Electronic Colloquium on Computational Complexity, Report No. 16 (2008)

The Sign-Rank of AC^0

A A. R *
IAS, School of Math. & Steklov Mathematical Institute, razborov@ias.edu

A A. S †
Univ. of Texas at Austin, Dept. of Comp. Sciences, sherstov@cs.utexas.edu

February 25, 2008

Abstract

The *sign-rank* of a matrix $A = [A_{ij}]$ with ± 1 entries is the least rank of a real matrix $B = [B_{ij}]$ with $A_{ij}B_{ij} > 0$ for all i, j. We obtain the first exponential lower bound on the sign-rank of a function in AC^0. Namely, let $f(x, y) = \bigwedge_{i=1}^{m} \bigvee_{j=1}^{m^2}(x_{ij} \wedge y_{ij})$. We show that the matrix $[f(x, y)]_{x,y}$ has sign-rank $2^{\Omega(m)}$. This in particular implies that $\Sigma_2^{cc} \not\subseteq \mathsf{UPP}^{cc}$, which solves a long-standing open problem posed by Babai, Frankl, and Simon (1986).

Our result additionally implies a lower bound in learning theory. Specifically, let $\phi_1, \ldots, \phi_r : \{0,1\}^n \to \mathbb{R}$ be functions such that every DNF formula $f : \{0,1\}^n \to \{-1,+1\}$ of polynomial size has the representation $f \equiv \operatorname{sign}(a_1\phi_1 + \cdots + a_r\phi_r)$ for some reals $a_1, \ldots, a_r$. We prove that then $r \geqslant 2^{\Omega(n^{1/3})}$, which essentially matches an upper bound of $2^{\tilde{O}(n^{1/3})}$ due to Klivans and Servedio (2001).

Finally, our work yields the first exponential lower bound on the size of *threshold-of-majority* circuits computing a function in AC^0. This substantially generalizes and strengthens the results of Krause and Pudlák (1997).

*Supported by NSF grant ITR-0324906 and by the Russian Foundation for Basic Research.

†Part of this work was done during the author's visit to the Institute for Advanced Study, Princeton, NJ.

ISSN 1433-8092

1 Introduction

The *sign-rank* of a real matrix $A = [A_{ij}]$ with nonzero entries is the least rank of a matrix $B = [B_{ij}]$ with $A_{ij}B_{ij} > 0$ for all i, j. In other words, the sign-rank measures the stability of the rank of A as its entries undergo arbitrary sign-preserving perturbations. This fundamental notion has been studied in contexts as diverse as matrix analysis, communication complexity, circuit complexity, and learning theory [3, 5, 12, 13, 23, 29, 38, 45, 48]. We will give a detailed overview of these applications shortly as they pertain to our work.

Despite its importance, progress in understanding the sign-rank has been slow and difficult. Indeed, we are aware of only a few nontrivial results on this subject. Alon et al. [3] obtained strong lower bounds on the sign-rank of random matrices. In a breakthrough result, Forster [12] proved strong lower bounds on the sign-rank of Hadamard matrices (more generally, Forster's result applies to any ± 1 matrix with small spectral norm). Several extensions and refinements of Forster's method were proposed in subsequent work [13, 14, 29]. More recently, Sherstov [48] obtained near-tight estimates of the sign-rank for all matrices of the form $[D(|x \wedge y|)]_{x,y}$, where $D : \{0, 1, \ldots, n\} \to \{-1, +1\}$ is given and x, y range over $\{0, 1\}^n$.

This paper focuses on AC^0, a prominent class whose sign-rank has seen no progress in previous work. For notational convenience, we view Boolean functions as mappings into $\{-1, +1\}$, as opposed to the usual range $\{0, 1\}$. The central objective of our study is to estimate the maximum sign-rank of a matrix $[f(x, y)]_{x,y}$, where $f : \{0, 1\}^n \times \{0, 1\}^n \to \{-1, +1\}$ a function in AC^0. An obvious upper bound is 2^n, while the best lower bound prior to this paper was quasipolynomial.[1] Our main result considerably tightens the gap by improving the lower bound to $2^{\Omega(n^{1/3})}$:

Theorem 1.1 (Main result). *Let* $f_m(x, y) = \bigwedge_{i=1}^{m} \bigvee_{j=1}^{m^2} (x_{ij} \wedge y_{ij})$. *Then the matrix* $[f_m(x, y)]_{x,y}$ *has sign-rank* $2^{\Omega(m)}$.

It is not difficult to show that the matrix in Theorem 1.1 has sign-rank $2^{O(m \log m)}$, i.e., the lower bound we prove is almost tight. (See Remark 6.1 for details.) Moreover, it is essential that the circuit in question has depth at least 3: one readily verifies that AC^0 circuits of depth 1 or 2 lead to at most polynomial sign-rank.

Our main result states that AC^0 contains matrices whose rank is rather stable in that it cannot be reduced below $2^{\Theta(n^{1/3})}$ by any sign-preserving changes to the matrix entries. We proceed to discuss applications of this fact to communication complexity, learning theory, and circuits.

[1] The quasipolynomial lower bound is immediate from Forster's work [12] and the fact that AC^0 can compute 2 on $\log^c n$ variables, for every constant $c > 1$.

1.1 Communication Complexity

The study of the sign-rank is synonymous with the study of *unbounded-error communication complexity,* a rich model introduced by Paturi and Simon [38]. Fix a function $f : X \times Y \to \{0,1\}$, where X and Y are some finite sets. Alice receives an input $x \in X$, Bob receives $y \in Y$, and their objective is to compute $f(x,y)$ with minimal communication. The two parties each have an unlimited private source of random bits which they can use in deciding what messages to send. Their protocol is said to *compute* f if on every input (x,y), the output is correct with probability greater than $1/2$. The *cost* of a protocol is the worst-case number of bits exchanged on any input (x,y). The *unbounded-error communication complexity* of f, denoted $U(f)$, is the least cost of a protocol that computes f.

The unbounded-error model occupies a special place in the study of communication because it is more powerful than almost any other standard model (deterministic, nondeterministic, randomized, quantum with or without entanglement). More precisely, the unbounded-error complexity $U(f)$ can be only negligibly greater than the complexity of f in any of these models—and often, $U(f)$ is exponentially smaller. We defer exact quantitative statements to Appendix A. The power of the unbounded-error model resides in its very liberal acceptance criterion: it suffices to produce the correct output with probability even slightly greater than $1/2$ (say, by an exponentially small amount). This contrasts with all other models, where the correct output is expected with probability at least $2/3$.

Also unlike other models of communication, the unbounded-error model has an exact matrix-analytic formulation. Let $f : X \times Y \to \{0,1\}$ be a given function and $M = [(-1)^{f(x,y)}]_{x \in X,\, y \in Y}$ its communication matrix. Paturi and Simon [38] showed that

$$U(f) \;=\; \log \text{sign-rank}(M) \;\pm\; O(1).$$

In other words, unbounded-error complexity and sign-rank are essentially equivalent notions. In this light, our main result gives the first polynomial lower bound on the unbounded-error complexity of AC^0:

Corollary 1.2 (Unbounded-error communication complexity of AC^0). *Let* $f_m(x,y) = \bigwedge_{i=1}^{m} \bigvee_{j=1}^{m^2} (x_{ij} \wedge y_{ij})$. *Then* $U(f_m) = \Omega(m)$.

Corollary 1.2 also solves a long-standing problem in communication complexity. Specifically, the only models for which simulations in the unbounded-error model were unknown had been developed in the seminal paper by Babai, Frankl, and Simon [4]. These models are based on alternation and mimicry classes PH and PSPACE, and Babai et al. asked [4, last par. on p. 345] whether $\Sigma_2^{cc} \subseteq \mathsf{UPP}^{cc}$. Forster [12] made substantial progress on this question, proving that

$\mathsf{PSPACE}^{cc} \not\subseteq \mathsf{UPP}^{cc}$. We resolve the original question completely: Corollary 1.2 immediately implies that $\Sigma_2^{cc} \not\subseteq \mathsf{UPP}^{cc}$.

1.2 Learning Theory

In a seminal paper [52], Valiant formulated the *probably approximately correct* (PAC) model of learning, now the primary model in computational learning theory. Research has shown that PAC learning is surprisingly difficult. (By "PAC learning," we shall always mean PAC learning under arbitrary distributions.) Indeed, the learning problem remains unsolved for such natural concept classes as DNF formulas of polynomial size and intersections of two halfspaces, whereas hardness results and lower bounds are abundant [11, 20, 22, 24–26].

