THE SIGN-RANK OF AC^0 *

ALEXANDER A. RAZBOROV[†] AND ALEXANDER A. SHERSTOV[‡]

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 $f(x, y) = \bigwedge_{i=1,\dots,m} \bigvee_{j=1,\dots,m^2}(x_{ij} \wedge y_{ij})$. We show that the matrix $[f(x, y)]_{x,y}$ has sign-rank $\exp(\Omega(m))$. This in particular implies that $\Sigma_2^{cc} \not\subseteq \mathsf{UPP}^{cc}$, which solves a longstanding open problem in communication complexity posed by Babai, Frankl, and Simon (1986).

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

Key words. Sign-rank; communication complexity; complexity classes Σ_2^{cc}, Π_2^{cc}, and UPP^{cc}; constant-depth AND/OR/NOT circuits.

AMS subject classifications. 03D15, 68Q15, 68Q17

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, 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 [40, 2, 4, 13, 14, 26, 32, 48, 51]. We will give a detailed overview of these applications shortly as they pertain to our work.

Despite its importance, progress in understanding sign-rank has been slow and difficult. Indeed, we are aware of only a few nontrivial results on this subject. Alon et al. [2] obtained strong lower bounds on the sign-rank of random matrices. No nontrivial results were available for any *explicit* matrices until the breakthrough work of Forster [13], who proved strong lower bounds on the sign-rank of Hadamard matrices and, more generally, all sign matrices with small spectral norm. Several extensions and refinements of Forster's method were proposed in subsequent work [14, 15, 32]. Near-tight estimates of the sign-rank were obtained in [51] for all symmetric problems, i.e., matrices of the form $[D(\sum x_i y_i)]_{x,y}$ where $D : \{0, 1, \dots, n\} \to \{-1, +1\}$ is a given predicate 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. It will henceforth be convenient to view Boolean functions as mappings of the form $\{0,1\}^n \to \{-1,+1\}$, where the elements -1 and $+1$ of the range represent "true" and "false," respectively. The central objective of our study is to estimate the maximum sign-rank of a matrix $[f(x,y)]_{x,y}$, where $f : \{0,1\}^n \times \{0,1\}^n \to$

* An extended abstract of this article appeared in *Proceedings of the 49th IEEE Symposium on Foundations of Computer Science (FOCS),* pages 57–66, 2008.

[†] Department of Computer Science, University of Chicago and Steklov Mathematical Institute (razborov@cs.uchicago.edu). Supported by NSF grant ITR-0324906 and by the Russian Foundation for Basic Research.

[‡] Department of Computer Science, The University of Texas at Austin (sherstov@cs.utexas.edu).

$\{-1,+1\}$ is a function in AC^0. An obvious upper bound is 2^n, while the best lower bound prior to this paper was quasipolynomial. (The quasipolynomial lower bound is immediate from Forster's work [13] and the fact that AC^0 can compute INNER PRODUCT MODULO 2 on $\log^c n$ variables, for every constant $c > 1$.) 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 that we prove is almost tight. (See Remark 6.1 for details.) Moreover, Theorem 1.1 is optimal with respect to circuit depth: AC^0 circuits of depth 1 and 2 lead to at most polynomial sign-rank. Indeed, the function $\bigvee_{i=1}^m (x_i \wedge y_i)$ which is universal in this context possesses the matrix representation $[\mathrm{sgn}(1/2 - \sum_i x_i y_i)]_{x,y}$ and thus has sign-rank at most $m+1$.

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.

1.1. Communication complexity. The study of sign-rank is synonymous with the study of *unbounded-error communication complexity,* a rich model introduced by Paturi and Simon [40]. Fix a function $f : X \times Y \to \{-1,+1\}$, 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 the more familiar, bounded-error models, where the correct output is expected with probability at least $2/3$.

Another compelling aspect of the unbounded-error model is that it has an exact matrix-analytic formulation. Let $f : X \times Y \to \{-1,+1\}$ be a given function and $M = [f(x,y)]_{x\in X,\, y \in Y}$ its communication matrix. Paturi and Simon [40] showed that

$$U(f) = \log_2(\text{sign-rank}(M)) \pm O(1).$$

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 solves a longstanding problem in communication complexity. Specifically, the only models for which efficient simulations in the unbounded-error model were unknown had been developed in the seminal paper by Babai, Frankl, and Simon [3]. These models are the communication analogues of the classes PH and PSPACE. Babai et al. asked [3, p. 345] whether $\Sigma_2^{cc} \subseteq \mathsf{UPP}^{cc}$. Forster [13] 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 implies that $\Pi_2^{cc} \not\subseteq \mathsf{UPP}^{cc}$ and hence (since UPP^{cc} is closed under complementation) $\Sigma_2^{cc} \not\subseteq \mathsf{UPP}^{cc}$. See Section 7 for detailed background on these various complexity classes.

