y)| \geq 2^{-d} \cdot 2^{-6n}$ for all $x, y \in \{-1,1\}^{3n}$. Equation (86) follows from the large average margin of R (Equation (81)).

Combining Equation (86) with Equation (82), we conclude that $r \geq 2^{d/2}$.

□

10.5.1 Communication applications

Since the parity function $f = \oplus_n$ on n has threshold degree n, and this is witnessed by the dual polynomial $2^{-n} \cdot \oplus_n$, which is perfectly smooth, we obtain an explicit communication problem $f \circ g$ with *linear* $\mathbf{UPP}^{\mathrm{cc}}$ complexity.

Corollary 91. Let $f = \oplus_n$ and $F = f \circ g$ be the composition of f with the 6-bit gadget from Theorem 89. The $\mathrm{rank}_{\pm}(M_F) \geq 2^{\Omega(n)}$ and hence $\mathbf{UPP}^{\mathrm{cc}}(F) \geq \Omega(n)$.

In fact, as per the proof of Corollary 84, the above lower bound also holds for the inner product function, recovering Forster's breakthrough result [For02].

$\mathbf{UPP}^{\mathrm{cc}}$ lower bounds for $\mathbf{AC}^0$. The function $\oplus_n \circ g$ in Corollary 91 is not in AC^0. It was open for several decades whether there is an AC^0 function F with $\mathbf{UPP}^{\mathrm{cc}}(F) \geq n^{\Omega(1)}$. This question was originally posed by Babai, Frankl, and Simon [BFS86] in a different but equivalent form; specifically, they asked for a function F solvable by $\mathbf{PH}^{\mathrm{cc}}$ protocols of polylogarithmic cost but not by $\mathbf{UPP}^{\mathrm{cc}}$ protocols of polylogarithmic cost. Here, $\mathbf{PH}^{\mathrm{cc}}$ denotes the communication analog of the polynomial hierarchy.

As described in Section 5.3, the question was resolved by Razborov and Sherstov [RS10]. They established the *existence* of a smooth dual witness for the fact that Minsky-Papert CNF $\mathsf{AND}_{n^{1/3}} \circ \mathsf{OR}_{n^{2/3}}$ has threshold degree $\Omega(n^{1/3})$ (see Theorem 26).

We wish to cover an *explicit construction* of a smooth dual witness for an AC^0 function. While such a construction is known for the Minsky-Papert DNF [SW19], we present one (based on techniques in [BT19a]) that we feel is somewhat easier to understand. Unfortunately, our dual witness Λ falls (very) slightly short of providing the necessary smoothness to apply Theorem 89. But we are able to obtain a dual witness ζ for the AC^0 function $\mathsf{AND}_{n^{1/3}} \circ \mathsf{OR}_{n^{2/3}} \circ \oplus_{\log^2 n}$ that *is* smooth enough to yield a $\mathbf{UPP}^{\mathrm{cc}}$ lower bound. This detailed proof sketch is particularly technical and may be skipped with no loss of continuity in this survey.

Theorem 92. Let $f = \mathsf{AND}_{n^{1/3}} \circ \mathsf{OR}_{n^{2/3}} \circ \oplus_{\log^2 n}$, which is defined over $N = n \log^2 n$ variables. Then $\deg_{\pm}(f) \geq D$ for some $D = \Omega(n^{1/3} \log n)$. Moreover, there is a dual polynomial ζ for this fact, such that $|\zeta(x)| \geq 2^{-D} \cdot 2^{-N}$ for al $x \in \{-1,1\}^N$.

Proof. We will first construct a smooth dual witness Λ for the fact that $\deg_{\pm}(\mathsf{AND}_{n^{1/3}} \circ \mathsf{OR}_{n^{2/3}}) \geq d$ for some $d \geq \Omega(n^{1/3}/\log n)$. Unfortunately, Λ is not quite smooth enough for Theorem 89: it satisfies $|\Lambda(x)| \geq n^{-2d} \cdot 2^{-n}$ for all x, whereas Theorem 89 requires $|\Lambda(x)| \geq 2^{-d/2} 2^{-n}$. We rectify this issue by giving a dual witness ζ for f that is just as smooth as Λ, but ζ witnesses a slightly larger degree lower bound of $D = d \log^2 n$. Thus, ζ *is* smooth enough relative to the degree bound it witnesses to apply Theorem 89.

Smooth dual polynomial for the Minsky-Papert CNF. Let $m = n^{1/3}$. We are going to start with the dual witness constructed in Theorem 47 with $g = \mathsf{OR}_{n^{2/3}}$, and $F = \mathsf{AND}_m \circ g$. As reviewed below, the proof of this theorem constructs a dual polynomial, let's call it $\mu^{(m)}$, witnessing $\widetilde{\deg}_{1-8^{-m}}(F) \geq d_1$ where $d_1 = \widetilde{\mathrm{odeg}}_{7/8}(g)$. Unfortunately, $\mu^{(m)}$ is insufficient to establish our desired result for two reasons. First, it makes some sign-errors, i.e., there are some inputs x for which $\mathrm{sgn}(\mu^{(m)}(x)) \neq F(x)$. Second, $\mu^{(m)}$ is not smooth. We are going modify the witness so as to both eliminate the sign errors and render it smooth.

Recall that $\mu^{(m)}$ equals the dual block composition $\psi_m \star \phi$ (Definition 38). Here, ϕ is any dual witness for $\widetilde{\mathrm{odeg}}_{7/8}(\mathsf{OR}_{n^{2/3}}) \geq d_1$ where $d_1 = \Omega(n^{1/3})$, and $\psi_m \colon \{-1,1\}^m \to \mathbb{R}$ is such that: $\psi_m(\mathbf{1}_m) = 1/2$, $\psi_m(-\mathbf{1}_m) = -1/2$, and $\psi_m(x) = 0$ otherwise. Note that $|\phi(\mathbf{1}_{n^{2/3}})| \geq 7/16$ (see Fact 32). Hence,

$$|\mu^{(m)}(\mathbf{1}_{m \cdot n^{2/3}})| \geq \frac{1}{2} \cdot (7/8)^m. \tag{87}$$

For the remainder of the proof, set $m = n^{1/3}$. We are going to construct our smooth dual for F in a three-step process. First, we will use it to build a dual witness γ such that $|\gamma(x)| \geq 2^{-m}$ for all inputs x of Hamming weight at most w, where we choose $w = \Omega(m/\log m)$ such that

$$n^w \leq 2^{m/4}. \tag{88}$$

Note that this implies that $w \leq m/2$. Second, we are going to "zero out" any sign errors that it makes (without introducing any new ones). Third, we are going to render it smooth.

Step 1: Achieving largeness on inputs of small Hamming weight. Let $x^* = (x^*_1, \ldots, x^*_m) \in (\{-1,1\}^{n^{2/3}})^m$ be an input of Hamming weight at most $w \leq m/2$. We are going to construct a dual witness μ_{x^*} satisfying

$$\mu_{x^*}(x^*) > (7/8)^m, \tag{89}$$

$$\mu_{x^*}(x) \geq 0 \text{ for all } x \text{ with } |x| \leq m, \text{ and,} \tag{90}$$

$$\mu_{x^*} \text{ witnesses that } \widetilde{\deg}_{1-4^{-m}}(F) \geq d. \tag{91}$$

Before constructing μ_{x^*}, we explain how to use it to achieve our goal in Step 1. Let $M = \binom{n}{\leq w} \leq n^w$ denote the number of inputs in $\{-1,1\}^n$ of Hamming weight at most w. Our final dual witness in Step 1 will be

$$\gamma = \frac{1}{M} \sum_{x^*\colon |x^*| \leq w} \mu_{x^*}.$$

This dual polynomial itself witnesses that $\widetilde{\deg}_{1-4^{-m}}(F) \geq d$ (since it is an average of dual witnesses for this statement), and for all x^* of Hamming weight at most w, γ satisfies:

$$\gamma(x^*) \geq \frac{1}{M} \cdot \mu_{x^*}(x^*) \geq (1/M) \cdot (1/2) \cdot (7/8)^m.$$

By Equation (88), this last expression is at least $(1/2)^m$.

We now turn to constructing μ_{x^*}. Let S be the set of $i \in [m]$ such that $|x^*_i| > 0$, and let $\ell = m - |S|$ be the number of i such that $|x^*_i| = 0$. Since $|x^*_i| > 0$ for at most $w \leq m/2$ values of i, $\ell \geq m/2$. Consider the restriction of F obtained by fixing the bits in each block $i \in S$ to x^*_i. This restricted function equals $\mathsf{AND}_\ell \circ \mathsf{OR}_{m^2}$. That is, for $y \in \left(\{-1,1\}^{m^2}\right)^\ell$, let $y \cup x^*|_S$ denote the input $x = (x_1, \ldots, x_m)$ for which $x_i = x^*_i$ for all $i \in S$, and for which y is interpreted as specifying $\{x_j\colon j \notin S\}$. Then $F(y \cup x^*|_S) = (\mathsf{AND}_\ell \circ \mathsf{OR}_{m^2})(y)$.

Define $\mu_{x^*}\colon \{-1,1\}^{m^3} \to \mathbb{R}$ via

$$\mu_{x^*}(x) = \begin{cases} 0 \text{ if } x|_S \neq x^*|_S \\ \mu^{(\ell)}(x|_{\bar{S}}) \text{ otherwise.} \end{cases}.$$

Effectively, μ_{x^*} is a dual witness for F that treats x^* exactly the way $\mu^{(\ell)}$ treats input $\mathbf{1}_{\ell \cdot n^{2/3}}$. By Equation (87), this implies that $\mu_{x^*}(x^*) \geq \frac{1}{2}(7/8)^\ell$, i.e., Equation (89) above holds. Clearly, μ_{x^*} has the same pure high degree and ℓ_1-norm as $\mu^{(\ell)}$, and since $F(y \cup x^*|_S) = (\mathsf{AND}_\ell \circ \mathsf{OR}_{m^2})(y)$, μ_{x^*} has the same correlation with F as $\mu^{(\ell)}$ does with $\mathsf{AND}_\ell \circ \mathsf{OR}_{m^2}$, namely $1 - 8^{-\ell} \geq 1 - 4^{-m}$. That is, Equation (91) above holds. Finally, it can be checked that all inputs x on which $\mathrm{sgn}(\mu_{x^*}(x)) \neq F(x)$ have Hamming weight strictly greater than m, and also that $F(x) = 1$ for all inputs x of Hamming weight strictly less than m. This ensures that Equation (90) above holds.

Step 2: Correcting errors. The dual witness γ constructed in Step 1 suffers the following two issues: it makes some sign-errors, i.e., it does not witness $\deg_{\pm}(F) \geq d$, and it is not smooth. We now correct these issues. Our tool to accomplish this is the following object (whose existence can be interpreted as a dual formulation of Lemma 56, though we do not explain this interpretation as it will not be necessary for our proof).

Lemma 93. Let $w < n$ be any integer. For any input $x \in \{-1,1\}^n$ of Hamming weight greater than w, there is a function $\beta_x \colon \{-1,1\}^n \to \mathbb{R}$ such that (1) β has pure high degree at least w, (2) $\beta_x(x) = 1$, (3) $\beta_x(x') = 0$ for all x' with Hamming weight greater than w, and (4) $|\beta_x(x')| \leq \binom{n}{w}$ for all x'.

Proof. We will prove this for $x = -\mathbf{1}_n$ and assuming n is even; the proof for other inputs x and for odd n is similar. Let $p(t) = \prod_{i=w+1}^{n-1} \frac{t-i}{n-i}$, and define $\beta_{-\mathbf{1}_n}(x) = \oplus_n(x) \cdot p(|x|)$. Property (1) follows from the same analysis as Lemma 31. Properties (2)-(3) are simple calculations. Property (4) follows from observing that $|p(t)|$ is maximized at $t = 0$ amongst all $t \in [n]^*$, and that $p(0) = \frac{(n-1)!}{w!(n-(w+1))!} = \binom{n-1}{w} < \binom{n}{w}$. □

Let $E = \{x' \colon \operatorname{sgn}(\gamma(x')) \neq F(x')\}$. Apply Lemma 93 (with w set as in Step 1) to each input $x' \in E$ (we observed in Step 1 that all $x' \in E$ have Hamming weight at least $m > w$, so that Lemma 93 applies to each such x'). We next use $\beta_{x'}$ to "zero-out" γ at x' as follows. Define a new witness $\Gamma(x) \coloneqq \gamma(x) + \sum_{x' \in E} |\gamma(x')| \cdot \beta_{x'}(x)$. Since the pure high degree of the sum of two functions is at least the minimum of their individual pure high degrees, Γ has pure high degree at least $d \coloneqq \min(d_1, w) = w \geq \Omega(m/\log m)$. We now show that Γ has perfect sign-agreement with F. By design, $\Gamma(x) = 0$ for all $x \in E$, and $\Gamma(x) = \gamma(x)$ for all $|x| > w$. It remains to analyze $\Gamma(x)$ for $|x| \leq w$. Since $\sum_{x' \in E} |\gamma(x')| \leq 4^{-m}$, and $|\beta_{x'}(x)| \leq \binom{n}{w}$ for all $x \in \{-1,1\}^n$, we have that for any x with $|x| \leq w$,

$$\Gamma(x) \geq \gamma(x) - \sum_{x' \in E} |\gamma(x')| \cdot \beta_{x'}(x) \geq (1/2)^m - 4^{-m} \cdot \binom{n}{w}.$$

As per Equation (88), we chose $w = \Omega(m/\log m)$ to be small enough that $\binom{n}{w} \leq 2^{m/2}$, and hence the above expression is at least $(1/2)^{m-1}$.