One concept class for which efficient PAC learning algorithms are available is the class of *halfspaces*, i.e., functions $f : \mathbb{R}^n \to \{-1, +1\}$ representable as $f(x) \equiv \operatorname{sign}(a_1 x_1 + \cdots + a_n x_n - \theta)$ for some reals $a_1, \ldots, a_n, \theta$. Halfspaces constitute one of the most studied classes in computational learning theory [6,31,35,44] and a major success story of the field. Indeed, a significant part of computational learning theory attempts to learn rich concept classes by reducing them to halfspaces. The reduction works as follows. Let C be a given *concept class*, i.e., a set of Boolean functions $\{0,1\}^n \to \{-1,+1\}$. One seeks functions $\phi_1, \ldots, \phi_r : \{0,1\}^n \to \mathbb{R}$ such that every $f \in C$ has the representation $f(x) \equiv \operatorname{sign}(a_1\phi_1(x) + \cdots + a_r\phi_r(x))$ for some reals $a_1, \ldots, a_r$. This process is technically described as *embedding C in halfspaces of dimension r*. Once this is accomplished, C can clearly be learned in time polynomial in n and r by any halfspace-learning algorithm.

For this approach to be practical, the number r of real functions needs to be reasonable (ideally, polynomial in n). It is therefore of interest to determine what natural concept classes can be embedded in halfspaces of low dimension [5, 25]. For brevity, we refer to the smallest dimension of such a representation as the *dimension complexity* of a given class. Formally, the dimension complexity $\operatorname{dc}(C)$ of a given class C of functions $\{0,1\}^n \to \{-1,+1\}$ is the least r for which there exist $\phi_1, \ldots, \phi_r : \{0,1\}^n \to \mathbb{R}$ such that every $f \in C$ is expressible as $f(x) \equiv \operatorname{sign}(a_1\phi_1(x) + \cdots + a_r\phi_r(x))$ for some reals $a_1, \ldots, a_r$. To relate this discussion to sign-rank, let $M_C = [f(x)]_{f \in C,\, x \in \{0,1\}^n}$ be the characteristic matrix of C. A moment's reflection reveals that $\operatorname{dc}(C) = \operatorname{sign\text{-}rank}(M_C)$, i.e., the dimension complexity of a concept class is precisely the sign-rank of its characteristic matrix. Indeed, the term "dimension complexity" has been used interchangeably with sign-rank in the recent literature [45, 50], which does not lead to confusion since concept classes are naturally identified with their characteristic matrices.

Thus, the study of sign-rank yields nontrivial PAC learning algorithms. In particular, the current fastest algorithm for learning polynomial-size DNF for-

mulas, due to Klivans and Servedio [23], was obtained precisely by placing an upper bound of $2^{\tilde{O}(n^{1/3})}$ on the dimension complexity of that concept class, with the functions ϕ_i corresponding to the monomials of degree up to $\tilde{O}(n^{1/3})$.

Klivans and Servedio also observed that their $2^{\tilde{O}(n^{1/3})}$ upper bound is best possible when the functions ϕ_i are taken to be the monomials up to a given degree. Our work gives a far-reaching generalization of the latter observation: we prove the same lower bound without assuming anything whatsoever about the embedding functions ϕ_i. That is, we have:

Corollary 1.3 (Dimension complexity of DNF). *Let C be the set of all read-once (hence, linear-size) DNF formulas $f : \{0,1\}^n \to \{-1,+1\}$. Then C has dimension complexity $2^{\Omega(n^{1/3})}$.*

Proof. Let $f_m(x,y)$ be the function from Theorem 1.1, where $m = \lfloor n^{1/3} \rfloor$. Then for any fixed y, the resulting function $f_y(x) = \neg f_m(x,y)$ is a read-once DNF formula. □

Learning polynomial-size DNF formulas was the original challenge posed in Valiant's paper [52]. More than twenty years later, this challenge remains a central open problem in computational learning theory despite active research [7, 23, 51]. To account for this lack of progress, several hardness results have been obtained based on complexity-theoretic assumptions [2,22]. Corollary 1.3 complements that line of work by exhibiting a concrete *structural* barrier to the efficient learning of DNF formulas. In particular, it rules out a $2^{o(n^{1/3})}$-time learning algorithm based on dimension complexity.

While restricted, the dimension-complexity paradigm is quite rich and captures many efficient PAC learning algorithm designed to date, with the notable exception of learning low-degree polynomials over GF(p). It is also worth noting [21, p. 124] that an unconditional superpolynomial lower bound for learning polynomial-size DNF formulas in the *standard* PAC model would imply that $\mathsf{P} \neq \mathsf{NP}$; thus, such a result is well beyond the reach of the current techniques.

1.3 Threshold Circuits

Recall that a *threshold gate* g with Boolean inputs $x_1, \ldots, x_n$ is a function of the form $g(x) = \text{sign}(a_1 x_1 + \cdots + a_n x_n - \theta)$, for some fixed reals $a_1, \ldots, a_n, \theta$. Thus, a threshold gate generalizes the familiar *majority* gate. A major unsolved problem in computational complexity is to exhibit a Boolean function that requires a depth-2 threshold circuit of superpolynomial size.

Communication complexity has been crucial to the progress on this problem. Through randomized communication complexity, many explicit functions have

been found [15, 16, 33, 46, 47] that require *majority-of-threshold* circuits of exponential size. This solves an important case of the general problem. Lower bounds for the unbounded-error model (or, equivalently, on the sign-rank) cover another important case [13], that of *threshold-of-majority* circuits. The following statement is immediate from our main result, in view of the work by Forster et al. [13, Lem. 5]:

Corollary 1.4 (Threshold circuits). *Let* $f_m(x, y) = \bigwedge_{i=1}^{m} \bigvee_{j=1}^{m^2}(x_{ij} \wedge y_{ij})$. *Let* C *be a depth-*2 *threshold circuit, with arbitrary weights at the top gate and integer weights of absolute value* $\leqslant w$ *at the bottom gates. If* C *computes* f_m, *then it has size* $2^{\Omega(m)}/w$.

This is the first exponential lower bound for *threshold-of-majority* circuits computing a function in AC^0. It substantially generalizes and strengthens an earlier result of Krause and Pudlák [27, Thm. 2], who proved an exponential lower bound for *threshold-of-MOD*$_r$ circuits (for any constant $r \geqslant 2$) computing a function in AC^0. Our work also complements exponential lower bounds for *majority-of-threshold* circuits computing functions in AC^0, obtained recently by Buhrman et al. [8] and Sherstov [47, 49].

Theorem 1.1 immediately implies lower bounds for other classes of depth-2 circuits, e.g., those with a threshold gate receiving inputs from arbitrary symmetric gates. Rather than formulate these statements as theorems, we refer the reader to the work by Forster et al. [13, §6] for details on how the sign-rank relates to those other circuit models.

1.4 Our Proof and Techniques

Figure 1 illustrates the main components of our proof. A starting point in our study is an elegant result due to Minsky and Papert [31], who constructed a linear-size DNF formula that cannot be sign-represented by polynomials of low degree. Minsky and Papert's observation has played an important role in several other works [27, 36, 47, 49].

Second, we revisit a fundamental technique from approximation theory, the *interpolation bound*, which bounds a univariate polynomial $p \in P_d$ on an interval based on the values of p at $d + 1$ distinct points. By combining the interpolation bound with an adapted version of Minsky and Papert's argument, we establish a key intermediate result (Lemma 3.3). This result concerns multivariate polynomials that have nonnegligible agreement with the Minsky-Papert function and constrains their behavior on a large fraction of the inputs.

We proceed by deriving a Fourier-theoretic property common to all low-degree multivariate polynomials on $\{0, 1\}^n$: we show that their values on $\{0, 1\}^n$ can

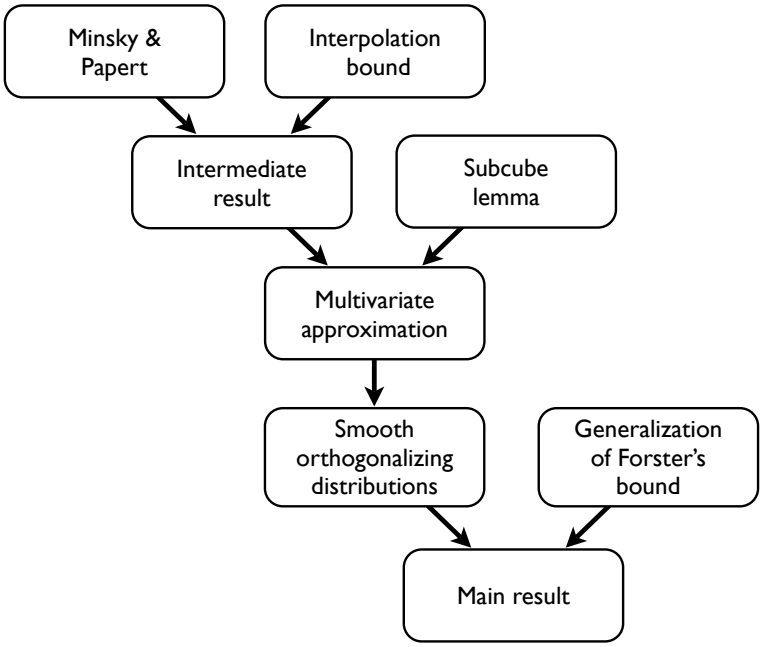

Figure 1: Proof outline.

be conveniently bounded in terms of their behavior on certain small subcubes (Lemma 3.2). In light of this Fourier-theoretic observation, our intermediate result on multivariate polynomials takes on a much stronger form. Namely, we prove that multivariate polynomials with any nontrivial agreement with the Minsky-Papert function are highly constrained *throughout* the hypercube (Theorem 3.4). With some additional work in Section 4, we are able to deduce the existence of a smooth distribution on $\{0,1\}^n$ with respect to which the Minsky-Papert function is orthogonal to all low-degree polynomials. From this, we obtain our main result by suitably modifying the analysis in Forster's fundamental paper on sign-rank (see Section 5 and Appendix B).

The techniques of our proof seem to be of independent interest. Multivariate polynomials on $\{0,1\}^n$ arise frequently in the complexity literature and pose a considerable analytical challenge. A solution that we introduce is to project a multivariate polynomial in several ways to univariate polynomials, study the latter objects, and recombine the results using Fourier-theoretic tools (see Section 3). To our knowledge, this approach is novel and shows promise in more general contexts.

2 Preliminaries

All Boolean functions in this paper are represented as mappings $\{0,1\}^n \to \{-1,+1\}$, where -1 corresponds to "true." For $x \in \{0,1\}^n$, we define $|x| = x_1 + x_2 + \cdots + x_n$. The symbol P_d stands for the set of all univariate real polynomials of degree up to d. By the degree of a *multivariate* polynomial, we will always mean its total degree, i.e., the greatest total degree of any monomial. The notation $[n]$ refers to the set $\{1, 2, \ldots, n\}$. Set membership notation, when used in the subscript of an expectation operator, means that the expectation is taken with respect to the *uniformly random* choice of an element from the indicated set.

2.1 Matrix Analysis

The symbol $\mathbb{R}^{m \times n}$ refers to the family of all $m \times n$ matrices with real entries. The (i, j)th entry of a matrix A is denoted by A_{ij}. We frequently use "generic-entry" notation to specify a matrix succinctly: we write $A = [F(i, j)]_{i,j}$ to mean that the (i, j)th entry of A is given by the expression $F(i, j)$. In most matrices that arise in this work, the exact ordering of the columns (and rows) is irrelevant. In such cases we describe a matrix by the notation $[F(i, j)]_{i \in I,\, j \in J}$, where I and J are some index sets.

Let $A = [A_{ij}] \in \mathbb{R}^{m \times n}$ be given. We let $\|A\|_\infty \stackrel{\text{def}}{=} \max_{i,j} |A_{ij}|$ and denote the singular values of A by $\sigma_1(A) \geqslant \sigma_2(A) \geqslant \cdots \geqslant \sigma_{\min\{m,n\}}(A) \geqslant 0$. The notation $\|\cdot\|_2$ refers to the Euclidean norm of vectors. Recall that the *spectral norm*, *trace norm*, and *Frobenius norm* of A are given by

$$\begin{aligned} \|A\| &= \max_{x \in \mathbb{R}^n,\, \|x\|_2 = 1} \|Ax\|_2 = \sigma_1(A), \\ \|A\|_\Sigma &= \sum \sigma_i(A), \\ \|A\|_{\mathrm{F}} &= \sqrt{\sum A_{ij}^2} = \sqrt{\sum \sigma_i(A)^2}. \end{aligned}$$

An essential property of these norms is their invariance under orthogonal transformations on the left and on the right, which incidentally explains the alternative expressions for the spectral and Frobenius norms given above. The following relationship follows at once by the Cauchy-Swartz inequality:

$$\|A\|_\Sigma \leqslant \|A\|_{\mathrm{F}} \sqrt{\mathrm{rank}(A)} \qquad (A \in \mathbb{R}^{m \times n}). \tag{2.1}$$

For $A, B \in \mathbb{R}^{m \times n}$, we write $\langle A, B \rangle \stackrel{\text{def}}{=} \sum_{i,j} A_{ij} B_{ij}$. A useful consequence of the singular value decomposition is:

$$\langle A, B \rangle \leqslant \|A\| \, \|B\|_\Sigma \qquad (A, B \in \mathbb{R}^{m \times n}). \tag{2.2}$$

The *Hadamard product* of A and B is the matrix $A \circ B = [A_{ij}B_{ij}]$. The symbol J stands for the all-ones matrix, whose dimensions will be apparent from the context. The notation $A \geqslant 0$ means that all the entries in A are nonnegative. The shorthand $A \neq 0$ means as usual that A is not the zero matrix.

2.2 The Fourier Transform over $\mathbb{Z}_2^n$

Consider the vector space of functions $\{0,1\}^n \to \mathbb{R}$, equipped with the inner product

$$\langle f, g \rangle \stackrel{\text{def}}{=} 2^{-n} \sum_{x \in \{0,1\}^n} f(x)g(x).$$

For $S \subseteq [n]$, define $\chi_S : \{0,1\}^n \to \{-1,+1\}$ by $\chi_S(x) = (-1)^{\sum_{i \in S} x_i}$. Then $\{\chi_S\}_{S \subseteq [n]}$ is an orthonormal basis for the inner product space in question. As a result, every function $f : \{0,1\}^n \to \mathbb{R}$ has a unique representation of the form

$$f(x) = \sum_{S \subseteq [n]} \hat{f}(S)\chi_S(x),$$

where $\hat{f}(S) \stackrel{\text{def}}{=} \langle f, \chi_S \rangle$. The reals $\hat{f}(S)$ are called the *Fourier coefficients of* f. The following fact is immediate from the definition of $\hat{f}(S)$:

Proposition 2.1. *Let $f : \{0,1\}^n \to \mathbb{R}$ be given. Then*

$$\max_{S \subseteq [n]} |\hat{f}(S)| \leqslant 2^{-n} \sum_{x \in \{0,1\}^n} |f(x)|.$$

2.3 Symmetric Functions

Let S_n denote the symmetric group on n elements. For $\sigma \in S_n$ and $x \in \{0,1\}^n$, we denote by σx the string $(x_{\sigma(1)}, \ldots, x_{\sigma(n)}) \in \{0,1\}^n$. A function $\phi : \{0,1\}^n \to \mathbb{R}$ is called *symmetric* if $\phi(x) = \phi(\sigma x)$ for every $x \in \{0,1\}^n$ and every $\sigma \in S_n$. Equivalently, ϕ is symmetric if $\phi(x)$ is uniquely determined by $|x|$. Observe that for every $\phi : \{0,1\}^n \to \mathbb{R}$ (symmetric or not), the derived function

$$\phi_{\text{sym}}(x) \stackrel{\text{def}}{=} \mathop{\mathbf{E}}_{\sigma \in S_n} \left[\phi(\sigma x)\right]$$

is symmetric. The symmetric functions on $\{0,1\}^n$ are intimately related to univariate polynomials, as demonstrated by Minsky and Papert's *symmetrization argument:*

Proposition 2.2 (Minsky & Papert [31]). *Let $\phi : \{0,1\}^n \to \mathbb{R}$ be representable by a real n-variate polynomial of degree r. Then there is a polynomial $p \in P_r$ with*

$$\mathop{\mathbf{E}}_{\sigma \in S_n}\Big[\phi(\sigma x)\Big] = p(|x|) \qquad \forall x \in \{0,1\}^n.$$

Minsky and Papert's observation has seen numerous uses in the literature [1,34,37]. We will need the following straightforward generalization.