Step 3: Ensuring smoothness. We have shown that Γ is a dual witness for $\deg_{\pm}(F) \geq \Omega(m/\log m)$, but Γ is not smooth. We render it smooth using the same technique above to "add mass" as necessary to each input in $\{-1,1\}^n$. Since Γ is already "big" on inputs of Hamming weight at most w, it suffices to add mass only to inputs of larger Hamming weight. To do so, we again use the object from Lemma 93.

Specifically, consider the dual witness

$$\tau(x) = \Gamma(x) + \sum_{x' \in \{-1,1\}^n \colon |x'| > w} (n^{-2w} \cdot 2^{-n} \cdot F(x)) \beta_{x'}(x).$$

Then $\operatorname{phd}(\tau) \geq w$, as τ is a sum of dual witnesses all of which have pure high degree at least w. A similar analysis to Step 2 shows that for all x with $|x| > w$, $\tau(x) \cdot F(x) > n^{-2w} \cdot 2^{-n} \beta_x(x) >$

$n^{-2w} \cdot 2^{-n}$. Also similar to Step 2, for x with $|x| \le w$, we have

$$\tau(x) \ge \Gamma(x) - \sum_{x' \in \{-1,1\}^n \colon |x'| > w} n^{-2w} 2^{-n} \cdot |\beta_{x'}(x)|$$
$$\ge (1/2)^{m-1} - n^{-2w} \cdot \binom{n}{w} \ge (1/2)^{m-1} - n^{-w} \ge (1/2)^{m-2}.$$

Hence, we have shown that τ witness $\deg_{\pm}(F) \ge \Omega(m/\log m)$, and moreover τ satisfies

$$|\tau(x)| \ge n^{-2w} \cdot 2^{-n} \text{ for all } x \in \{-1,1\}^n. \tag{92}$$

Final Step. Let $t = \log^2 n$, and let $\alpha = 2^{-t} \cdot \oplus_t$ denote the natural dual witness for the fact that $\deg_{\pm}(\oplus_t) \ge t$. Define $\zeta = \tau \star \alpha$. Then ζ witnesses the fact that $\deg_{\pm}(F \circ \oplus_t) \ge d \cdot t = \Omega(n^{1/3} \log n)$. Moreover, Equation (92) implies that $|\zeta(x)| \ge n^{-2w} \cdot 2^{-N}$ as desired, where recall from the statement of the theorem that $N = n \log^2 n$ is the number of variables over which $f = \mathsf{AND}_{n^{1/3}} \circ \mathsf{OR}_{n^{2/3}} \circ \oplus_{\log^2 n}$ is defined.

A detail that is suppressed in this proof sketch is that the ℓ_1-norms of the corrected dual witnesses Γ and τ are not exactly 1, so we need to normalize ζ to ensure this property. Fortunately, this does not ruin its smoothness because the calculations above imply that the total mass of the correction terms is $o(1)$.

□

10.5.2 $\mathsf{LTF} \circ \mathsf{MAJ}$ circuit lower bounds

We saw in Section 10.4.4 that a $\mathbf{PP}^{\mathrm{cc}}$ lower bound for F implies F cannot be computed by small $\mathsf{MAJ} \circ \mathsf{LTF}$ circuits. Here, we show that a $\mathbf{UPP}^{\mathrm{cc}}$ lower bound for F implies that F cannot be computed by small $\mathsf{LTF} \circ \mathsf{MAJ}$ circuits.

Theorem 94. Suppose $F(x, y)$ is computed by a $\mathsf{LTF} \circ \mathsf{MAJ}$ circuit $\mathcal{C}$ of size $s + 1 \ge n$. Then $\mathbf{UPP}^{\mathrm{cc}}(F) \le O(\log(s))$.

Proof. Write the output of the circuit as $\mathrm{sgn}(w_0 + \sum_{i=1}^{s} w_i \cdot \mathsf{MAJ}^i(x, y))$, where $\mathsf{MAJ}^i(x, y)$ denotes the output of the ith MAJ gate in $\mathcal{C}$ (here, by a majority gate, we mean a gate that takes as input a subset of the variables of (x, y), each possibiy negated, and outputs -1 if the majority of those inputs equal -1). By perturbing weights if necessary, we may assume without loss of generality that $w_0 + \sum_{i=1}^{s} w_i \mathsf{MAJ}^i(x, y) \ne 0$ for any input pair (x, y). In the $\mathbf{UPP}^{\mathrm{cc}}$ protocol for F, Alice randomly chooses an $i \in [s]$ with probability proportional to $|w_i|$ and sends i to Bob, which costs $O(\log s)$ bits. Alice and Bob then output $\mathrm{sgn}\left(w_i \cdot \mathsf{MAJ}^i(x, y)\right)$. Note that this costs $O(\log n)$ additional bits of communication. This is because it suffices for Alice to send to Bob the number of variables in x that are equal to -1 and connected by a wire to MAJ^i, as this enables Bob to determine the exact number of inputs to MAJ^i that are equal to -1.

Similar to the proof of Fact 71, this protocol outputs $F(x, y)$ with probability at least

$$\frac{1}{2}\left(1 + \frac{F(x, y) \cdot \sum_{i=1}^{s} w_i}{\sum_{i=1}^{s} |w_i|}\right) > 1/2.$$

□

Combining Theorem 94 with Theorem 92 and Theorem 89 reveals that an AC^0 function that cannot be computed by $\mathsf{LTF} \circ \mathsf{MAJ}$ circuits of size less than $2^{n^{1/3}}$.

Extending to $\mathsf{LTF} \circ \mathsf{LTF}$ circuits? Theorem 94 shows that every function F computable by a polynomial size $\mathsf{LTF} \circ \mathsf{MAJ}$ circuit has a $\mathbf{UPP}^{\mathrm{cc}}$ protocol of logarithmic cost. One may wonder whether the same is true for $\mathsf{LTF} \circ \mathsf{LTF}$ circuits. If so, one would resolve the notorious open problem of obtaining superpolynomial $\mathsf{LTF} \circ \mathsf{LTF}$ lower bounds for an explicit function. Unfortunately, this is not the case. Chattopadhyay and Mande [CM18] showed that there is a polynomial size $\mathsf{LTF} \circ \mathsf{LTF}$ circuit computing a function F over n bits with $\mathbf{UPP}^{\mathrm{cc}}(F) \geq \Omega(n^{1/4})$.

We briefly sketch their construction of F and a way one can prove that $\mathbf{UPP}^{\mathrm{cc}}(F) \geq \Omega(n^{1/4})$. Section 10.4.3 (see Footnote 39) mentioned a function called $\mathsf{ODDMAXBIT}$ (OMB for short) that is a linear threshold function, yet has large $\mathbf{PP}$ complexity. Since OMB itself has threshold degree 1, it obviously does not by itself have large $\mathbf{UPP}$ complexity. But they show that by composing OMB with AND and lifting with the two-bit $\oplus$ gadget, one does obtain such a function, namely $F = \mathsf{OMB}_{n^{3/4}} \circ \mathsf{AND}_{n^{1/4}} \circ \oplus_2$.

Let us explain why this function is computable by small $\mathsf{LTF} \circ \mathsf{LTF}$ circuits. As previously mentioned, OMB has threshold degree 1 and hence is computable by a single LTF gate. Meanwhile, letting $N = n^{1/4}$, observe that $(\mathsf{AND}_N \circ \oplus_2)(x, y)$ evaluates to -1 if and only if $x_1 \neq y_1$, $x_2 \neq y_2$, $\ldots$, $x_N \neq y_N$. This is equivalent to requiring that

$$\sum_{i=1}^{N} 3^{i-1} \cdot (x_i + y_i) = 0.$$

Let $W(x, y) = \sum_{i=1}^{N} 3^{i-1} \cdot (x_i + y_i)$, and consider the following two functions, both of which are LTFs:

$$h(x, y) = \begin{cases} -1 \text{ if } W(x, y) \geq 0 \\ 1 \text{ otherwise} \end{cases}$$

and

$$g(x, y) = \begin{cases} -1 \text{ if } W(x, y) \geq 1/2 \\ 1 \text{ otherwise.} \end{cases}$$

Then $1 - h(x, y) + g(x, y)$ equals -1 if $W(x, y) = 0$ and otherwise equals 1. Hence, an $\mathsf{LTF} \circ \mathsf{LTF}$ circuit for $\mathsf{OMB}_{N^3} \circ \mathsf{AND}_N \circ \oplus_2$ consists of the LTF gate for OMB_{N^3}, but with each input to it replaced by $g(x_i, y_i) - h(x_i, y_i)$, where (x_i, y_i) is the input to the ith copy of $\mathsf{AND}_N \circ \oplus_2$. Since g and h are both LTFs, this is an $\mathsf{LTF} \circ \mathsf{LTF}$ circuit.

We do not prove the $\mathbf{UPP}^{\mathrm{cc}}$ lower bound for $F = \mathsf{OMB}_{n^{3/4}} \circ \mathsf{AND}_{n^{1/4}} \circ \oplus_2$, but to give some idea of how the proof goes, we sketch how to prove the $\mathbf{UPP}^{\mathrm{cc}}$ lower bound for the slightly more complicated function $G = \mathsf{OMB}_{n^{3/4}} \circ \mathsf{AND}_{n^{1/4}} \circ g$ where $g = \mathsf{idx} \circ \oplus_2$ is the gadget from Theorems 77 and 89. Unfortunately, owing to the inclusion of the idx function in the gadget g, G is *not* (known to be) computed by an $\mathsf{LTF} \circ \mathsf{LTF}$ circuit, so the analysis presented here is not sufficient to recover the result of Chattopadhyay and Mande [CM18].

Let $N = n^{1/4}$. To prove that $\mathbf{UPP}^{\mathrm{cc}}(G) \geq \Omega(N)$, the proof of Theorem 89 implies that it suffices to give a smooth dual witness for the fact that, for every $\varepsilon < 1$, the ε-approximate degree of $\mathsf{OMB}_{N^3} \circ \mathsf{AND}_N$ is at least $2^{\Omega(N)}$. In more detail, for some $d \geq \Omega(N)$, it is enough to give a function $\psi \colon \{-1, 1\}^n \to \mathbb{R}$ such that that $\|\psi\|_1 = 1$, ψ has pure high degree at least $d \geq \Omega(N)$,

$$\langle \psi, \mathsf{OMB}_{N^3} \circ \mathsf{AND}_N \rangle = 1, \tag{93}$$

and

$$|\psi(x)| \geq 2^{-d/2} \cdot 2^{-n}. \tag{94}$$

Actually, the same analysis can be used to conclude that $\mathbf{UPP}^{\mathrm{cc}}(F) \geq \Omega(n^{1/4})$ even if the right hand side of Equation (93) is $1 - 2^{-d}$ rather than 1, and even if Equation (94) fails to hold for a 2^{-d} fraction of inputs $x \in \{-1,1\}^n$.

Here is how to construct a dual witness ψ satisfying these relaxed properties. First, Thaler [Tha16] gave a dual witness ϕ for the fact that $\mathsf{OMB}_{n^{3/4}}$ satisfies $\widetilde{\deg}_{\varepsilon}(\mathsf{OMB}_{n^{3/4}}) \geq d$ where $d \geq \Omega(n^{1/4})$ and $\varepsilon = 1 - 2^{-d}$. Moreover, his dual witness satisfies

$$|\phi(\mathbf{1}_{n^{3/4}})| \geq 2^{-d/2}. \tag{95}$$

Let $\gamma\colon \{-1,1\}^N \to \mathbb{R}$ be the following function:

$$\gamma(x) = \begin{cases} -1/2 \text{ if } x = \mathbf{1}_N \\ 1/(2^N - 1) \text{ otherwise.} \end{cases}$$

It is easy to check that γ is balanced (hence has pure high degree at least 1) and is perfectly correlated with AND; hence, it is a dual witness for the fact that $\deg_{\pm}(\mathsf{AND}_N) \geq 1$. The desired dual witness for $\mathsf{OMB}_{N^3} \circ \mathsf{AND}_N$ is simply the dual block composition $\phi \star \gamma$.

We now sketch why $\phi \star \gamma$ satisfies the requisite properties. First, by Lemma 40, $\phi \star \gamma$ has ℓ_1-norm 1, and by Lemma 39, it has pure high degree at least $\mathrm{phd}(\phi) \cdot \mathrm{phd}(\gamma) \geq d \cdot 1 = d$. Next, observe that by Theorem 43,

$$\langle \phi \star \gamma, \mathsf{OMB}_{N^3} \circ \mathsf{AND}_N \rangle = \langle \phi, \mathsf{OMB}_{N^3} \rangle \geq 1 - 2^{-d},$$

yielding the required relaxed version of Equation (93). Finally, Equation (95) implies that for any input $x = (x_1, \ldots, x_{N^3}) \in \left(\{-1,1\}^N\right)^{N^3}$ with each $x_i \in \mathsf{AND}^{-1}(+1)$, $|(\phi \star \gamma)(x)| \geq 2^{-n} \cdot 2^{-d/2}$. Since only a $1 - 2^{-N}$ fraction of inputs in $\{-1,1\}^N$ are in $\mathsf{AND}^{-1}(+1)$, at least a $1 - \frac{N^3}{2^{-N}}$ fraction on inputs x in $\{-1,1\}^n$ satisfy $|(\phi \star \gamma)(x)| \geq 2^{-n} \cdot 2^{-d/2}$. This yields the required relaxed version of Equation (94) so long as $N \geq 2d$.