Proposition 2.3. *Let $n_1, \ldots, n_k$ be positive integers with $n \stackrel{\text{def}}{=} n_1 + \cdots + n_k$. Let $\phi : \{0,1\}^n \to \mathbb{R}$ be representable by a real n-variate polynomial of degree r. Write $x \in \{0,1\}^n$ as $x = (x^{(1)}, \ldots, x^{(k)})$, where $x^{(i)} = (x_{n_1+\cdots+n_{i-1}+1}, \ldots, x_{n_1+\cdots+n_i})$. Then there is a polynomial p on $\mathbb{R}^k$ of degree at most r such that*

$$\mathop{\mathbf{E}}_{\sigma_1 \in S_1, \ldots, \sigma_k \in S_k}\Big[\phi(\sigma_1 x^{(1)}, \ldots, \sigma_k x^{(k)})\Big] = p\Big(|x^{(1)}|, \ldots, |x^{(k)}|\Big) \qquad \forall x \in \{0,1\}^n.$$

2.4 Sign-rank

The *sign-rank* of a real matrix $A = [A_{ij}]$ is the least rank of a matrix $B = [B_{ij}]$ such that $A_{ij}B_{ij} > 0$ for all i, j with $A_{ij} \neq 0$. (Note that this definition generalizes the one given above in the abstract and introduction, which applied only to matrices A with nonzero entries.) In a breakthrough result, Forster [12] proved the first nontrivial lower bound on the sign-rank of an explicit ± 1 matrix. The centerpiece of Forster's argument is the following theorem, which is a crucial starting point for our work.

Theorem 2.4 (Forster [12], implicit). *Let X, Y be finite sets and $M = [M_{xy}]_{x \in X, y \in Y}$ a real matrix ($M \neq 0$). Put $r = \text{sign-rank}(M)$. Then there is a matrix $R = [R_{xy}]_{x \in X, y \in Y}$ such that:*

$$\begin{aligned} \text{rank}(R) &= r, \\ M \circ R &\geqslant 0, \\ \|R\|_\infty &\leqslant 1, \\ \|R\|_{\text{F}} &= \sqrt{|X|\,|Y|/r}. \end{aligned}$$

Appendix B provides a detailed explanation of how Theorem 2.4 is implicit in Forster's work.

2.5 Pattern Matrices

Pattern matrices arose in earlier works by Sherstov [47, 49] and proved useful in obtaining strong lower bounds on communication complexity. Relevant definitions and results from [49] follow.

Let n and N be positive integers with $n \mid N$. Split $[N]$ into n contiguous blocks, with N/n elements each:

$$[N] = \left\{1, 2, \ldots, \frac{N}{n}\right\} \cup \left\{\frac{N}{n} + 1, \ldots, \frac{2N}{n}\right\} \cup \cdots \cup \left\{\frac{(n-1)N}{n} + 1, \ldots, N\right\}.$$

Let $\mathcal{V}(N, n)$ denote the family of subsets $V \subseteq [N]$ that have exactly one element from each of these blocks (in particular, $|V| = n$). Clearly, $|\mathcal{V}(N, n)| = (N/n)^n$. For a bit string $x \in \{0,1\}^N$ and a set $V \in \mathcal{V}(N, n)$, define the *projection of x onto V* by

$$x|_V \stackrel{\text{def}}{=} (x_{i_1}, x_{i_2}, \ldots, x_{i_n}) \in \{0,1\}^n,$$

where $i_1 < i_2 < \cdots < i_n$ are the elements of V.

Definition 2.5 (Pattern matrix). For $\phi : \{0,1\}^n \to \mathbb{R}$, the (N, n, ϕ)-*pattern matrix* is the real matrix A given by

$$A = \Big[\phi(x|_V \oplus w)\Big]_{x \in \{0,1\}^N,\, (V,w) \in \mathcal{V}(N,n) \times \{0,1\}^n}.$$

In words, A is the matrix of size 2^N by $(N/n)^n 2^n$ whose rows are indexed by strings $x \in \{0,1\}^N$, whose columns are indexed by pairs $(V, w) \in \mathcal{V}(N, n) \times \{0,1\}^n$, and whose entries are given by $A_{x,(V,w)} = \phi(x|_V \oplus w)$.

The logic behind the term "pattern matrix" is as follows: a mosaic arises from repetitions of a pattern in the same way that A arises from applications of ϕ to various subsets of the variables. We will need the following expression for the spectral norm of a pattern matrix.

Theorem 2.6 (Sherstov [49, Thm. 4.3]). *Let $\phi : \{0,1\}^n \to \mathbb{R}$ be given. Let A be the (N, n, ϕ)-pattern matrix. Then*

$$\|A\| = \sqrt{2^{N+n}\left(\frac{N}{n}\right)^n} \max_{S \subseteq [n]} \left\{|\hat{\phi}(S)| \left(\frac{n}{N}\right)^{|S|/2}\right\}.$$

3 A Result on Multivariate Approximation

The purpose of this section is to establish a certain property of low-degree polynomials on $\mathbb{R}^m$ (Theorem 3.4). This property is the backbone of our main proof.

A starting point in our discussion is an *interpolation bound*, i.e., a bound on the values of a polynomial on an interval given its values on a finite set of points. Results of this general form arise routinely in approximation theory. To prove the specific statement of interest to us, we follow the classical technique of interpolating the polynomial at strategically chosen nodes. For other uses of this technique, see Cheney [9, §7, Lem. 1] and Rivlin [43, Thm. 3.9].

Lemma 3.1 (Interpolation bound). *Let $I \subset \mathbb{R}$ be an interval of length L. Let p be a polynomial of degree $d \leqslant L$ such that*

$$|p(x_i)| \leqslant 1 \qquad (i = 0, 1, \ldots, d),$$

where $x_0, x_1, \ldots, x_d \in I$ are some points with pairwise distances at least 1*. Then*

$$\max_{x \in I} |p(x)| \;\leqslant\; 2^d \binom{L}{d}.$$

Proof. Without loss of generality, assume that $x_0 < x_1 < \cdots < x_d$. Fix $x \in I$. For any $k \in \{0, 1, \ldots, d\}$, we have:

$$\prod_{\substack{i=0 \\ i \neq k}}^{d} |x - x_i| \leqslant L(L-1)\cdots(L-d+1)$$

and (since $|x_i - x_k| \geqslant |i - k|$)

$$\prod_{\substack{i=0 \\ i \neq k}}^{d} |x_k - x_i| \geqslant k!(d-k)!.$$

Therefore,

$$\prod_{\substack{i=0 \\ i \neq k}}^{d} \frac{|x - x_i|}{|x_k - x_i|} \leqslant \frac{L(L-1)\cdots(L-d+1)}{k!(d-k)!} = \binom{L}{d}\binom{d}{k}.$$

It remains to substitute this estimate in the Lagrange interpolation formula:

$$|p(x)| = \left| \sum_{k=0}^{d} p(x_k) \prod_{\substack{i=0 \\ i \neq k}}^{d} \frac{x - x_i}{x_k - x_i} \right| \leqslant \binom{L}{d} \sum_{k=0}^{d} \binom{d}{k} = 2^d \binom{L}{d}. \qquad \square$$

We now establish another auxiliary fact. It provides a convenient means to bound a function whose Fourier transform is supported on low-order characters, in terms of its behavior on low-weight inputs.

Lemma 3.2. *Let k be an integer, $0 \leqslant k \leqslant n-1$. Let $f : \{0,1\}^n \to \mathbb{R}$ be given with $\hat{f}(S) = 0$ for $|S| > k$. Then*

$$|f(1^n)| \leqslant 2^k \binom{n}{k} \max_{|x| \leqslant k} |f(x)|.$$

Proof. Define the symmetric function $g : \{0,1\}^n \to \mathbb{R}$ by $g(x) = \chi_{[n]}(x)p(|x|)$, where

$$p(t) \stackrel{\text{def}}{=} \prod_{k<i<n} \frac{t-i}{n-i}.$$

The following properties of g are immediate:

$$g(1^n) = (-1)^n, \tag{3.1}$$

$$g(x) = 0 \qquad (k < |x| < n), \tag{3.2}$$

$$\hat{g}(S) = 0 \qquad (|S| \leqslant k). \tag{3.3}$$

Furthermore,

$$\sum_{|x| \leqslant k} |g(x)| = \sum_{t=0}^{k} \binom{n}{t} |p(t)| = \sum_{t=0}^{k} \binom{n}{t} \binom{n-t-1}{n-k-1} \leqslant 2^k \binom{n}{k}. \tag{3.4}$$

We are now prepared to analyze f. By (3.3),

$$\sum_{x \in \{0,1\}^n} f(x)g(x) = 0. \tag{3.5}$$

On the other hand, (3.1) and (3.2) show that

$$\sum_{x \in \{0,1\}^n} f(x)g(x) = (-1)^n f(1^n) + \sum_{|x| \leqslant k} f(x)g(x). \tag{3.6}$$

The lemma follows at once from (3.4)–(3.6). □

We are now in a position to study the approximation problem of interest to us. Define the sets

$$Z = \{0, 1, 2, \ldots, 4m^2\}^m, \qquad Z^+ = \{1, 2, \ldots, 4m^2\}^m.$$

Define $F : Z \to \{-1, +1\}$ by

$$F(z) = \begin{cases} -1 & \text{if } x \in Z^+, \\ 1 & \text{otherwise.} \end{cases}$$

For $u, z \in Z$, let $\Delta(u, z) = |\{i : u_i \neq z_i\}|$ be the ordinary Hamming distance. We shall prove the following intermediate result, inspired by Minsky and Papert's analysis [31] of the threshold degree of CNF formulas.

Lemma 3.3. *Let Q be a degree-d real polynomial in m variables, where $d \leqslant m/3$. Assume that*

$$F(z)Q(z) \geqslant -1 \qquad (z \in Z). \tag{3.7}$$

Then $|Q(z)| \leqslant 2^{m+2d}$ at every point $z \in Z^+$ with $\Delta(u, z) < m/3$, where $u = (1^2, 3^2, 5^2, \ldots, (2m-1)^2) \in Z^+$.

Proof. Fix $z \in Z^+$ with $\Delta(u, z) < m/3$. Define $p \in P_{2d}$ by

$$p(t) = Q(p_1(t), p_2(t), \ldots, p_m(t)),$$

where

$$p_i(t) = \begin{cases} (t - 2i + 1)^2 & \text{if } z_i = u_i \text{ (equivalently, } z_i = (2i-1)^2), \\ z_i & \text{otherwise.} \end{cases}$$

Letting $S = \{i : u_i = z_i\}$, inequality (3.7) implies that

$$p(2i - 1) \geqslant -1 \qquad (i \in S), \tag{3.8}$$

$$p(2i) \leqslant 1 \qquad (i = 0, 1, \ldots, m). \tag{3.9}$$

Claim 3.3.1. *Let $i \in S$. Then $|p(\xi)| \leqslant 1$ for some $\xi \in [2i - 2, 2i - 1]$.*

Proof. The claim is trivial if p vanishes at some point in $[2i - 2, 2i - 1]$. In the contrary case, p maintains the same sign throughout this interval. As a result, (3.8) and (3.9) show that $\min\{|p(2i - 2)|, |p(2i - 1)|\} \leqslant 1$. □

Claim 3.3.1 provides $|S| > 2m/3 \geqslant 2d \geqslant \deg(p)$ points in $[0, m]$, with pairwise distances at least 1, at which p is bounded in absolute value by 1. By Lemma 3.1,

$$\max_{0 \leqslant t \leqslant m} |p(t)| \leqslant 2^{\deg(p)} \binom{m}{\deg(p)} \leqslant 2^{m+2d}.$$

This completes the proof since $Q(z) = p(0)$. □

Finally, we remove the restriction on $\Delta(u, z)$, thereby establishing the main result of this section:

Theorem 3.4. *Let Q be a degree-d real polynomial in m variables, where $d < m/3$. Assume that*

$$F(z)Q(z) \geqslant -1 \qquad (z \in Z).$$

Then

$$|Q(z)| \leqslant 8^m \qquad (z \in Z^+).$$

Proof. As before, put $u = (1^2, 3^2, 5^2, \ldots, (2m-1)^2)$. Fix $z \in Z^+$ and define the "interpolating" function $f : \{0,1\}^m \to \mathbb{R}$ by

$$f(x) = Q(x_1 z_1 + (1 - x_1)u_1, \ldots, x_m z_m + (1 - x_m)u_m).$$

In this notation, we know from Lemma 3.3 that $|f(x)| \leqslant 2^{m+2d}$ for every $x \in \{0,1\}^m$ with $|x| < m/3$, and our goal is to show that $|f(1^m)| \leqslant 8^m$. Since Q has degree d, the Fourier transform of f is supported on characters of order up to d. As a result,

$$\begin{aligned} |f(1^m)| &\leqslant 2^d \binom{m}{d} \max_{|x| \leqslant d} |f(x)| && \text{by Lemma 3.2} \\ &\leqslant 2^{m+3d} \binom{m}{d} && \text{by Lemma 3.3} \\ &\leqslant 8^m. \end{aligned}$$

□

4 A Smooth Orthogonalizing Distribution

An important concept in our work is that of an orthogonalizing distribution. Let $f : \{0,1\}^n \to \{-1,+1\}$ be given. A distribution μ on $\{0,1\}^n$ is *d-orthogonalizing* for f if

$$\mathop{\mathbf{E}}_{x \sim \mu} \left[f(x)\chi_S(x) \right] = 0 \qquad (|S| < d).$$

In words, a distribution μ is d-orthogonalizing for f if with respect to μ, the function f is orthogonal to every character or order less than d.

This section focuses on the following function from $\{0,1\}^{4m^3}$ to $\{-1,+1\}$:

$$\mathrm{MP}_m(x) = \bigwedge_{i=1}^{m} \bigvee_{j=1}^{4m^2} x_{i,j}.$$

(Recall that we interpret -1 as "true".) This function was originally studied by Minsky and Papert [31] and has played an important role in later works [27, 36, 47, 49]. An explicit m-orthogonalizing distribution for MP_m is known [47]. However, our main result requires a $\Theta(m)$-orthogonalizing distribution for MP_m that is additionally *smooth*, i.e., places substantial weight on all but a tiny fraction of the points, and the distribution given in [47] severely violates the latter property. Proving the existence of a distribution that is simultaneously $\Theta(m)$-orthogonalizing and smooth is the goal of this section (Theorem 4.1).

We will view an input $x \in \{0,1\}^n = \{0,1\}^{4m^3}$ to MP_m as composed of blocks: $x = (x^{(1)}, \ldots, x^{(m)})$, where the ith block is $x^{(i)} = (x_{i,1}, x_{i,2}, \ldots, x_{i,4m^2})$. The proof that is about to start refers to the sets Z, Z^+ and the function F as defined in Section 3.

Theorem 4.1. *There is a $\frac{1}{3}m$-orthogonalizing distribution μ for MP_m such that $\mu(x) \geqslant \frac{1}{2}8^{-m}\,2^{-n}$ for all inputs $x \in \{0,1\}^n$ with $\mathrm{MP}_m(x) = -1$.*

Proof. Let X be the set of all inputs with $\mathrm{MP}_m(x) = -1$, i.e.,

$$X = \{x \in \{0,1\}^n : x^{(1)} \neq 0, \quad \ldots, \quad x^{(m)} \neq 0\}.$$

It suffices to show that the following linear program has optimum at least $\frac{1}{2}8^{-m}$:

variables:	$\epsilon \geqslant 0$; $\quad \mu(x) \geqslant 0$ for $x \in \{0,1\}^n$	
maximize:	ϵ	
subject to:	$\displaystyle\sum_{x\in\{0,1\}^n} \mu(x)\mathrm{MP}_m(x)\chi_S(x) = 0$	for $\lvert S\rvert < m/3$,
	$\displaystyle\sum_{x\in\{0,1\}^n} \mu(x) \leqslant 1$,	
	$\mu(x) \geqslant \epsilon 2^{-n}$	for $x \in X$.

(LP1)

For $x \in \{0,1\}^n$, we let $z(x) = (|x^{(1)}|, \ldots, |x^{(m)}|)$; note that $\mathrm{MP}_m(x) = F(z(x))$. Since the function MP_m is invariant under the action of the group $S_{4m^2} \times \cdots \times S_{4m^2}$, in view of Proposition 2.3, the dual of (LP1) can be simplified as follows:

variables:	a polynomial Q on $\mathbb{R}^m$ of degree $< m/3$;	
	$\eta \geqslant 0$; $\quad \delta_z \geqslant 0$ for $z \in Z^+$	
minimize:	η	
subject to:	$\displaystyle\sum_{x\in X} \delta_{z(x)} \geqslant 2^n$,	
	$F(z)Q(z) \geqslant -\eta$	for $z \in Z$,
	$F(z)Q(z) \geqslant -\eta + \delta_z$	for $z \in Z^+$.