10.5.3 Open problems on threshold degree and sign-rank

It is open whether Theorem 89 holds *without* the smoothness condition $|\psi(x)| \geq 2^{-d/2} \cdot 2^{-n}$ on the threshold degree dual witness ψ. This may seem like a low-level technical question, but conceptually it is asking whether there is a generic query-to-communication-lifting theorem for **UPP**. That is, the threshold degree of a function $f\colon \{-1,1\}^n \to \{-1,1\}$ characterizes its **UPP** query complexity (Fact 71), while the sign-rank of a function $F\colon \{-1,1\}^n \times \{-1,1\}^n \to \{-1,1\}$ characterizes its **UPP** communication complexity (Fact 75). If the requirement that $|\psi(x)| \geq 2^{-d/2} \cdot 2^{-n}$ could be removed from Theorem 89, it would generically translate **UPP** query lower bounds into **UPP** communication lower bounds, rather than requiring "extra properties" of the dual object that was used to prove the query lower bound. It would also drastically simplify known sign-rank lower bounds (e.g., Theorem 92).

Open Problem 95. Identify a "gadget" function g on a constant number of bits, such that for any function $f\colon \{-1,1\}^n \to \{-1,1\}$ with $\deg_{\pm}(f) \geq d$, the following holds for the composed function $F = f \circ g$: $\mathrm{rank}_{\pm}(M_F) \geq 2^d$.

A potentially easier open question is the following. It is known that **UPP** query complexity is not closed under intersection, i.e., there is a function h with threshold degree 1 such that $H(x,y) = h(x) \wedge h(y)$ has threshold degree $\Omega(n)$ [She13b, She13c, She21]. It is open whether a similar result holds for **UPP** communication complexity. A generic lifting theorem for **UPP** would resolve this question, but so would giving a dual witness ψ for the large threshold degree of $H(x,y)$ such that ψ has the requisite smoothness properties to invoke Theorem 89. The best-known result in this direction is that there exists a function F with **UPP** communication complexity $O(\log n)$, such that computing the AND of two copies of F evaluated over disjoint inputs requires **UPP** communication complexity $\Omega(\log^2 n)$ [BMT21].

Open Problem 96. Is the class of problems with polylogarithmic **UPP** communication complexity closed under intersection?

10.6 Extensions to multiparty communication complexity

A long line of work has shown that approximate degree lower bounds imply not only two-party randomized and quantum communication lower bounds as covered in this survey, but also state-of-the-art multi-party number-on-forehead lower bounds. These results have their own set of applications in circuit complexity and elsewhere. The interested reader is directed to [LS09b, Section 8.3] for a survey on some early results in this line of work, as well as more recent papers [She14, She18c].

11 Assorted Applications

11.1 Secret Sharing Schemes

Suppose a party, called the *dealer*, wishes to distribute a secret bit $b \in \{-1,1\}$ amongst n parties, in the following sense. The deal will hand the ith party a bit x_i, and it is required that all n parties together can reconstruct b by applying some known "reconstruction" function f_n to their shares, while no "small" coalition of parties (say at most d parties) can guess b with any advantage over random guessing. This is called a secret-sharing scheme (for a 1-bit secret), and is a fundamental object of study in cryptography.

As observed by [BIVW16], a dual polynomial ψ for $\widetilde{\deg}_\varepsilon(f) \geq d$ precisely yields such a secret sharing scheme. Recall (Section 6) that ψ can be decomposed into two distributions $\psi_{-1} = 2\max\{-\psi(x), 0\}$ and $\psi_{+1} = 2\max\{\psi(x), 0\}$. The dealer simply draws $x \sim \psi_b$, and distributes x_i to part i. To reconstruct the secret, the parties output $f(x)$. The fact that $\langle f, \psi\rangle > \varepsilon$ implies that if the secret bit b itself is chosen at random, then the probability of successful reconstruction is:

$$\begin{aligned}\frac{1}{2}\sum_{b\in\{-1,1\}} \Pr_{x\sim\psi_b}[f(x)=b] &= \frac{1}{2}\sum_{b\in\{-1,1\}}\sum_{x\in\{-1,1\}^n:\ \mathrm{sgn}(\psi(x))=b} 2|\psi(x)|\cdot\frac{1+\mathrm{sgn}(\psi(x))f(x)}{2} \\ &= \sum_{x\in\{-1,1\}^n}|\psi(x)|\cdot\frac{1+\mathrm{sgn}(\psi(x))f(x)}{2} = \frac{1}{2}+\frac{1}{2}\sum_{x\in\{-1,1\}^n}\psi(x)f(x) = \frac{1+\langle f,\psi\rangle}{2} \\ &> \frac{1+\varepsilon}{2}.\end{aligned}$$

In particular, if ψ is a dual polynomial for $\deg_\pm(f) \geq 1$, then reconstruction is successful with probability 1 (this is referred to as perfect reconstruction). Meanwhile, the fact that ψ has pure

high degree at least d means that no coalition of size less than d can achieve *any* advantage over random guessing at reconstructing the secret b. Indeed, any function of at most d bits of x is obviously a degree-d polynomial p in x, and hence the calculation above shows that

$$\frac{1}{2}\sum_{b\in\{-1,1\}} \Pr_{x\sim\psi_b}[p(x)=b] = \frac{1+\langle p,\psi\rangle}{2} = 1/2.$$

Cryptographers are often interested in ensuring that the reconstruction function is "simple", e.g., computable by a constant-depth circuit. The fact that, for any constant $\delta > 0$, we know an AC^0 function with threshold degree $\Omega(n^{1-\delta})$ (see Section 8.3) means that there is a secret sharing scheme with perfect reconstruction, in which the reconstruction function is in AC^0, and no coalition of fewer than $\Omega(n^{1-\delta})$ parties gains any advantage over random guessing in reconstructing the secret.

Several works [CIL17, BMTW19, BW17, BDF+22] have studied variants of the above, such as allowing coalitions of size d to have a small but non-zero advantage over random guessing (which is related to ε-approximate weight (Section 10.4.1)), reconstruction by even simpler functions such as OR (which yields something called a *visual* secret sharing scheme), and the simplicity of sampling from the distribution ψ_b.

11.2 Learning Algorithms

PAC Learning. Valiant's *Probably Approximately Correct* (PAC) model [Val84] is intended to capture the task of supervised learning. In this model, there is some target function $f\colon \{-1,1\}^n \to \{-1,1\}$ that is unknown to the learning algorithm, but is assumed to come from some class of functions $\mathcal{C}$. Here, $\mathcal{C}$ is referred to as a *concept class*. The learning algorithm is given access to *labeled training data*. This means the learner is fed some number m of labeled examples $(x, f(x))$—here, x is the example, and $f(x)$ is the label. The number of labeled examples m consumed by the algorithm is referred to as the *sample complexity* of the learner.

Each example x is assumed to be drawn independently from some fixed distribution $\mathcal{D}$ over $\{-1,1\}^n$. We will be interested in *distribution-independent* PAC learning, which means that $\mathcal{D}$ is unknown to the learning algorithm and could be any distribution over $\{-1,1\}^n$. The goal is for the learning algorithm to manage to predict f's labels on as-yet-unseen examples that are drawn from the same distribution as the training data. This means that the learner must output a *hypothesis* h, which is a function mapping $\{-1,1\}^n \to \{-1,1\}$.[41] The hypothesis h should be an *ε-approximately correct* predictor for f under distribution $\mathcal{D}$. This means that $\Pr_{x\sim\mathcal{D}}[h(x) \neq f(x)] \leq \varepsilon$ for some desired accuracy parameter ε. The probability $\Pr_{x\sim\mathcal{D}}[h(x) \neq f(x)]$ is referred to as the *generalization error* of h, or sometimes as the *loss* of h (under the so-called *zero-one loss function*).[42]

The learning algorithm is "probably approximately correct" with parameters ε and δ, if, with probability at least $1-\delta$ over the random generation of labeled training data (and any internal randomness of the learning algorithm), the algorithm outputs an ε-approximately correct hypothesis.

[41] For the learning algorithm to be considered computationally efficient, h should have a polynomial size representation, and it should be possible to efficiently evaluate h at any desired input x.

[42] When learning real-valued rather than Boolean-valued functions, other loss functions make sense and often lead to more tractable learning problems. These include ℓ_2-loss, ℓ_1-loss, hinge loss, etc. In this survey, we only consider learning Boolean-valued functions under zero-one loss.

Prominent examples of concept classes studied in the learning theory literature include *disjunctions* and *conjunctions*, in which $\mathcal{C}$ consists of all functions of the form OR or AND applied to a subset of input variables or their negations, DNF formulas of size polynomial in n, and constant-depth circuits of size polynomial in n (i.e., AC^0).

As we will see shortly, another prominent concept class is the set of all *halfspaces*. This concept class is both interesting in its own right, and useful for "capturing" other concept classes, including some of those mentioned above.

Agnostic learning. The PAC learning model described above makes the often-unrealistic assumption that the target function f resides in the concept class $\mathcal{C}$. This is sometimes referred to as the "realizable case", referring to the assumption there is some concept in $\mathcal{C}$ that "fully realizes" f by agreeing with f at *all* inputs $x \in \{-1,1\}^n$.

In many realistic scenarios, $\mathcal{C}$ is merely a good rather than perfect description of structure within f. In this situation, the optimal concept c^* from $\mathcal{C}$ will itself fail to fit f perfectly, i.e., will have non-zero loss. This means that $\Pr_{x\sim\mathcal{D}}[c^*(x) \neq f(x)]$ is small but not zero. Let us refer to this probability as opt. In the *agnostic learning* model [KSS94, Hau92], the goal is to output a hypothesis h such that $\Pr_{x\sim\mathcal{D}}[h(x) \neq f(x)] \leq \mathrm{opt} + \varepsilon$. In words, the algorithm should output a hypothesis h that fits f nearly as well as the optimal concept from $\mathcal{C}$.

Another way to view the agnostic learning setting is to think of it as equivalent to the realizable setting, but where a small fraction (namely, opt) of adversarially training examples may be corrupted. In general, agnostic learning of any given concept class is a much more challenging task than (realizable) PAC learning, because the latter does not require learning in the presence of *any* noise in the training data, much less adversarially chosen classification noise.

Bounding sample complexity via VC dimension. A common approach to learning algorithm design is for the algorithm to find a hypothesis h that "fits" the training data. This means that the *training error*, i.e., the fraction of training inputs x such that $h(x) \neq f(x)$, is small. The algorithm that identifies a hypothesis h from a class of functions $\mathcal{H}$ that *best* fits the training data is typically referred to in machine learning as *empirical risk minimization* (ERM).

There is, however, the risk of "overfitting" to the training data. This means the algorithm finds an h that fits the training data well, but h has poor generalization error. This means that, although h describes f well on the training data, h is not actually an accurate predicator under the "true" data distribution $\mathcal{D}$. Intuitively, this can happen if h "hones in on" artifacts in the training data, i.e., properties of the specific sample used for training that are not representative of the true data distribution $\mathcal{D}$.

Suppose that the learning algorithm outputs a hypothesis h that is a member of some *hypothesis class* $\mathcal{H}$. It turns out that overfitting to training data is unlikely to occur if $\mathcal{H}$ is a "simple" or "not-very-expressive" class of functions.

One way of formalizing this is via so-called *Vapnik-Chervonenkis* dimension (VC-dimension). The definition of VC dimension is outside of the scope of this survey, but the salient point is the following. If a hypothesis class $\mathcal{H}$ has VC dimension at most D, then with sample complexity polynomial in D, $\log(1/\delta)$, and $1/\varepsilon$, the generalization error and test error of all hypotheses differ by at most ε with probability at least $1-\delta$ over the random choice of training data. This means that, if a learning algorithm outputs a hypothesis h that fits the training data well, then h will also have good generalization error.

Halfspace learners. Recall that halfspaces (also called linear threshold functions, see Section 10.4.4) are functions of the form $f(x_1, \ldots, x_n) = \text{sgn}(w_0 + \sum_{i=1}^n w_i x_i)$ for some weights $w_0, \ldots, w_n \in \mathbb{R}$. It is known that the VC-dimension of halfspaces over n variables is $n+1$.

A set of labeled training examples is said to be *linearly-separable* if there is some halfspace f such that each training example x is assigned label $f(x)$. There are many algorithms known that, given any set of linearly-separable examples, find a halfspace that correctly labels them. This can be done, for example, via any linear programming algorithm. But much simpler algorithms are known. These include the so-called *Perceptron* [Ros61] and *Winnow* [Lit88] algorithms.

Both Perceptron and Winnow find *some* weight vector that is consistent with the training data. In contrast, so-called *support vector machines* (SVM) find a particular weight vector, namely the one that maximizes the *margin*, which roughly corresponds to the closest to zero that $w_0 + \sum_{i=1}^n w_i x_i$ gets across all training examples x (the notion of margin also arose in Section 10.5).

Since the VC dimension of halfspaces is just $n+1$, to PAC-learn the concept class of all halfspaces over n variables, it suffices to take a sample of size $\text{poly}(n)$ and run any of the above algorithms to identify a halfspace consistent with the sample. In fact, there are halfspace-learning algorithms that are known that are robust to *random classification noise* in the training data. This refers to a variant of the PAC learning setting in which each training example has its label flipped with some small probability. Note that it is not known how to agnostically learn halfspaces in sub-exponential time—halfspace learning algorithms are not known to be robust to *adversarial* classification noise.

Learning polynomial threshold functions. Given a labeled example $x = (x_1, \ldots, x_n) \in \{-1, 1\}^n$, one can think of each x_i as a (binary) *feature*. For example, if attempting to classify an email as spam or not spam, each coordinate of x may indicate certain properties of the email such as "does it contain the word 'bank'?" or "does it contain spelling errors?" or "has the sender previously corresponded with the recipient?"

Unfortunately, many natural functions f are not halfspaces—there is no weight vector $w = (w_0, \ldots, w_n)$ such that $f(x) = \text{sgn}\,(w_0 + \sum_{i=1}^n w_i x_i)$ for all $x \in \{-1, 1\}^n$. In such situations, one can imagine taking such a feature-vector x and expanding it into a larger vector, in which each coordinate contains a "derived feature" corresponding to a "combination" of the n "basic" features $x_1, \ldots, x_n$. Then even if f is not itself a halfspace, it may be one in the higher-dimensional space of "expanded feature vectors".

To make this concrete, recall that a degree-d polynomial threshold function f is the sign of a polynomial $p\colon \{-1, 1\}^n \to \mathbb{R}$ of total degree at most d. Any PTF f can be thought of as a halfspace over an expanded feature space of size $\binom{n}{\leq d}$, whereby each "expanded feature" is a product of at most d features in x.

The above reframing of degree-d PTFs as halfspaces over expanded feature vectors immediately yields an algorithm that runs in time $n^{O(d)}$ for PAC learning degree-d PTFs. The algorithm samples a training set of size $n^{O(d)}$, and for each example x in the training set, it expands x into the size-$\binom{n}{\leq d}$ vector obtained by evaluating all monomials of degree at most d at x. One then applies a halfspace-learning algorithm to the resulting sample of expanded vectors. The total runtime of the algorithm—to expand each example in the sample and then apply the halfspace-learning algorithm to the expanded samples—is $n^{O(d)}$.[43]

[43] If the halfspace-learner used is a support-vector machine, the above algorithm roughly corresponds to running an SVM using the so-called polynomial kernel. See [KKMS08, Section 3.2] for additional discussion.

Learning DNFs via PTFs. A famous challenge problem posed in Valiant's seminal paper introducing the notion of PAC learning is to learn polynomial size DNF formulas in polynomial time. Unfortunately, we are very far from achieving this goal. The best-known algorithm currently runs in time exponential in $n^{1/3}$.

It is easy to see via the Chebyshev-polynomial-based technique of Lemma 8 that the threshold degree of any DNF or CNF of size s is at most $O(\sqrt{n \log s})$. This immediately yields an DNF learning algorithm that runs in time exponential in $\sqrt{n \log s}$. Klivans and Servedio [KS04] improved the above degree bound from $\tilde{O}(n^{1/2})$ to $\tilde{O}(n^{1/3})$, and the resulting algorithm remains the fastest known today.

Limitations on learning algorithms from threshold degree of sign-rank lower bounds. The fact that Minsky and Papert's DNF has threshold degree $\Omega(n^{1/3})$ (Theorem 26) means that Klivans and Servedio's threshold degree upper bound for DNFs cannot be improved. One may wonder, though, whether one can improve on their algorithm by applying a halfspace learning algorithm to a *different* set of derived features, one that does not consist of all monomials of degree at most d. Are there any "derived feature sets" of size smaller than $\binom{n}{\leq d}$ over which all DNFs are halfspaces?

It turns out that degree-at-most-d monomials are in fact the optimal feature set for learning DNFs via the above halfspace-based approach. As we now explain, this follows from the $\exp(\Omega(n^{1/3}))$ sign-rank lower bound of Razoborov and Sherstov [RS10] (Section 10.5.1)for the composition of the Minsky-Papert DNF with a constant-sized gadget.

Specifically, Razborov and Sherstov construct a binary matrix M such that each row of M is the evaluation table of a DNF defined over n variables (more specifically, each row is the evaluation table of the Minsky-Papert DNF applied to a subset of input variables or their negations) and such that M has sign-rank $r \geq 2^{\Omega(n^{1/3})}$.

Let $\mathcal{C}$ denote the class of polynomial size DNFs over n-bit inputs. Let M be the matrix whose rows are indexed by DNFs $f \in \mathcal{C}$ and whose columns are indexed by inputs $x \in \{-1,1\}^n$, with $M_{f,x} = f(x)$.

Now suppose that $\mathcal{F}$ is a set of "feature functions" $\phi\colon \{-1,1\}^n \to \{-1,1\}$ such that every $f \in \mathcal{C}$ can be expressed as a halfspace over the features. In other words, letting $s+1 = |\mathcal{F}|$ and $\phi_0, \phi_1, \ldots, \phi_s$ be an enumeration of the feature functions in $\mathcal{F}$, with $\phi_0(x) = 1$ for all x, suppose that for each DNF f, there exist weights $w_0, \ldots, w_s \in \mathbb{R}$ such that

$$f(x) = \operatorname{sgn}\left(\sum_{i=0}^{s} w_i \phi_i(x)\right). \tag{96}$$

This means that the sign-rank of M is at most s. To see this, let A be the $|\mathcal{C}| \times (s+1)$ matrix with rows indexed by DNFs f in $\mathcal{C}$ and columns indexed by feature functions $\phi_i \in \mathcal{F}$, with A_{f,ϕ_i} equal to the coefficient w_i of ϕ_i in Equation (96). Let B be the $(s+1) \times 2^n$ matrix, with rows indexed by $\phi_i \in \mathcal{F}$ and columns indexed by $x \in \{-1,1\}^n$, with $B_{\phi_i,x} = \phi_i(x)$. Then Equation (96) implies that M equals the entry-wise sign of the product matrix $A \cdot B$. This proves that the sign-rank of M is at most $s+1$. Razborov and Sherstov's result then implies that $s \geq \exp(\Omega(n^{1/3}))$ as claimed.

Agnostic Learning via Approximate Degree Upper Bounds. We have seen that if every function f in a concept class $\mathcal{C}$ has threshold degree at most d, then $\mathcal{C}$ can be PAC-learned in

time $n^{O(d)}$. However, threshold degree upper bounds are not known to yield learning algorithms in the more challenging agnostic setting. The fundamental challenge is that the empirical risk minimization problem is intractable when the hypothesis class consists of all halfspaces. That is, while there are efficient procedures for finding a halfspace that *perfectly* fits the training data under the assumption that such a halfspace exists, without this assumption there is no analogous procedure to find the best-fitting halfspace (and in fact the problem is known to be NP-hard [GR09]).

To obtain efficient agnostic learners, we turn to approximate degree upper bounds. Suppose that every $c \in \mathcal{C}$ is approximated to error $\varepsilon/4$ by a polynomial of degree at most d. Then $\mathcal{C}$ can be agnostically learned in time $n^{O(d)}$ using the so-called ℓ_1-*polynomial regression* algorithm of Kalai, Klivans, Mansour, and Servedio (KKMS) [KKMS08].

Conceptually, the algorithm of KKMS solves a *convex relaxation* of the empirical risk minimization problem for degree-d polynomial threshold functions. Whereas finding the degree-d polynomial p that minimizes the *zero-one* loss of $\operatorname{sgn}(p(x))$ is intractable, the algorithm of KKMS instead roughly finds the polynomial p minimizing the ℓ_1-loss (see Equation (97) below). This minimization problem can be written as a linear program and hence can be solved in polynomial time. However, in order to guarantee that an optimal solution to the linear program actually yields an accurate hypothesis, we will need to require that every function c in the concept class $\mathcal{C}$ being learned have approximate degree at most d (instead of merely requiring threshold degree at most d as in the realizable case).

In more detail, the algorithm first draws a sample of labeled training data of size $n^{O(d)}$. Then, using linear programming, the algorithm finds the degree-d polynomial p that best fits the data under the ℓ_1 loss. This means that p minimizes

$$\sum_{(x,f(x)) \text{ in training data}} |p(x) - f(x)| \tag{97}$$

amongst all degree-d polynomials, where recall that m is the number of training examples consumed by the algorithm. Finally, the algorithm chooses a threshold $t \in [-1, 1]$ in a manner specified momentarily, and outputs the hypothesis $h(x) = \operatorname{sgn}(p(x) - t)$. The threshold t is chosen to maximize the number of accurate predictions the algorithm makes on the training data, i.e., to minimize the number of training examples x such that $h(x) \neq f(x)$.

One may wish to think of $p(x)$ as a kind of weighted prediction for $f(x)$. If $p(x) \geq 1$ or $p(x) \leq -1$, then p is indicating, with high confidence, that $f(x) = 1$ or $f(x) = -1$ (though $p(x)$ may make incorrect predictions on a small fraction of inputs x under distribution $\mathcal{D}$, even when it indicates confidence therein). If $p(x) \in (-1, 1)$, then p is "indicating some uncertainty" regarding its prediction for $f(x)$. The threshold t yields a "cutoff" to turn uncertain predictions into hard predictions.

Sketch of the accuracy analysis for the algorithm. Following the analysis of [KKMS08], we first show that for the hypothesis h output by the algorithm, the number of training examples x such that $h(x) \neq f(x)$ is at most $1/2$ the ℓ_1-error in Expression (97). To see this, observe that $h(x) \neq f(x)$ only if the threshold t "separates" $p(x)$ and $f(x)$, i.e., only if $p(x) < t < f(x)$ or $f(x) < t < p(x)$. If t were chosen uniformly at random from the interval $[-1, 1]$ (which has length 2) the probability that t splits $p(x)$ and $f(x)$ is at most $|p(x) - f(x)|/2$. Hence, if t were chosen uniformly at random from $[-1, 1]$, the expected number of errors of h on the training set would

be at most one half of Expression (97). This implies the *existence* of a threshold t achieving this training error. Since t is chosen to minimize the training error, the selected threshold does at least as well as this expectation.

Second, we show that the fact that there exists a degree-d $\varepsilon/4$-approximating polynomial for every function in $\mathcal{C}$ implies that, in expectation over the random sample drawn from $\mathcal{D}$, Expression (97) is at most $m \cdot (2\text{opt} + \varepsilon/4)$ for the polynomial p selected by the algorithm. To see this, let $c \in \mathcal{C}$ be the optimal classifier for the target function f, i.e., c minimizes $\Pr_{x \sim S}[c(x) \neq f(x)]$ amongst all concepts in $\mathcal{C}$. Let p^* be an $\varepsilon/4$-approximating polynomial for c. Then

$$\sum_{(x,f(x)) \text{ in training data}} |p^*(x) - f(x)| \leq \sum_{(x,f(x)) \text{ in training data}} |p^*(x) - c(x)| + |c(x) - f(x)|.$$

Because p^* is an $\varepsilon/4$-approximation for c, $|p^*(x) - c(x)| \leq \varepsilon/4$ for all $x \in \{-1,1\}^n$. Meanwhile,

$$\mathbb{E}\left[\frac{1}{m} \sum_{(x,f(x)) \text{ in training set}} |c(x) - f(x)|\right] = 2\text{opt},$$

where the expectation is over the randomly sampled training data. This follows by definition of opt and the fact that if $c(x) = f(x)$ then $|c(x) - f(x)| = 0$, while if $c(x) \neq f(x)$, then $|c(x) - f(x)| = 2$. So the expected value of Expression (97) with $p = p^*$ is at most $m \cdot (2\text{opt} + \varepsilon/4)$. Since the algorithm chooses the polynomial p that minimizes Expression (97) (and hence the p selected by the algorithm always does at least as well as p^* in minimizing Expression (97)), the expected value of Equation (97) is in turn at most $m \cdot (2\text{opt} + \varepsilon/4)$ as claimed.

Combined with the first step of the analysis, we conclude that the *expected* number of errors that the algorithm's chosen hypothesis h makes on the training set is at most $m\,(\text{opt} + \varepsilon/8)$ (again, the expectation here is taken over the randomly sampled training data). Standard probabilistic analyses then imply that with noticeable probability, specifically at least $\varepsilon/4$, h achieves training error at most $\text{opt} + \varepsilon/2$, and in this event, VC theory implies that h achieves generalization error at most $\text{opt} + \varepsilon$ as desired.

One still needs to reduce the failure probability from $1 - \varepsilon/4$, to δ, where by failure we mean that the algorithm outputs an h with generalization error more than $\text{opt} + \varepsilon$. To achieve this, one can run the above algorithm $O(\log(1/\delta)/\varepsilon)$ times independently, to ensure that with high probability, at least one of the runs indeed outputs an h with the desired training error of at most $\text{opt} + \varepsilon/2$.

Example applications. Perhaps the most prominent example application of the above agnostic learning result is to disjunctions and conjunctions (OR or AND applied to a subset of features or their negations). Since OR and AND have ε-approximate degree $\Theta(\sqrt{n \log(1/\varepsilon)})$, the above yields an agnostic learning algorithm that runs in time $n^{O\left(\sqrt{n \log(1/\varepsilon)}\right)}$. As another prominent example, it is known that all De Morgan formulas of size s have approximate degree $O(\sqrt{s})$ [Rei11], which yields an $n^{O(\sqrt{s})}$-time agnostic learning algorithm for this class of functions for constant $\varepsilon > 0$.