(LP2)

The programs are both feasible and therefore have the same finite optimum. Fix an optimal solution η, Q, δ_z to (LP2). For the sake of contradiction, assume that $\eta \leqslant \frac{1}{2}8^{-m}$. Then $|Q(z)| \leqslant \frac{1}{2}$ for each $z \in Z^+$, by Theorem 3.4. From the constraints of the third type in (LP2) we conclude that $\delta_z \leqslant \frac{1}{2} + \eta < 1$ ($z \in Z^+$). This contradicts the first constraint. Thus, the optimum of (LP1) and (LP2) is at least $\frac{1}{2}8^{-m}$. □

5 A Generalization of Forster's Bound

Using Theorem 2.4, Forster gave a simple proof of the following fundamental result [12, Thm. 2.2]: for any matrix $A = [A_{xy}]_{x \in X,\, y \in Y}$ with ± 1 entries,

$$\text{sign-rank}(A) \geqslant \frac{\sqrt{|X|\,|Y|}}{\|A\|}.$$

Forster et al. [13, Thm. 3] generalized this bound to arbitrary real matrices $A \neq 0$:

$$\text{sign-rank}(A) \geqslant \frac{\sqrt{|X|\,|Y|}}{\|A\|} \cdot \min_{x,y} |A_{xy}|. \tag{5.1}$$

Forster and Simon [14, §5] considered a different generalization, inspired by the notion of matrix rigidity. Let A be a given ± 1 matrix, and let $\tilde{A}$ be obtained from A by changing some h entries in an arbitrary fashion ($h < |X|\,|Y|$). Forster and Simon showed that

$$\text{sign-rank}(\tilde{A}) \geqslant \frac{\sqrt{|X|\,|Y|}}{\|A\| + 2\sqrt{h}}. \tag{5.2}$$

The above generalizations are not sufficient for our purposes. Before we can proceed, we need to prove the following "hybrid" bound, which combines the ideas of the previous work.

Theorem 5.1. *Let $A = [A_{xy}]_{x \in X,\, y \in Y}$ be a real matrix with $s = |X|\,|Y|$ entries ($A \neq 0$). Assume that all but h of the entries of A satisfy $|A_{xy}| \geqslant \gamma$, where h and $\gamma > 0$ are arbitrary parameters. Then*

$$\text{sign-rank}(A) \geqslant \frac{\gamma s}{\|A\|\sqrt{s} + \gamma h}.$$

Proof. Let r denote the sign-rank of A. Theorem 2.4 supplies a matrix $R = [R_{xy}]$ with

$$\text{rank}(R) = r, \tag{5.3}$$

$$A \circ R \geqslant 0, \tag{5.4}$$

$$\|R\|_\infty \leqslant 1, \tag{5.5}$$

$$\|R\|_{\text{F}} = \sqrt{s/r}. \tag{5.6}$$

The crux of the proof is to estimate $\langle A, R\rangle$ from below and above. On the one hand,

$$\begin{aligned}
\langle A, R\rangle &\geqslant \sum_{x,y:\,|A_{xy}|\geqslant\gamma} A_{xy}R_{xy} && \text{by (5.4)}\\
&\geqslant \gamma\left(\sum_{x,y}|R_{xy}| - h\right) && \text{by (5.4), (5.5)}\\
&\geqslant \gamma\|R\|_{\mathrm{F}}^2 - \gamma h && \text{by (5.5)}\\
&= \frac{\gamma s}{r} - \gamma h && \text{by (5.6).}
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\langle A, R\rangle &\leqslant \|A\| \cdot \|R\|_{\Sigma} && \text{by (2.2)}\\
&\leqslant \|A\| \cdot \|R\|_{\mathrm{F}}\ \sqrt{r} && \text{by (2.1), (5.3)}\\
&= \|A\|\sqrt{s} && \text{by (5.6).}
\end{aligned}$$

Comparing these lower and upper bounds on $\langle A, R\rangle$ yields the claimed estimate of $r = \text{sign-rank}(A)$. □

Remark 5.2. Using the method of Theorem 5.1, one can improve (5.2) to

$$\text{sign-rank}(\tilde{A}) \geqslant \frac{s}{\|A\|\sqrt{s} + 2h},$$

where $s = |X|\,|Y|$ as before. This improvement becomes significant for $h \ll s$.

6 Main Result

At last, we are in a position to prove the main result of this work.

Theorem 1.1 (Restated from p. 1). *Define* $f_m(x, y) = \bigwedge_{i=1}^{m}\bigvee_{j=1}^{m^2}(x_{ij} \wedge y_{ij})$. *Then the matrix* $[f_m(x, y)]_{x,y}$ *has sign-rank* $2^{\Omega(m)}$.

Proof. Let M be the (N, n, MP_m)-pattern matrix, where $N = 10^6 n$. Let P be the (N, n, μ)-pattern matrix, where μ is the distribution from Theorem 4.1. We are going to estimate the sign-rank of $M \circ P$.

By Theorem 4.1, all but a $2^{-\Omega(m^2)}$ fraction of the inputs $x \in \{0,1\}^n$ satisfy $\mu(x) \geqslant \frac{1}{2}8^{-m}\,2^{-n}$. As a result, all but a $2^{-\Omega(m^2)}$ fraction of the entries of $M \circ P$ are at least $\frac{1}{2}8^{-m}\,2^{-n}$ in absolute value. Theorem 5.1 at once implies that

$$\text{sign-rank}(M) \geqslant \text{sign-rank}(M \circ P) \geqslant \min\left\{\frac{8^{-m}\,2^{-n}\sqrt{s}}{4\,\|M \circ P\|},\ 2^{\Omega(m^2)}\right\}, \tag{6.1}$$

where $s = 2^{N+n}\left(\frac{N}{n}\right)^n$ denotes the number of entries in $M \circ P$.

We proceed to bound the spectral norm of $M \circ P$. Note first that $M \circ P$ is the (N, n, ϕ)-pattern matrix, where $\phi : \{0,1\}^n \to \mathbb{R}$ is given by $\phi(x) = \mathrm{MP}_m(x)\mu(x)$. Since μ is a $\frac{1}{3}m$-orthogonalizing distribution for MP_m, we have

$$\hat{\phi}(S) = 0 \qquad \text{for } |S| < \tfrac{1}{3}m. \tag{6.2}$$

Since $\sum_{x\in\{0,1\}^n} |\phi(x)| = 1$, Proposition 2.1 shows that

$$|\hat{\phi}(S)| \leqslant 2^{-n} \qquad \text{for each } S \subseteq [n]. \tag{6.3}$$

Theorem 2.6 implies, in view of (6.2) and (6.3), that

$$\|M \circ P\| \leqslant \sqrt{s} \cdot 2^{-n}\left(\frac{N}{n}\right)^{-m/6} = 10^{-m}\, 2^{-n}\, \sqrt{s}.$$

Substituting this estimate in (6.1) shows that the sign-rank of M is at least $2^{\Omega(m)}$. It remains to note that M is a submatrix of $[f_{cm}(x,y)]_{x,y}$, where $c = 4N/n = 4{\cdot}10^6$. □

Remark 6.1. The lower bound in Theorem 1.1 is essentially optimal. To see this, note that the matrix $[f_m(x,y)]_{x,y}$ has the same sign pattern as the matrix

$$R = \left[\frac{1}{2} - \prod_{i=1}^{m}\left(\sum_{j=1}^{m^2} x_{ij}y_{ij}\right)\right]_{x,y}.$$

Therefore, the sign-rank of $[f_m(x,y)]_{x,y}$ does not exceed

$$\operatorname{rank}(R) \leqslant 1 + m^{2m} = 2^{O(m\log m)}.$$

7 Open Problems

Our work is closely related to several natural and important problems. The first is a well-known and challenging open problem in complexity theory. Are there matrices computable in AC^0 that have low spectral norm? More precisely, does one have $\|[f(x,y)]_{x\in X, y\in Y}\| \leqslant 2^{-n^{\Omega(1)}}\sqrt{|X||Y|}$ for some choice of an AC^0 function $f : \{0,1\}^n \times \{0,1\}^n \to \{-1,+1\}$ and some multisets X, Y of n-bit Boolean strings? An affirmative answer to this question would subsume our results and additionally imply that AC^0 is not learnable in Kearns' important *statistical query model* [19]. A suitable *lower* bound on the spectral norm of every such matrix, on the other

hand, would result in the breakthrough separation of PH^{cc} and PSPACE^{cc}. See [4, 30, 41, 45] for relevant background.