These are the fastest known agnostic learning algorithms known for the above concept classes. Moreover, the approximate-rank lower bounds covered in Section 10.4 implies barriers to obtaining improved algorithms via ℓ_1-regression [KS07]. These barriers are analogous to the sign-rank-derived barriers discussed earlier for improving Klivans and Servedio's $2^{\tilde{O}(n^{1/3})}$-time algorithm for PAC learning DNFs.

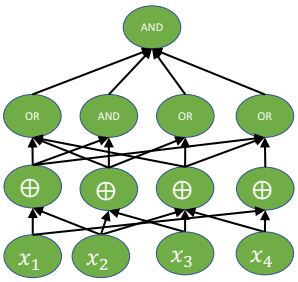


Figure 5: A depth-2 AC^0 circuit with a layer of parity gates at the bottom.

11.3 Circuit Lower Bounds from Approximate Degree Upper Bounds

Let $\mathcal{C}$ be a class of circuits, and imagine that for some function $s(n)$ one has managed to prove that for every circuit $C \in \mathcal{C}$ of size at most $s(n)$ over n-bit inputs, the function computed by C has threshold degree at most $n-1$ (just one below the maximum possible value, n). One can immediately conclude that no circuit in $\mathcal{C}$ of size s is capable of computing the parity function $\oplus_n$, as $\deg_{\pm}(\oplus_n) = n$. For example, recall that a deep result established by Reichardt [Rei11] states that any De Morgan formula of size s has approximate degree (and hence also threshold degree) $O(\sqrt{s})$. This means that no De Morgan formula of size $o(n^2)$ can compute the parity function.

If one shows that every $C \in \mathcal{C}$ of size at most $s(n)$ has *approximate* degree $o(n)$, then one can conclude that no $C \in \mathcal{C}$ can compute the majority function, since $\widetilde{\deg}(\mathsf{MAJ}) = \Theta(n)$. So, for example, by Reichardt's result, no De Morgan formula of size $o(n^2)$ can compute the majority function.

But many circuit classes of interest are powerful enough to compute parity and majority. For example, an important circuit class that we will consider shortly is $\mathrm{AC}^0 \circ \mathsf{MOD}_2$, which denotes the class of all AC^0 circuits augmented with a layer of parity gates above the inputs. Clearly, the parity function $\oplus_n$ is computable with a size-1 circuit in $\mathrm{AC}^0 \circ \mathsf{MOD}_2$ since $\oplus_n$ itself is in MOD_2. Since parity has maximal threshold degree, it may seem that approximate degree and threshold degree are useless for proving lower bounds against such a circuit classes. But it turns out this is inaccurate.

In Sections 11.3.1 and 11.3.2, we focus on implications for circuits computing the so-called inner-product-mod-2 function, $\mathsf{IP2}$. The ideas in this section are due to Tal [Tal17].

11.3.1 Worst-case lower bounds from threshold degree upper bounds

For a given circuit class $\mathcal{C}$ and given class of functions $\mathcal{G}$, let $\mathcal{C} \circ \mathcal{G}$ denote the circuit class in which circuits from $\mathcal{C}$ are augmented with a layer of leaf gates containing functions from $\mathcal{G}$. So for example, if $\mathcal{C}$ is AC^0, and $\mathcal{G}$ is the set of all parity functions on $2n$ bits, i.e., $\mathcal{G} = \mathsf{MOD}_2 \coloneqq \{\chi_S \colon S \subseteq [2n]\}$, then $\mathrm{AC}^0 \circ \mathcal{G}$ denotes the circuit class $\mathrm{AC}^0 \circ \mathsf{MOD}_2$ discussed above. See Figure 5 for an example.

Henceforth, define $\mathcal{G}$ to be the class of all functions on $2n$-bit inputs (x, y) with two-party deterministic communication complexity at most $O(\log n)$, and note that MOD_2 is a subset of $\mathcal{G}$, since the parity function has constant communication complexity. Let us assume that $s(n) = \mathrm{poly}(n)$ is a function such that every circuit C in $\mathcal{C}$ of size $O(s(n))$ defined over $s(n)$ inputs has threshold degree $d \leq o(n/\log n)$. For example, if $\mathcal{C}$ is the class of De Morgan formulas, by Reichardt's result we may take any $s(n) = o((n/\log n)^2)$.

Then, as explained in the next paragraph, every circuit $C' \in \mathcal{C} \circ \mathcal{G}$ of size at most $s(n)$ computes a function $C'(x,y)$ with $\mathbf{UPP}^{\mathrm{cc}}$ complexity $o(n)$. Since $\mathbf{UPP}^{\mathrm{cc}}(\mathsf{IP2}) \geq \Omega(n)$ (see the remark after Corollary 91), it follows that no $C \in \mathcal{C}$ of size at most $s(n)$ computes $\mathsf{IP2}$.

The $\mathbf{UPP}^{\mathrm{cc}}$ protocol for C' is analogous to the one in Fact 71, operating as follows. Since C' has size at most $s(n)$, it has at most $s(n)$ leaf gates, say, $G_1, \ldots, G_\ell \in \mathcal{G}$ for $\ell \leq s(n)$. Let us write $C' = C(G_1(x,y), \ldots, G_\ell(x,y))$, where the circuit $C \in \mathcal{C}$ itself has size at most $s(n)$. Let $z \in \{-1,1\}^\ell$ denote the vector $(G_1(x,y), \ldots, G_\ell(x,y))$. By assumption, there is some polynomial $p(x)$ of degree $d = o(n/\log n)$ that sign-represents C, say, $p(z) = \sum_{S \subseteq [\ell]} \hat{p}(S) \cdot \chi_S(z)$. Alice can use her private randomness to pick a parity χ_S with probability proportional to $|\hat{p}(S)|$. She can send S to Bob (this costs at most $\log_2(\binom{s(n)}{\leq d}) \leq \log_2(s(n)^d) \leq o(n)$ bits), and then Alice and Bob together can compute $z_i = G_i(x,y)$ for each $i \in S$. Since each G_i has communication cost $O(\log n)$, this costs at most $O(|S| \cdot \log n) = o(n)$ bits of communication. Bob can then output $\mathrm{sgn}(\hat{p}(S)) \cdot \chi_S(z)$. Exactly as analyzed in Fact 71, this protocol outputs $C(z) = C'(x)$ with probability strictly greater than $1/2$.

An immediate consequence of the above (along with Reichardt's result [Rei11]) is that, if $\mathcal{C}$ denotes the class of De Morgan formulas, circuits in $\mathcal{C}' = \mathcal{C} \circ \mathcal{G}$ require size at least $\Omega((n/\log n)^2)$ to compute $\mathsf{IP2}$ [Tal17]. This answered a question of Jukna [Juk12], and it is essentially tight since $\mathsf{IP2}$ is computed by quadratic size De Morgan formulas, even without any leaf gates from $\mathcal{G}$.

The next subsection shows that if $\mathcal{C}$ has sublinear approximate degree rather than just sublinear threshold degree, then circuits from $\mathcal{C} \circ \mathcal{G}$ cannot compute $\mathsf{IP2}$ even *on average*.

11.3.2 Average-case lower bounds from approximate degree upper bounds

For simplicity, in this section we describe lower bounds only for circuit classes of the form $\mathcal{C} \circ \mathsf{MOD}_2$. It is well-known that $\mathsf{IP2}$ has correlation 2^{-n} under the uniform distribution with any parity function χ_S, i.e., $2^{-n} \left| \sum_{x,y \in \{-1,1\}^n} \mathsf{IP2}(x,y) \cdot \chi_S(X,Y) \right| = 2^{-n}$. To see this, explicitly calculate the Fourier coefficients of $\mathsf{IP2}(x,y) = \oplus_n(x \wedge y)$ by expressing its multilinear extension as $\prod_{i=1}^n \frac{1}{2}(1 + x_i + y_i - x_i \cdot y_i)$. Applying the distributive law to express this polynomial as a linear combination of parity functions reveals that each Fourier coefficient has magnitude 2^{-n}. The claim then follows from the fact that the correlation of $\mathsf{IP2}$ with the parity function χ_S is precisely the Fourier coefficient $\widehat{\mathsf{IP2}}(S)$.[44]

Let $\varepsilon > 2^{-o(n)}$, and let us assume that $s(n)$ is a function such that every circuit C in $\mathcal{C}$ of size $O(s(n))$ defined over $s(n)$ inputs has ε-approximate degree $d \leq o(n/\log n)$. Let us assume for simplicity that $s(n) \leq n^c$ for some constant $c > 0$. For example, if $\mathcal{C}$ is the class of De Morgan formulas, by the fact that any De Morgan formula of size s has approximate degree $O(\sqrt{s} \cdot \log(1/\varepsilon))$, we can take $s(n) = o((n/(\log(n) \cdot \log(1/\varepsilon)))^2)$.

Let $C' \colon \{-1,1\}^{2n} \to \{-1,1\}$ be a $\mathcal{C} \circ \mathsf{MOD}_2$ circuit of size s^*, and let

$$q = \Pr_{x,y \in \{-1,1\}^n}[C'(x,y) = \mathsf{IP2}(x,y)].$$

Suppose that $q \geq 1/2 + \varepsilon$. Our goal is to show that $s^* > s(n)$. By way of contradiction, let us suppose that $s^* \leq s(n)$ and show that this would imply that $\mathsf{IP2}$ is impossibly well-correlated with some parity function.

[44] In fact, $\mathsf{IP2}(x,y)$ has correlation $2^{-\Omega(n)}$ under the uniform distribution with any function $f(x,y)$ computed by a two-party deterministic communication protocol of cost $O(1)$. We do not cover the proof of this fact in this survey, but it can be used to extend the results of this section from $\mathcal{C} \circ \mathsf{MOD}_2$ to $\mathcal{C} \circ \mathcal{G}$.

Analogous to the previous section, let $\ell \leq s^*$ denote the number of parity gates in $\mathcal{C}$, with the ith parity gate denoted by $G_i(x)\colon \{-1,1\}^n \to \{-1,1\}$. Let us write $C'(x,y) = C(G_1(x,y),\ldots,G_\ell(x,y))$, where $C \in \mathcal{C}$ is a circuit of size at most s^* defined over $\ell \leq s^*$ inputs. By assumption, there exists a polynomial p of degree at most $d = o(n/\log n)$ such that, for all $w \in \{-1,1\}^\ell$, $|p(w) - C(w)| \leq \varepsilon$.

Next, we show that under the uniform distribution, $\mathsf{IP2}(x,y)$ correlates well with $p(G_1(x),\ldots,G_\ell(x))$. We decompose the expectation $\mathbf{E}_{x,y\in\{-1,1\}^n}[p(x,y)\cdot\mathsf{IP2}(x,y)]$ according to whether or not $\mathsf{IP2}(x,y) = C'(x,y)$:

$$\begin{aligned}
&\mathbf{E}_{x,y\in\{-1,1\}^n}[p(G_1(x,y),\ldots,G_\ell(x,y))\cdot\mathsf{IP2}(x,y)] = \\
&\mathbf{E}_{x,y\in\{-1,1\}^n}[p(G_1(x,y),\ldots,G_\ell(x,y))\cdot\mathsf{IP2}(x,y)|\mathsf{IP2}(x,y) = C'(x,y)]\cdot\Pr[\mathsf{IP2}(x,y) = C'(x,y)]+ \\
&\mathbf{E}_{x,y\in\{-1,1\}^n}[p(G_1(x,y),\ldots,G_\ell(x,y))\cdot\mathsf{IP2}(x,y)|\mathsf{IP2}(x,y) \neq C'(x,y)]\cdot\Pr[\mathsf{IP2}(x,y) \neq C'(x,y)] \\
&\qquad\qquad \geq (1-\varepsilon)\cdot q + (-1-\varepsilon)\cdot(1-q) \\
&\qquad\qquad = 2q - 1 - \varepsilon \geq 2\cdot(1/2+\varepsilon) - 1 - \varepsilon = \varepsilon. \qquad (98)
\end{aligned}$$

Next, we write $p(z)$ as a multilinear polynomial: $p(z) = \sum_{S\subseteq[\ell],|S|\leq d}\hat{p}(S)\cdot\prod_{i\in S} z_i$. Since $\hat{p}(S) = \mathbf{E}_{z\in\{-1,1\}^\ell}[p(z)\cdot\prod_{i\in S} z_i]$, we have that $|\hat{p}(S)| \leq 1+\varepsilon$ for every S. Note that there are at most $\binom{\ell}{\leq d}$ monomials in p. Invoking Equation (98), we have:

$$\begin{aligned}
\varepsilon &\leq \mathbf{E}_{x,y\in\{-1,1\}^n}\left[p(G_1(x,y),\ldots,G_\ell(x,y))\cdot\mathsf{IP2}(x,y)\right] \\
&= \mathbf{E}_{x,y\in\{-1,1\}^n}\left[\sum_{S\subseteq[\ell],|S|\leq d}\hat{p}(S)\prod_{i\in S}G_i(x,y)\cdot\mathsf{IP2}(x,y)\right] \\
&= \sum_{S\subseteq[\ell],|S|\leq d}\hat{p}(S)\cdot\mathbf{E}_{x,y\in\{-1,1\}^n}\left[\prod_{i\in S}G_i(x,y)\cdot\mathsf{IP2}(x,y)\right] \\
&\leq \sum_{S\subseteq[\ell],|S|\leq d}(1+\varepsilon)\left|\mathbf{E}_{x,y\in\{-1,1\}^n}\left[\prod_{i\in S}G_i(x,y)\cdot\mathsf{IP2}(x,y)\right]\right|.
\end{aligned}$$

Hence there must exist a set $S \subseteq [\ell]$ with size at most d such that

$$\left|\mathbf{E}_{x,y\in\{-1,1\}^n}\left[\prod_{i\in S}G_i(x,y)\cdot\mathsf{IP2}(x,y)\right]\right| \geq \frac{\varepsilon}{\binom{\ell}{\leq d}(1+\varepsilon)} \geq (\varepsilon/2)\cdot(s^*)^{-d} \geq \varepsilon\cdot 2^{o(n)} \geq 2^{o(n)}.$$

Here, the final two inequalities exploited the assumptions that $\varepsilon \geq 2^{-o(n)}$ and $s^* \leq n^c$ for some constant $c > 0$. Since $\prod_{i\in S} G_i(x,y)$ is a product of parity functions and hence is itself a parity function, this contradicts the fact that $\mathsf{IP2}$ has correlation 2^{-n} with any parity function. We conclude that s^* must be greater than $s(n)$.