The second problem concerns the sign-rank of arbitrary pattern matrices. For a Boolean function $f : \{0,1\}^n \to \{-1,+1\}$, its *threshold degree* $\deg(f)$ is the least degree of a multivariate polynomial $p(x_1, \ldots, x_n)$ such that $f(x) \equiv \operatorname{sign}(p(x))$. Let M_f denote the (n^c, n, f)-pattern matrix, where $c \geqslant 1$ is a sufficiently large constant. It is straightforward to verify that the sign-rank of M_f does not exceed $n^{O(\deg(f))}$. Is that upper bound close to optimal? Specifically, does M_f have sign-rank $\exp(\deg(f)^{\Omega(1)})$ for every f? Evidence in this paper and prior work suggests an answer in the affirmative. For example, our main result confirms this hypothesis for the Minsky-Papert function, $f = \mathrm{MP}$. For $f = \mathrm{PARITY}$ the hypothesis immediately follows from the seminal work of Forster [12]. More examples were discovered in [48].

In the field of communication complexity, we were able to resolve the main question left open by Babai, Frankl, and Simon [4], but only in one direction: $\mathsf{PH}^{cc} \not\subseteq \mathsf{UPP}^{cc}$. The other direction remains wide open despite much research, i.e., no lower bounds are known for PH^{cc} or even Σ_2^{cc}. The latter question is in turn closely related to such important concepts as *matrix rigidity* [41] and *graph complexity* (e.g., see [17, 39, 40] and the literature cited therein).

Acknowledgments

The authors would like to thank Adam Klivans and Yaoyun Shi for helpful feedback on an earlier version of this manuscript.

References

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

[2] M. Alekhnovich, M. Braverman, V. Feldman, A. Klivans, and T. Pitassi. Learnability and automatizability. In *Proceedings of the 45th Symposium on Foundations of Computer Science (FOCS)*, 2004.

[3] N. Alon, P. Frankl, and V. Rödl. Geometrical realization of set systems and probabilistic communication complexity. In *Proc. of the 26th Symposium on Foundations of Computer Science (FOCS)*, pages 277–280, 1985.

[4] L. Babai, P. Frankl, and J. Simon. Complexity classes in communication complexity theory. In *Proc. of the 27th Symposium on Foundations of Computer Science (FOCS)*, pages 337–347, 1986.

[5] S. Ben-David, N. Eiron, and H. U. Simon. Limitations of learning via embeddings in Euclidean half spaces. *J. Mach. Learn. Res.*, 3:441–461, 2003.

[6] A. Blum, A. M. Frieze, R. Kannan, and S. Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22(1/2):35–52, 1998.

[7] N. H. Bshouty. A subexponential exact learning algorithm for DNF using equivalence queries. *Inf. Process. Lett.*, 59(1):37–39, 1996.

[8] H. Buhrman, N. K. Vereshchagin, and R. de Wolf. On computation and communication with small bias. In *Proc. of the 22nd Conf. on Computational Complexity (CCC)*, pages 24–32, 2007.

[9] E. W. Cheney. *Introduction to Approximation Theory*. Chelsea Publishing, New York, 2nd edition, 1982.

[10] R. de Wolf. *Quantum Computing and Communication Complexity*. PhD thesis, University of Amsterdam, 2001.

[11] V. Feldman, P. Gopalan, S. Khot, and A. K. Ponnuswami. New results for learning noisy parities and halfspaces. In *Proceedings of the 47th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 563–574, 2006.

[12] J. Forster. A linear lower bound on the unbounded error probabilistic communication complexity. *J. Comput. Syst. Sci.*, 65(4):612–625, 2002.

[13] J. Forster, M. Krause, S. V. Lokam, R. Mubarakzjanov, N. Schmitt, and H.-U. Simon. Relations between communication complexity, linear arrangements, and computational complexity. In *Proc. of the 21st Conf. on Foundations of Software Technology and Theoretical Computer Science (FST TCS)*, pages 171–182, 2001.

[14] J. Forster and H. U. Simon. On the smallest possible dimension and the largest possible margin of linear arrangements representing given concept classes. *Theor. Comput. Sci.*, 350(1):40–48, 2006.

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

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

[17] S. Jukna. On graph complexity. *Combinatorics, Probability and Computing*, 15:1–22, 2006.

[18] B. Kalyanasundaram and G. Schnitger. The probabilistic communication complexity of set intersection. *SIAM J. Discrete Math.*, 5(4):545–557, 1992.

[19] M. Kearns. Efficient noise-tolerant learning from statistical queries. In *Proc. of the 25th Symposium on Theory of Computing (STOC)*, pages 392–401, 1993.

[20] M. Kearns and L. Valiant. Cryptographic limitations on learning boolean formulae and finite automata. *J. ACM*, 41(1):67–95, 1994.

[21] M. J. Kearns and U. V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, Cambridge, 1994.

[22] M. Kharitonov. Cryptographic hardness of distribution-specific learning. In *Proc. of the 25th Symposium on Theory of Computing*, pages 372–381, 1993.

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

[24] A. R. Klivans and A. A. Sherstov. Cryptographic hardness for learning intersections of halfspaces. In *Proc. of the 47th Symposium on Foundations of Computer Science (FOCS)*, pages 553–562, 2006.

[25] A. R. Klivans and A. A. Sherstov. A lower bound for agnostically learning disjunctions. In *Proc. of the 20th Conf. on Learning Theory (COLT)*, pages 409–423, 2007.

[26] A. R. Klivans and A. A. Sherstov. Unconditional lower bounds for learning intersections of halfspaces. *Machine Learning*, 69(2–3):97–114, 2007.

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

[28] E. Kushilevitz and N. Nisan. *Communication complexity*. Cambridge University Press, New York, 1997.

[29] N. Linial, S. Mendelson, G. Schechtman, and A. Shraibman. Complexity measures of sign matrices. *Combinatorica*, 2006. To appear. Manuscript at `http://www.cs.huji.ac.il/~nati/PAPERS/complexity_matrices.ps.gz`.

[30] S. V. Lokam. Spectral methods for matrix rigidity with applications to size-depth trade-offs and communication complexity. *J. Comput. Syst. Sci.*, 63(3):449–473, 2001.

[31] M. L. Minsky and S. A. Papert. *Perceptrons: expanded edition*. MIT Press, Cambridge, Mass., 1988.

[32] I. Newman. Private vs. common random bits in communication complexity. *Inf. Process. Lett.*, 39(2):67–71, 1991.

[33] N. Nisan. The communication complexity of threshold gates. In *Combinatorics, Paul Erdős is Eighty*, pages 301–315, 1993.

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

[35] A. B. J. Novikoff. On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, volume XII, pages 615–622, 1962.

[36] R. O'Donnell and R. A. Servedio. New degree bounds for polynomial threshold functions. In *Proc. of the 35th Symposium on Theory of Computing (STOC)*, pages 325–334, 2003.

[37] R. Paturi. On the degree of polynomials that approximate symmetric Boolean functions. In *Proc. of the 24th Symposium on Theory of Computing*, pages 468–474, 1992.

[38] R. Paturi and J. Simon. Probabilistic communication complexity. *J. Comput. Syst. Sci.*, 33(1):106–123, 1986.

[39] P. Pudlák, V. Rödl, and P. Savický. Graph complexity. *Acta Inf.*, 25(5):515–535, 1988.

[40] A. A. Razborov. Bounded-depth formulae over the basis $\{\&, \oplus\}$ and some combinatorial problems. *Complexity Theory and Applied Mathematical Logic*, vol. "Problems of Cybernetics":146–166, 1988. In Russian, available at `http://www.mi.ras.ru/~razborov/graph.pdf`.

[41] A. A. Razborov. On rigid matrices. Manuscript in Russian, available at `http://www.mi.ras.ru/~razborov/rigid.pdf`, June 1989.

[42] A. A. Razborov. On the distributional complexity of disjointness. *Theor. Comput. Sci.*, 106(2):385–390, 1992.

[43] T. J. Rivlin. *An Introduction to the Approximation of Functions*. Dover Publications, New York, 1981.

[44] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–408, 1958.

[45] A. A. Sherstov. Halfspace matrices. In *Proc. of the 22nd Conf. on Computational Complexity (CCC)*, pages 83–95, 2007.

[46] A. A. Sherstov. Powering requires threshold depth 3. *Inf. Process. Lett.*, 102(2–3):104–107, 2007.

[47] A. A. Sherstov. Separating AC^0 from depth-2 majority circuits. In *Proc. of the 39th Symposium on Theory of Computing (STOC)*, pages 294–301, 2007.