Instantiations of the average-case lower bound. A first instantiation was previously indicated, namely if $\mathcal{C}$ denotes the class of De Morgan formulas, then we can set $s(n) = o(n^2/(\log^2(n)\log^2(1/\varepsilon)))$ to conclude that any $\mathcal{C}\circ\mathsf{MOD}_2$ circuit computing $\mathsf{IP2}$ on a $1/2+\varepsilon$ fraction of inputs requires size at least $\Omega(n^2/(\log^2(n)\log^2(1/\varepsilon)))$ [Tal17].

A second application is to a well-known frontier problem in circuit complexity, which is to prove superpolynomial lower bounds for the size of $\mathrm{AC}^0 \circ \mathsf{MOD}_2$ circuits computing the $\mathsf{IP2}$ function. This question has been related to a variety of frontier open problems in communication complexity and pseudorandomness, see for example [ER21].[45] Unfortunately, the best lower bound on the size of $\mathrm{AC}^0 \circ \mathsf{MOD}_2$ circuits computing $\mathsf{IP2}$ are only slightly superlinear. Superlinear worst-case lower bounds were first proved in [CGJ+18], while Bun, Kothari, and Thaler [BKT21] extended the bounds to hold in the average case, as follows.

If $D \geq 1$ is a constant and $\mathcal{C}$ denotes the class of depth-D AC^0 circuits, then [BKT21] proved that the approximate degree of any linear-sized circuit over n inputs in $\mathcal{C}$ has ε-approximate degree at most $O(n^{1-2^{-D}} \log^{2^{-D}}(1/\varepsilon))$. Hence, can set $s(n) = o\left(n/\left(\log(n) \cdot \log^{2^{-D}}(1/\varepsilon)\right)\right)^{1/(1-2^{-D})}$. In particular, if $\varepsilon = n^{-\log n}$, then we can set $s(n) = \left(n/\log^3 n\right)^{1/(1-2^{-D})} > \Omega(n^{1+2^{-D}})$. We conclude that any depth-D AC^0 circuit with a layer of parity gates at the bottom that computes $\mathsf{IP2}$ on more than a $1/2 + n^{-\log n}$ fraction of inputs requires (slightly) superlinear size $\Omega(n^{1+2^{-D}})$.

Additional applications of these techniques for proving average-case lower bounds were considered by Kabanets et al. [KKL+20].

11.4 Parity is not in $\mathsf{LTF} \circ \mathsf{AC}^0$

A famous result in circuit complexity is that parity, $\oplus_n$, is not in AC^0 [FSS84a], i.e., constant-depth circuits of unbounded fan-in that compute parity require exponential size. We cover one such proof of this result, due to [ABFR94], which is based on threshold degree. In fact, the analysis establishes the stronger result that parity is not in $\mathsf{LTF} \circ \mathsf{AC}^0$, the class of polynomial size AC^0 circuits with a threshold gate at the top (i.e., the output gate computes a linear threshold function).

The proof proceeds in two steps. First, show that any polynomial p of degree $o(\sqrt{n})$ disagrees in sign with parity on at least a $1/2 - o(1)$ fraction of inputs. That is, $p(x) \cdot \oplus_n(x) < 0$ for at least $(1/2 - o(1)) \cdot 2^n$ inputs x in $\{-1, 1\}^n$.

Second, show that for any $\mathsf{LTF} \circ \mathsf{AC}^0$ circuit $\mathcal{C}$, there is a polynomial p of polylogarithmic degree that agrees in sign with $\mathcal{C}$ on 99% of inputs.[46] Together, these two results imply that parity cannot be computed by any $\mathsf{LTF} \circ \mathsf{AC}^0$ circuit of polynomial size (in fact, $\mathsf{LTF} \circ \mathsf{AC}^0$ circuits require size at least $\exp\left(n^{1/O(d)}\right)$ to compute parity).

Step 1. Before proving the first step, let us first explain that the bound is tight: one can *exactly* compute parity on, say, 99% of all inputs by "interpolating" the middle $O(\sqrt{n})$ Hamming layers of the Boolean hypercube. That is, standard bounds on the Binomial coefficients imply that there is some constant $c > 0$ such that 99% of the 2^n inputs in $\{-1, 1\}^n$ have Hamming weight between $n/2 - c\sqrt{n}$ and $n/2 + c\sqrt{n}$. Define $P(t)$ via interpolation to be the unique polynomial of degree at most $2c\sqrt{n}$ that evaluates to $(-1)^t$ at all integers $t \in [n/2 - c\sqrt{n}, n/2 + c\sqrt{n}]$. Let $p(x) = P(|x|)$. Then p sign-represents (in fact, exactly computes) parity on 99% of all inputs.

This construction turns out to be *exactly* optimal in terms of the number of inputs at which it agrees in sign with parity. It can be checked that p is a degree-d polynomial that disagrees in sign

[45] Superpolynomial lower bounds against $\mathrm{AC}^0 \circ \mathsf{MOD}_2$ circuits computing other functions such as MAJ are known, but the techniques used to prove these lower bounds totally break down for $\mathsf{IP2}$.

[46] More precisely, for any $\mathsf{LTF} \circ \mathsf{AC}^0$ circuit of size s and depth d, there is a polynomial p of degree $(\log s)^{O(d)} \cdot \log(1/\delta)$ that agrees with $\mathcal{C}$ on a $1 - \delta$ fraction of inputs [HS19].

with parity on $\binom{n}{\leq k}$ inputs where $k = (n-d)/2$.[47] If $d = o(\sqrt{n})$, then $\binom{n}{\leq k} \geq (1/2 - o(1)) \cdot 2^n$.

To prove that this is optimal, it suffices to show that for any set $S \subseteq \{-1,1\}^n$ of size at most $\binom{n}{\leq k}$, any degree $d = n - 2k$ polynomial p must disagree in sign with parity on at least one input outside of S. This turns out to be equivalent to constructing a dual polynomial $\psi \colon \{-1,1\}^n \to \mathbb{R}$ of pure high degree at least $d = n - 2k$ such that:

(a) $\psi(x) = 0$ for all $x \in S$, i.e., ψ vanishes on S.

(b) ψ has perfect correlation with parity, i.e., $\psi(x) \cdot \oplus_n(x) \geq 0$ for all $x \in \{-1,1\}^n$.

Intuitively, such a ψ is a dual witness to the high threshold degree of $\oplus_n$ that "ignores" inputs outside of S. Hence, ψ witnesses that any PTF p for parity must have degree at least d, *regardless* of how p behaves at inputs in S.

Here is the construction of the dual polynomial ψ. Let S be any subset of size at most $\binom{n}{\leq k}$. Note that $\binom{n}{\leq k}$ is the number of parities of degree at most k. By elementary linear algebra, there exists a non-zero polynomial q of degree at most k that vanishes on S. That is, since there are $\binom{n}{\leq k}$ coefficients of q, we can choose them so as to ensure that $q(x) = 0$ for all $x \in S$. The polynomial q^2 then has degree at most $2k$, and $q(x) \geq 0$ for all $x \in \{-1,1\}^n$. The function $\psi(x) \coloneqq \oplus_n(x) \cdot q^2(x)$ is a non-zero dual polynomial with pure high degree $n - 2k$ (see the proof of Lemma 31) that is perfectly correlated with parity, as required.

Step 2. The key is to show that for any AC^0 circuit $\mathcal{C}$ of depth d and size s, there is a polynomial p of degree $(\log s/\delta)^{O(d)}$ that exactly computes $\mathcal{C}$ on a $1 - \delta$ fraction of inputs. One then obtains a polynomial that sign-represents any $\mathsf{LTF} \circ \mathsf{AC}^0$ circuit of size by s on 99% of inputs by, first, exactly computing each constituent AC^0 circuit on all but a $1/(100s)$ fraction of inputs—which requires degree just $(\log s)^{O(d)}$—and then taking the appropriate linear combination of the resulting polynomials, i.e., with the coefficients of the linear combination given by the weights of the LTF gate.

Probabilistic Degree. The technical core of the key result is an upper bound of $O(\log(n) \cdot \log(1/\delta))$ on the so-called *δ-error probabilistic degree of* OR. Here, a δ-error probabilistic polynomial of degree d over the reals for a circuit $\mathcal{C}(x_1, \ldots, x_n)$ is a *random* polynomial $P(x_1, \ldots, x_n)$ such that for any $x \in \{-1,1\}^n$, $\Pr_P[\mathcal{C}(x) \neq P(x)] \leq \delta$. That is, P is drawn at random from some distribution over degree-d polynomials, and the requirement is that for every fixed input x, $P(x)$ is very likely to *exactly equal* $\mathcal{C}(x)$. Before proving the upper bound of $O(\log(n) \cdot \log(1/\delta))$ on the probabilistic degree of OR, we explain why it implies an upper bound of $O(\log^{O(d)}(n))$ on the probabilistic degree of any AC^0 circuit of depth d.

From Probabilistic Polynomials for OR and AND to AC^0. Let $\mathcal{C}$ be an AC^0 circuit of size s. Each AND and OR gate of $\mathcal{C}$ has a $1/(100s)$-error probabilistic polynomial of degree $O(\log^2(s))$. For each gate, draw such a polynomial at random and consider the gate-by-gate composition of the resulting polynomials. This composed polynomial p has degree $\log(s)^{O(d)}$. For each input x, by a union bound over all s gates of $\mathcal{C}$, the probability that $p(x) \neq \mathcal{C}(x)$ is at most $s \cdot (1/(100s)) \leq 1/100$. Hence, the expected number of inputs x at which $p(x) = \mathcal{C}(x)$ is at least $.99 \cdot 2^n$. By averaging,

[47] Recall from Section 2.1 that $\binom{n}{\leq k}$ denotes $\sum_{i=0}^{k} \binom{n}{i}$.

there exists *some* polynomial p of degree at most $\log(s)^{O(d)}$ that agrees with $\mathcal{C}$ on at least 99% of all inputs.

Probabilistic Polynomial for OR. The δ-error probabilistic polynomial for OR that we cover is due to [BRS+90]. The construction is reminiscent of an earlier one over finite fields due to Razborov [Raz87].

We begin with a construction that achieves $\delta = 1/(2e)$. Let $\mathcal{F}$ be a random collection of $1+\log n$ subsets of $[n]$ generated as follows. For $i = 0, 1, \ldots, \log n$, the i'th subset, S_i, of $\mathcal{F}$ is generated by independently including each $j \in [n]$ in S_i with probability 2^{-i}. Then define

$$P(x_1, \ldots, x_n) = -1 + 2 \prod_{S \in \mathcal{F}} \left(1 + \sum_{j \in S} (x_j - 1)/2 \right).$$

Observe that P has degree $|\mathcal{F}| = 1 + \log n$. Showing that this indeed yields a $(1/(2e))$-error probabilistic polynomial for OR relies on the following two observations. First, if $x \in \mathsf{OR}^{-1}(1)$, so that $x = \mathbf{1}_n$, then $P(x) = 1$ with probability 1 over the choice of P. This holds because $(x_j - 1)/2$ will equal 0 for every j. Hence, P always has the desired behavior on $x = \mathbf{1}_n$.

Second, so long as there is at least one set $S \in \mathcal{F}$ such that there is *exactly one* $j \in S$ with $x_j = -1$, then $P(x) = -1$. We now show that this occurs with probability at least $1/(2e)$ over the random choice of subsets in $\mathcal{F}$. This guarantees that the resulting distribution over polynomials P has the desired behavior on any $x \in \mathsf{OR}^{-1}(-1)$.

Specifically, fix an $x \in \{-1, 1\}^n$ with $|x| \geq 1$, and let T denote the set of coordinates of x that are equal to -1. Let 2^i be the smallest power of 2 greater than or equal to $|x|$. The probability that exactly one coordinate j from T is in S_i is

$$\sum_{j \in T} \Pr[j \in S_i] \cdot \Pr[\text{no other elements of } T \text{ are in } S_i] = |x| \cdot 2^{-i} \cdot (1 - 2^{-i})^{|x|-1}.$$

This probability is at least $1/(2e)$. To see this, observe that $|x| \cdot 2^{-i} > 1/2$, while

$$(1 - 2^{-i})^{|x|-1} \geq (1 - 2^{-|x|})^{|x|-1} \geq 1/e,$$

where the final inequality invokes Fact 2.

To drive the error down from $1/(2e)$ to δ, one generates $\ell = O(\log(1/\delta))$ independent "copies" of each set S_i and adds all of the copies to $\mathcal{F}$. That is, for each copy of S_i, each coordinate $j \in [n]$ is included independently with probability 2^{-i}. The probability that *none* of the copies of S_i contain exactly one element of T is at most $1/(2e)^{\ell}$. Hence, for a suitable $\ell = O(\log(1/\delta))$, this probability is at most δ. Hence, P is a δ-error probabilistic approximation for OR of degree at most $O(\log(n) \cdot \log(1/\delta))$.

Acknowledgements. We are grateful to Lane Hemaspaandra for inviting us to write the SIGACT News column that served as the genesis of this manuscript. We would also like to thank Lane, as well as Karthik Gajulapalli, Justin Goldstein, Samuel King, Satyajeet Nagargoje, Sidhant Saraogi, and Shuchen Zhu for their thoughtful and detailed comments on earlier versions of this manuscript. Finally, we are especially grateful to the two anonymous reviewers who provided insightful and immensely valuable feedback on the entirety of this manuscript.

References

[Aar02] Scott Aaronson. Quantum lower bound for the collision problem. In John H. Reif, editor, *Proceedings on 34th Annual ACM Symposium on Theory of Computing, May 19-21, 2002, Montréal, Québec, Canada*, pages 635–642. ACM, 2002.

[Aar08] Scott Aaronson. The polynomial method in quantum and classical computing. In *Foundations of Computer Science*, page 3, 2008.

[Aar12] Scott Aaronson. Impossibility of succinct quantum proofs for collision-freeness. *Quantum Information & Computation*, 12(1-2):21–28, 2012.

[ABDK+21] Scott Aaronson, Shalev Ben-David, Robin Kothari, Shravas Rao, and Avishay Tal. Degree vs. approximate degree and quantum implications of Huang?s sensitivity theorem. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1330–1342, 2021.

[ABFR94] James Aspnes, Richard Beigel, Merrick Furst, and Steven Rudich. The expressive power of voting polynomials. *Combinatorica*, 14(2):135–148, 1994.

[ABP19] Srinivasan Arunachalam, Jop Briët, and Carlos Palazuelos. Quantum query algorithms are completely bounded forms. *SIAM Journal on Computing*, 48(3):903–925, 2019.

[AKKT20] Scott Aaronson, Robin Kothari, William Kretschmer, and Justin Thaler. Quantum lower bounds for approximate counting via laurent polynomials. In Shubhangi Saraf, editor, *35th Computational Complexity Conference, CCC 2020, July 28-31, 2020, Saarbrücken, Germany (Virtual Conference)*, volume 169 of *LIPIcs*, pages 7:1–7:47. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020.

[All89] Eric Allender. A note on the power of threshold circuits. In *Foundations of Computer Science*, pages 580–584, 1989.

[Amb05] Andris Ambainis. Polynomial degree and lower bounds in quantum complexity: Collision and element distinctness with small range. *Theory of Computing*, 1(1):37–46, 2005.

[Amb06] Andris Ambainis. Polynomial degree vs. quantum query complexity. *Journal of Computer and System Sciences*, 72(2):220–238, 2006.

[Amb07] Andris Ambainis. Quantum walk algorithm for element distinctness. *SIAM Journal on Computing*, 37(1):210–239, 2007.

[Amb18] Andris Ambainis. Understanding quantum algorithms via query complexity. In *Proceedings of the International Congress of Mathematicians*, May 2018.

[AS04] Scott Aaronson and Yaoyun Shi. Quantum lower bounds for the collision and the element distinctness problems. *Journal of the ACM*, 51(4):595–605, 2004.

[AW09] Scott Aaronson and Avi Wigderson. Algebrization: A new barrier in complexity theory. *ACM Transactions on Computation Theory (TOCT)*, 1(1):1–54, 2009.

[BBBV97] Charles H Bennett, Ethan Bernstein, Gilles Brassard, and Umesh Vazirani. Strengths and weaknesses of quantum computing. *SIAM Journal on Computing*, 26(5):1510–1523, 1997.

[BBC+01] Robert Beals, Harry Buhrman, Richard Cleve, Michele Mosca, and Ronald De Wolf. Quantum lower bounds by polynomials. *Journal of the ACM*, 48(4):778–797, 2001.

[BBGK18] Shalev Ben-David, Adam Bouland, Ankit Garg, and Robin Kothari. Classical lower bounds from quantum upper bounds. In *Foundations of Computer Science*, pages 339–349, 2018.

[BCDWZ99] Harry Buhrman, Richard Cleve, Ronald De Wolf, and Christof Zalka. Bounds for small-error and zero-error quantum algorithms. In *Foundations of Computer Science*, pages 358–368, 1999.

[BCH+19] Adam Bouland, Lijie Chen, Dhiraj Holden, Justin Thaler, and Prashant Nalini Vasudevan. On the power of statistical zero knowledge. *SIAM Journal on Computing*, 49(4):1–58, 2019.

[BCW98] Harry Buhrman, Richard Cleve, and Avi Wigderson. Quantum vs. classical communication and computation. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 63–68, 1998.

[BDF+22] Andrej Bogdanov, Krishnamoorthy Dinesh, Yuval Filmus, Yuval Ishai, Avi Kaplan, and Akshayaram Srinivasan. Bounded indistinguishability for simple sources. In Mark Braverman, editor, *13th Innovations in Theoretical Computer Science Conference, ITCS 2022, January 31 - February 3, 2022, Berkeley, CA, USA*, volume 215 of *LIPIcs*, pages 26:1–26:18. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022.

[BdW01] Harry Buhrman and Ronald de Wolf. Communication complexity lower bounds by polynomials. In *Proceedings 16th Annual IEEE Conference on Computational Complexity*, pages 120–130. IEEE, 2001.

[Bei94] Richard Beigel. Perceptrons, PP, and the polynomial hierarchy. *Computational complexity*, 4(4):339–349, 1994.

[Bel12] Aleksandrs Belovs. Learning-graph-based quantum algorithm for k-distinctness. In *Foundations of Computer Science*, pages 207–216, 2012.

[Bel15] Aleksandrs Belovs. Quantum algorithms for learning symmetric juntas via the adversary bound. *Computational Complexity*, 24(2):255–293, 2015.

[Ben20] Gal Beniamini. The approximate degree of bipartite perfect matching. *arXiv preprint arXiv:2004.14318*, 2020. To appear in CCC 2022.

[BFS86] László Babai, Peter Frankl, and Janos Simon. Complexity classes in communication complexity theory. In *Foundations of Computer Science*, pages 337–347, 1986.

[BG22] Jop Briët and Francisco Escudero Gutiérrez. On converses to the polynomial method. *arXiv preprint arXiv:2204.12303*, 2022.

[BHMT02] Gilles Brassard, Peter Høyer, Michele Mosca, and Alain Tapp. Quantum amplitude amplification and estimation. In *Quantum computation and information (Washington, DC, 2000)*, volume 305 of *Contemp. Math.*, pages 53–74. Amer. Math. Soc., Providence, RI, 2002.

[BHT97] Gilles Brassard, Peter Hoyer, and Alain Tapp. Quantum algorithm for the collision problem. *ACM SIGACT News (Cryptology Column)*, 28:14–19, 1997. quant-ph/9705002.

[BHT98] Gilles Brassard, Peter Høyer, and Alain Tapp. Quantum counting. In Kim Guldstrand Larsen, Sven Skyum, and Glynn Winskel, editors, *Automata, Languages and Programming, 25th International Colloquium, ICALP'98, Aalborg, Denmark, July 13-17, 1998, Proceedings*, volume 1443 of *Lecture Notes in Computer Science*, pages 820–831. Springer, 1998.

[BIVW16] Andrej Bogdanov, Yuval Ishai, Emanuele Viola, and Christopher Williamson. Bounded indistinguishability and the complexity of recovering secrets. In *International Cryptology Conference*, volume 9816, pages 593–618, 2016.

[BKT18] Mark Bun, Robin Kothari, and Justin Thaler. The polynomial method strikes back: Tight quantum query bounds via dual polynomials. In *Symposium on Theory of Computing*, pages 297–310, 2018.

[BKT21] Mark Bun, Robin Kothari, and Justin Thaler. Quantum algorithms and approximating polynomials for composed functions with shared inputs. *Quantum*, 5:543, September 2021.

[BM12] Paul Beame and Widad Machmouchi. The quantum query complexity of ac^0. *Quantum Inf. Comput.*, 12(7-8):670–676, 2012.

[BMT21] Mark Bun, Nikhil S Mande, and Justin Thaler. Sign-rank can increase under intersection. *ACM Transactions on Computation Theory (TOCT)*, 13(4):1–17, 2021.

[BMTW19] Andrej Bogdanov, Nikhil S. Mande, Justin Thaler, and Christopher Williamson. Approximate degree, secret sharing, and concentration phenomena. In Dimitris Achlioptas and László A. Végh, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2019, September 20-22, 2019, Massachusetts Institute of Technology, Cambridge, MA, USA*, volume 145 of *LIPIcs*, pages 71:1–71:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2019.

[BN21] Gal Beniamini and Noam Nisan. Bipartite perfect matching as a real polynomial. In Samir Khuller and Virginia Vassilevska Williams, editors, *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*, pages 1118–1131. ACM, 2021.

[BNRdW07] Harry Buhrman, Ilan Newman, Hein Rohrig, and Ronald de Wolf. Robust polynomials and quantum algorithms. *Theory of Computing Systems*, 40(4):379–395, 2007.

[BRS+90] Richard Beigel, Nick Reingold, Daniel Spielman, et al. *The perceptron strikes back*. Yale University, Department of Computer Science, 1990.

[BRS95] Richard Beigel, Nick Reingold, and Daniel A. Spielman. PP is closed under intersection. *Journal of Computer and System Sciences*, 50(2):191–202, 1995.

[BS92] Jehoshua Bruck and Roman Smolensky. Polynomial threshold functions, ac^0 functions, and spectral norms. *SIAM Journal on Computing*, 21(1):33–42, 1992.

[BT15a] Mark Bun and Justin Thaler. Dual lower bounds for approximate degree and Markov–Bernstein inequalities. *Information and Computation*, 243:2–25, 2015.

[BT15b] Mark Bun and Justin Thaler. Hardness amplification and the approximate degree of constant-depth circuits. In *International Colloquium on Automata, Languages, and Programming*, pages 268–280, 2015.

[BT18a] Mark Bun and Justin Thaler. Approximate degree and the complexity of depth three circuits. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[BT18b] Mark Bun and Justin Thaler. Approximate Degree and the Complexity of Depth Three Circuits. In Eric Blais, Klaus Jansen, José D. P. Rolim, and David Steurer, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2018)*, volume 116 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 35:1–35:18, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[BT19a] Mark Bun and Justin Thaler. The large-error approximate degree of AC^0. In *International Conference on Randomization and Computation*, volume 145 of *LIPIcs*, pages 55:1–55:16. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2019.

[BT19b] Mark Bun and Justin Thaler. A nearly optimal lower bound on the approximate degree of AC^0. *SIAM Journal on Computing*, 49(4):59–96, 2019.

[BVdW07] Harry Buhrman, Nikolay Vereshchagin, and Ronald de Wolf. On computation and communication with small bias. In *Twenty-Second Annual IEEE Conference on Computational Complexity (CCC'07)*, pages 24–32. IEEE, 2007.

[BW17] Andrej Bogdanov and Christopher Williamson. Approximate bounded indistinguishability. In Ioannis Chatzigiannakis, Piotr Indyk, Fabian Kuhn, and Anca Muscholl, editors, *44th International Colloquium on Automata, Languages, and Programming, ICALP 2017, July 10-14, 2017, Warsaw, Poland*, volume 80 of *LIPIcs*, pages 53:1–53:11. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017.

[CGJ+18] Mahdi Cheraghchi, Elena Grigorescu, Brendan Juba, Karl Wimmer, and Ning Xie. $\mathrm{AC}^0 \circ \mathrm{MOD}_2$ lower bounds for the Boolean inner product. *J. Comput. Syst. Sci.*, 97:45–59, 2018.

[CIL17] Kuan Cheng, Yuval Ishai, and Xin Li. Near-optimal secret sharing and error correcting codes in AC^0. In *Theory of Cryptography Conference*, pages 424–458. Springer, 2017.

[CM18] Arkadev Chattopadhyay and Nikhil Mande. A short list of equalities induces large sign rank. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 47–58. IEEE, 2018.

[CR92] Don Coppersmith and Theodore J Rivlin. The growth of polynomials bounded at equally spaced points. *SIAM Journal on Mathematical Analysis*, 23(4):970–983, 1992.

[CTUW14] Karthekeyan Chandrasekaran, Justin Thaler, Jonathan Ullman, and Andrew Wan. Faster private release of marginals on small databases. In *Innovations in Theoretical Computer Science*, pages 387–402, 2014.