[48] A. A. Sherstov. Unbounded-error communication complexity of symmetric functions. Technical Report TR-07-53, The Univ. of Texas at Austin, Dept. of Computer Sciences, September 2007.

[49] A. A. Sherstov. The pattern matrix method for lower bounds on quantum communication. In *Proc. of the 40th Symposium on Theory of Computing (STOC)*, 2008. To appear.

[50] N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. In *Proc. of the 18th Conf. on Learning Theory (COLT)*, pages 545–560, 2005.

[51] J. Tarui and T. Tsukiji. Learning DNF by approximating inclusion-exclusion formulae. In *Proc. of the 14th Conf. on Computational Complexity*, pages 215–221, 1999.

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

A More on the Unbounded-Error Model

Readers with background in communication complexity will note that the unbounded-error model is exactly the same as the *private-coin randomized model* [28, Chap. 3], with one exception: in the latter case the correct answer is expected with probability at least $2/3$, whereas in the former case the success probability need only *exceed* $1/2$ (say, by an exponentially small amount). This difference has far-reaching implications. For example, the fact that the parties in the unbounded-error model do not have a *shared* source of random bits is crucial: allowing shared randomness would make the complexity of every function a constant, as one can easily verify. By contrast, introducing shared randomness into the randomized model has minimal impact on the complexity of any given function [32].

As one might expect, the weaker success criterion in the unbounded-error model also has a drastic impact on the complexity of certain functions. For example, the well-known function on n-bit strings has complexity $O(\log n)$ in the unbounded-error model and $\Omega(n)$ in the randomized model [18, 42]. Furthermore, explicit functions are known [8,45] with unbounded-error complexity $O(\log n)$ that require $\Omega(\sqrt{n})$ communication in the randomized model to even achieve advantage $2^{-\sqrt{n}/5}$ over random guessing.

More generally, the unbounded-error complexity of a function $f : X \times Y \to \{0,1\}$ is never much more than its complexity in the other standard models. For example, it is not hard to see that

$$\begin{aligned} U(f) &\leqslant \min\{N^0(f), N^1(f)\} + O(1) \\ &\leqslant D(f) + O(1), \end{aligned}$$

where D, N^0, and N^1 refer to communication complexity in the *deterministic*, 0-*nondeterministic*, and 1-*nondeterministic* models, respectively. Continuing,

$$\begin{aligned} U(f) &\leqslant R_{1/3}(f) + O(1) \\ &\leqslant O\left(R^{\text{pub}}_{1/3}(f) + \log\log\left[|X| + |Y|\right]\right), \end{aligned}$$

where $R_{1/3}$ and $R^{\text{pub}}_{1/3}$ refer to the *private-* and *public-coin randomized* models, respectively. As a matter of fact, one can show that

$$U(f) \leqslant O\left(Q^*_{1/3}(f) + \log\log\left[|X| + |Y|\right]\right),$$

where $Q^*_{1/3}$ refers to the *quantum model with prior entanglement*. An identical inequality is clearly valid for the quantum model *without* prior entanglement. See [10, 28] for rigorous definitions of these various models; our sole intention was to point out that the unbounded-error model is at least as powerful.

B Details on Forster's Method

The purpose of this section is to explain in detail how Theorem 2.4 is implicit in Forster's work.

Recall that vectors $v_1, \dots, v_n$ in $\mathbb{R}^r$ are said to be *in general position* if no r of them are linearly dependent. Forster proved that any set of vectors in general position can be balanced in a useful way:

Theorem B.1 (Forster [12, Thm. 4.1]). *Let $U \subset \mathbb{R}^r$ be a finite set of vectors in general position, $|U| \geqslant r$. Then there is a nonsingular transformation $A \in \mathbb{R}^{r\times r}$ such that*

$$\sum_{u\in U} \frac{1}{\|Au\|^2} (Au)(Au)^{\mathsf{T}} = \frac{|U|}{r} I_r.$$

(The vector norm $\|\cdot\|$ above and throughout this section is the Euclidean norm $\|\cdot\|_2$.) Theorem B.1 is the main technical tool needed to establish the statement of interest to us (cf. [12, Thm. 2.2]):

Theorem 2.4 (Restated from p. 9). *Let X, Y be finite sets and $M = [M_{xy}]_{x\in X, y\in Y}$ a real matrix ($M \neq 0$). Put $r = \text{sign-rank}(M)$. Then there is a matrix $R = [R_{xy}]_{x\in X, y\in Y}$ such that:*

$$\text{rank}(R) = r, \tag{B.1}$$

$$M \circ R \geqslant 0, \tag{B.2}$$

$$\|R\|_\infty \leqslant 1, \tag{B.3}$$

$$\|R\|_{\text{F}} = \sqrt{|X|\,|Y|/r}. \tag{B.4}$$

Proof. Since $M \neq 0$, it follows that $r \geqslant 1$. Fix a matrix $Q = [Q_{xy}]$ of rank r such that

$$Q_{xy} M_{xy} > 0 \qquad \text{whenever} \qquad M_{xy} \neq 0. \tag{B.5}$$

Write

$$Q = \Big[\langle u_x, v_y\rangle\Big]_{x\in X,\, y\in Y}$$

for suitable collections of vectors $\{u_x\} \subset \mathbb{R}^r$ and $\{v_y\} \subset \mathbb{R}^r$. If the vectors $\{u_x : x \in X\}$ are not already in general position, we can replace them with their slightly perturbed versions $\{\tilde{u}_x\}$ that *are* in general position. Provided that the perturbations are small enough, property (B.5) will still hold, i.e., we will have $\langle \tilde{u}_x, v_y\rangle M_{xy} > 0$ whenever $M_{xy} \neq 0$. As a result, we can assume w.l.o.g. that $\{u_x\}$ are in general position. Furthermore, a moment's reflection reveals that the vectors $\{v_y\}$ can be assumed to be all nonzero.

Since sign-rank$(M) \leqslant \operatorname{rank}(M)$, we infer that $|X| \geqslant r$. Theorem B.1 is therefore applicable to the set $\{u_x\}$ and yields a nonsingular matrix A with

$$\sum_{x \in X} \frac{1}{\|Au_x\|^2} (Au_x)(Au_x)^{\mathsf{T}} = \frac{|X|}{r} I_r. \tag{B.6}$$

Define

$$R = \left[\frac{\langle u_x, v_y \rangle}{\|Au_x\| \, \|(A^{-1})^{\mathsf{T}} v_y\|} \right]_{x \in X,\, y \in Y}.$$

It remains to verify properties (B.1)–(B.4). Property (B.1) follows from the representation $R = D_1 Q D_2$, where D_1 and D_2 are diagonal matrices with strictly positive diagonal entries. By (B.5), we know that $R_{xy} M_{xy} > 0$ whenever $M_{xy} \neq 0$, which immediately gives us (B.2). Property (B.3) holds because

$$\frac{|\langle u_x, v_y \rangle|}{\|Au_x\| \, \|(A^{-1})^{\mathsf{T}} v_y\|} = \frac{|\langle Au_x, (A^{-1})^{\mathsf{T}} v_y \rangle|}{\|Au_x\| \, \|(A^{-1})^{\mathsf{T}} v_y\|} \leqslant 1.$$

Finally, property (B.4) will follow once we show that $\sum_x R_{xy}^2 = |X|/r$ for every $y \in Y$. So, fix $y \in Y$ and consider the unit vector $v = (A^{-1})^{\mathsf{T}} v_y / \|(A^{-1})^{\mathsf{T}} v_y\|$. We have:

$$\begin{aligned}
\sum_{x \in X} R_{xy}^2 &= \sum_{x \in X} \frac{\langle u_x, v_y \rangle^2}{\|Au_x\|^2 \, \|(A^{-1})^{\mathsf{T}} v_y\|^2} \\
&= \sum_{x \in X} \frac{(v_y^{\mathsf{T}} A^{-1})(Au_x)(Au_x)^{\mathsf{T}} (A^{-1})^{\mathsf{T}} v_y}{\|Au_x\|^2 \, \|(A^{-1})^{\mathsf{T}} v_y\|^2} \\
&= v^{\mathsf{T}} \left(\sum_{x \in X} \frac{1}{\|Au_x\|^2} (Au_x)(Au_x)^{\mathsf{T}} \right) v \\
&= \frac{|X|}{r} \qquad \text{by (B.6).}
\end{aligned}$$

□

ECCC ISSN 1433-8092

http://eccc.hpi-web.de/