[DGJ+10] Ilias Diakonikolas, Parikshit Gopalan, Ragesh Jaiswal, Rocco A Servedio, and Emanuele Viola. Bounded independence fools halfspaces. *SIAM Journal on Computing*, 39(8):3441–3462, 2010.

[DGRMT22] Marcel Dall'Agnol, Tom Gur, Subhayan Roy Moulik, and Justin Thaler. Quantum Proofs of Proximity. *Quantum*, 6:834, October 2022.

[dW08] Ronald de Wolf. A note on quantum algorithms and the minimal degree of epsilon-error polynomials for symmetric functions. *arXiv preprint arXiv:0802.1816*, 2008.

[ER21] Michael Ezra and Ron D. Rothblum. Small circuits imply efficient Arthur-Merlin protocols. In *ECCCTR: Electronic Colloquium on Computational Complexity, technical reports*, 2021.

[For02] Jürgen Forster. A linear lower bound on the unbounded error probabilistic communication complexity. *Journal of Computer and System Sciences*, 65(4):612–625, 2002.

[FSS84a] Merrick Furst, James B Saxe, and Michael Sipser. Parity, circuits, and the polynomial-time hierarchy. *Mathematical systems theory*, 17(1):13–27, 1984.

[FSS84b] Merrick L. Furst, James B. Saxe, and Michael Sipser. Parity, circuits, and the polynomial-time hierarchy. *Math. Syst. Theory*, 17(1):13–27, 1984.

[Gil77] John Gill. Computational complexity of probabilistic Turing machines. *SIAM Journal on Computing*, 6(4):675–695, 1977.

[GR09] Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. *SIAM Journal on Computing*, 39(2):742–765, 2009.

[GS10] Dmitry Gavinsky and Alexander A. Sherstov. A separation of NP and coNP in multiparty communication complexity. *Theory of Computing*, 6(1):227–245, 2010.

[Hau92] David Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inf. Comput.*, 100(1):78–150, 1992.

[HS19] Prahladh Harsha and Srikanth Srinivasan. On polynomial approximations to AC. *Random Struct. Algorithms*, 54(2):289–303, 2019.

[Hua19] Hao Huang. Induced subgraphs of hypercubes and a proof of the sensitivity conjecture. *Annals of Mathematics*, 190(3):949–955, 2019.

[Juk12] Stasys Jukna. *Boolean Function Complexity - Advances and Frontiers*, volume 27 of *Algorithms and combinatorics*. Springer, 2012.

[KKL+20] Valentine Kabanets, Sajin Koroth, Zhenjian Lu, Dimitrios Myrisiotis, and Igor Carboni Oliveira. Algorithms and lower bounds for De Morgan formulas of low-communication leaf gates. In Shubhangi Saraf, editor, *35th Computational Complexity Conference, CCC 2020, July 28-31, 2020, Saarbrücken, Germany (Virtual Conference)*, volume 169 of *LIPIcs*, pages 15:1–15:41. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020.

[KKMS08] Adam Tauman Kalai, Adam R Klivans, Yishay Mansour, and Rocco A Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

[Kla03] Hartmut Klauck. Rectangle size bounds and threshold covers in communication complexity. In *18th Annual IEEE Conference on Computational Complexity (Complexity 2003), 7-10 July 2003, Aarhus, Denmark*, pages 118–134. IEEE Computer Society, 2003.

[Kla11] Hartmut Klauck. On arthur merlin games in communication complexity. In *2011 IEEE 26th Annual Conference on Computational Complexity*, pages 189–199. IEEE, 2011.

[KLS96] Jeff Kahn, Nathan Linial, and Alex Samorodnitsky. Inclusion-exclusion: Exact and approximate. *Combinatorica*, 16(4):465–477, 1996.

[Ko89] Ker-I Ko. Constructing oracles by lower bound techniques for circuits, 1989.

[Kre95] Ilan Kremer. *Quantum communication*. Citeseer, 1995.

[KS04] Adam R Klivans and Rocco A Servedio. Learning DNF in time $2^{\tilde{O}(n^{1/3})}$. *Journal of Computer and System Sciences*, 68(2):303–318, 2004.

[KS07] Adam R Klivans and Alexander A Sherstov. A lower bound for agnostically learning disjunctions. In *International Conference on Computational Learning Theory*, pages 409–423. Springer, 2007.

[KSS94] Michael J Kearns, Robert E Schapire, and Linda M Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2):115–141, 1994.

[Lee09] Troy Lee. A note on the sign degree of formulas. *arXiv preprint arXiv:0909.4607*, 2009.

[Lit88] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2(4):285–318, 1988.

[Lok09] Satyanarayana V Lokam. *Complexity lower bounds using linear algebra*. Now Publishers Inc, 2009.

[LS09a] Troy Lee and Adi Shraibman. An approximation algorithm for approximation rank. In *2009 24th Annual IEEE Conference on Computational Complexity*, pages 351–357. IEEE, 2009.

[LS09b] Troy Lee and Adi Shraibman. Lower bounds in communication complexity. *Foundations and Trends in Theoretical Computer Science*, 3(4):263–399, 2009.

[LS09c] Nathan Linial and Adi Shraibman. Learning complexity vs communication complexity. *Comb. Probab. Comput.*, 18(1-2):227–245, 2009.

[LS09d] Nati Linial and Adi Shraibman. Lower bounds in communication complexity based on factorization norms. *Random Struct. Algorithms*, 34(3):368–394, 2009.

[LZ10] Troy Lee and Shengyu Zhang. Composition theorems in communication complexity. In Samson Abramsky, Cyril Gavoille, Claude Kirchner, Friedhelm Meyer auf der Heide, and Paul G. Spirakis, editors, *Automata, Languages and Programming, 37th International Colloquium, ICALP 2010, Bordeaux, France, July 6-10, 2010, Proceedings, Part I*, volume 6198 of *Lecture Notes in Computer Science*, pages 475–489. Springer, 2010.

[Man18] Nikhil Shekhar Mande. *Communication Complexity of XOR Functions*. PhD thesis, Tata Institute of Fundamental Research Mumbai, 2018.

[Mar90] Andrei Andreyevich Markov. On a question by DI Mendeleev. *Zapiski Imperatorskoi Akademii Nauk*, 62(1-24):12, 1890.

[MP69] Marvin Minsky and Seymour Papert. *Perceptrons: An introduction to computational geometry*. MIT Press, 1969.

[MTT61] Saburo Muroga, Iwao Toda, and Satoru Takasu. Theory of majority switching elements. *J. Franklin Institute*, 271(5):376–418, 1961.

[MTZ20] Nikhil S. Mande, Justin Thaler, and Shuchen Zhu. Improved approximate degree bounds for k-distinctness. In *Theory of Quantum Computation, Communication and Cryptography*, volume 158, pages 2:1–2:22, 2020.

[MW05] Chris Marriott and John Watrous. Quantum Arthur–Merlin games. *Computational Complexity*, 14(2):122–152, 2005.

[Nis93] Noam Nisan. The communication complexity of threshold gates. *Combinatorics, Paul Erdos is Eighty*, 1:301–315, 1993.

[NS94] Noam Nisan and Mario Szegedy. On the degree of Boolean functions as real polynomials. *Computational Complexity*, 4(4):301–313, 1994.

[O'D11] Ryan O'Donnell. Linear and semidefinite programming (advanced algorithms) fall 2011 lecture notes, 2011. `Available at: https://www.cs.cmu.edu/afs/cs.cmu.edu/academic/class/15859-f11/www/`.

[Pat92] Ramamohan Paturi. On the degree of polynomials that approximate symmetric boolean functions (preliminary version). In *Symposium on Theory of Computing*, pages 468–474, 1992.

[Pod08] Vladimir V Podolskii. A uniform lower bound on weights of perceptrons. In *International Computer Science Symposium in Russia*, pages 261–272. Springer, 2008.

[Pod09] Vladimir V Podolskii. Perceptrons of large weight. *Problems of Information Transmission*, 45(1):46–53, 2009.

[PS86] Ramamohan Paturi and Janos Simon. Probabilistic communication complexity. *Journal of Computer and System Sciences*, 33(1):106–123, 1986.

[Raz87] Alexander A Razborov. Lower bounds on the size of bounded depth circuits over a complete basis with logical addition. *Mathematical Notes of the Academy of Sciences of the USSR*, 41(4):333–338, 1987.

[Raz03] Alexander A Razborov. Quantum communication complexity of symmetric predicates. *Izvestiya: Mathematics*, 67(1):145, 2003.

[Rei11] Ben Reichardt. Reflections for quantum query algorithms. In Dana Randall, editor, *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2011, San Francisco, California, USA, January 23-25, 2011*, pages 560–569. SIAM, 2011.

[Ros61] Frank Rosenblatt. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, Cornell Aeronautical Lab Inc Buffalo NY, 1961.

[RS10] Alexander A Razborov and Alexander A Sherstov. The sign-rank of AC^0. *SIAM Journal on Computing*, 39(5):1833–1855, 2010.

[RST15] Benjamin Rossman, Rocco A. Servedio, and Li-Yang Tan. Complexity theory column 89: The polynomial hierarchy, random oracles, and boolean circuits. *SIGACT News*, 46(4):50–68, 2015.

[She08] Alexander A Sherstov. Halfspace matrices. *Computational Complexity*, 17(2):149–178, 2008.

[She09] Alexander A. Sherstov. Separating AC^0 from depth-2 majority circuits. *SIAM Journal on Computing*, 38(6):2113–2129, 2009.

[She11a] Alexander A Sherstov. The pattern matrix method. *SIAM Journal on Computing*, 40(6):1969–2000, 2011.

[She11b] Alexander A. Sherstov. The unbounded-error communication complexity of symmetric functions. *Comb.*, 31(5):583–614, 2011. Preliminary version in *FOCS* 2008.

[She12a] Alexander A Sherstov. Making polynomials robust to noise. In *Symposium on Theory of Computing*, pages 747–758, 2012. Journal version in Theory of Computing, 2013.

[She12b] Alexander A Sherstov. Strong direct product theorems for quantum communication and query complexity. *SIAM Journal on Computing*, 41(5):1122–1165, 2012.

[She13a] Alexander A Sherstov. Approximating the AND-OR tree. *Theory of Computing*, 9(1):653–663, 2013.

[She13b] Alexander A Sherstov. The intersection of two halfspaces has high threshold degree. *SIAM Journal on Computing*, 42(6):2329–2374, 2013.

[She13c] Alexander A Sherstov. Optimal bounds for sign-representing the intersection of two halfspaces by polynomials. *Combinatorica*, 33(1):73–96, 2013.

[She14] Alexander A Sherstov. Communication lower bounds using directional derivatives. *Journal of the ACM (JACM)*, 61(6):1–71, 2014.

[She18a] Alexander A Sherstov. Algorithmic polynomials. In *Symposium on Theory of Computing*, pages 311–324, 2018.

[She18b] Alexander A Sherstov. Breaking the Minsky–Papert barrier for constant-depth circuits. *SIAM Journal on Computing*, 47(5):1809–1857, 2018.

[She18c] Alexander A Sherstov. On multiparty communication with large versus unbounded error. *Theory of Computing*, 14(1):1–17, 2018.

[She18d] Alexander A Sherstov. The power of asymmetry in constant-depth circuits. *SIAM Journal on Computing*, 47(6):2362–2434, 2018.

[She21] Alexander A Sherstov. The hardest halfspace. *computational complexity*, 30(2):1–85, 2021.

[She22] Alexander A Sherstov. The approximate degree of dnf and cnf formulas. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1194–1207, 2022.

[Špa08] Robert Špalek. A dual polynomial for OR. *arXiv preprint arXiv:0803.4516*, 2008.

[ST19] Alexander A Sherstov and Justin Thaler. Vanishing-error approximate degree and qma complexity. *arXiv preprint arXiv:1909.07498*, 2019.

[SW19] Alexander A Sherstov and Pei Wu. Near-optimal lower bounds on the threshold degree and sign-rank of AC^0. In *Symposium on Theory of Computing*, pages 401–412, 2019.

[SZ09] Yaoyun Shi and Yufan Zhu. Quantum communication complexity of block-composed functions. *Quantum Information & Computation*, 9(5):444–460, 2009.

[Tal17] Avishay Tal. Formula lower bounds via the quantum method. In *Symposium on Theory of Computing*, pages 1256–1268, 2017.

[Tha16] Justin Thaler. Lower bounds for the approximate degree of block-composed functions. In *43rd International Colloquium on Automata, Languages, and Programming (ICALP 2016)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

[TUV12] Justin Thaler, Jonathan Ullman, and Salil Vadhan. Faster algorithms for privately releasing marginals. In *International Colloquium on Automata, Languages, and Programming*, pages 810–821, 2012.

[Val84] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[Vaz01] Vijay V Vazirani. *Approximation algorithms*, volume 1. Springer, 2001.

[Vya03] Mikhail Vyalyi. QMA= PP implies that PP contains PH. In *ECCCTR: Electronic Colloquium on Computational Complexity, technical reports*. Citeseer, 2003.

[Yao93] A Chi-Chih Yao. Quantum circuit complexity. In *Proceedings of 1993 IEEE 34th Annual Foundations of Computer Science*, pages 352–361. IEEE, 1993.

[Zha15] Mark Zhandry. A note on the quantum collision and set equality problems. *Quantum Information & Computation*, 15(7&8):557–567, 2015.