1.2. Learning theory. In a seminal paper [55], Valiant formulated the *probably approximately correct* (PAC) model of learning, now a central 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 [22, 24, 28, 12, 29, 27].

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

$$f(x) \equiv \mathrm{sgn}(a_1 x_1 + \cdots + a_n x_n - \theta)$$

for some reals $a_1, \ldots, a_n, \theta$. Halfspaces constitute one of the most studied classes in computational learning theory [46, 37, 34, 5] 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 $\mathcal{C}$ 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 $f \in \mathcal{C}$ has a representation

$$f(x) \equiv \mathrm{sgn}(a_1 \phi_1(x) + \cdots + a_r \phi_r(x))$$

for some reals $a_1, \ldots, a_r$. This process is technically described as *embedding $\mathcal{C}$ in halfspaces of dimension r.* Once this is accomplished, $\mathcal{C}$ can 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 [4, 27]. For brevity, we refer to the smallest dimension of such a representation as the *dimension complexity* of a given class. Formally, the dimension complexity $\mathrm{dc}(\mathcal{C})$ of a given class $\mathcal{C}$ of functions $\{0,1\}^n \to \{-1, +1\}$ is the least r for which there exist real functions $\phi_1, \ldots, \phi_r \colon \{0,1\}^n \to \mathbb{R}$ such that every $f \in \mathcal{C}$ is expressible as

$$f(x) \equiv \mathrm{sgn}(a_1(f)\phi_1(x) + \cdots + a_r(f)\phi_r(x)) \tag{1.1}$$

for some reals $a_1(f), \ldots, a_r(f)$ depending on f only. To relate this discussion to sign-rank, let $M_{\mathcal{C}} = [f(x)]_{f \in \mathcal{C},\, x \in \{0,1\}^n}$ be the characteristic matrix of $\mathcal{C}$. A moment's reflection reveals that the identity (1.1) is yet another way of saying that $M_{\mathcal{C}}$ has the same sign pattern as the matrix $[\sum_{i=1}^r a_i(f)\phi_i(x)]_{f,x}$ of rank at most r, whence 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 [53, 48], 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 formulas, due to Klivans and Servedio [26], 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 $\mathcal{C}$ be the set of all read-once* (*hence, linear-size*) *DNF formulas $f : \{0,1\}^n \to \{-1,+1\}$. Then $\mathcal{C}$ 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 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 [55]. More than twenty years later, this challenge remains a central open problem in computational learning theory despite active research, e.g., [7, 54, 26]. To account for this lack of progress, several hardness results have been obtained based on complexity-theoretic assumptions [24, 1]. Corollary 1.3 complements that line of work by exhibiting an unconditional, *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 PAC learning algorithms designed to date, with the notable exception [18, 6] of learning low-degree polynomials over GF(p). Furthermore, it is known [23, 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.

The lower bound in this work also points to the importance of generalizing the dimension complexity framework while maintaining algorithmic efficiency. Unfortunately, natural attempts at such a generalization do not work, including the idea of sign-representations that err on a small fraction of the inputs [12].

1.3. Threshold circuits. Recall that a *threshold gate* g with Boolean inputs $x_1, \ldots, x_n$ is a function of the form $g(x) = \operatorname{sgn}(a_1x_1 + \cdots + a_nx_n - \theta)$, 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, where by the size of a circuit we mean the number of gates.

Communication complexity has been crucial to the progress on this problem. Through randomized communication complexity, many explicit functions have been found [17, 16, 36, 47, 49] 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, that of *threshold-of-majority* circuits. More precisely, Forster et al. [14] proved the following:

Lemma 1.4 ([14, Lemma 5]). *Let $f : \{0,1\}^n \times \{0,1\}^n \to \{-1,+1\}$ be a Boolean function computed by a depth-2 threshold circuit with arbitrary weights at the top gate* *and integer weights of absolute value at most w at the bottom. Then the sign-rank of $F = [f(x,y)]_{x,y}$ is $O(snw)$, where s is the number of gates.*

Combined with our main result, this immediately gives the following.

Corollary 1.5 (Threshold circuits). *In every depth-2 threshold circuit that computes the function $f_m(x,y) = \bigwedge_{i=1}^{m} \bigvee_{j=1}^{m^2} (x_{ij} \wedge y_{ij})$, the sum of the absolute values of the (integer) weights at the bottom level must be of magnitude $2^{\Omega(m)}$.*

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 [30, 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 by Buhrman et al. [8] and independently in [49, 50].

1.4. Our proof and techniques.

At a high level, we adopt the approach introduced in [51] in the context of determining the sign-rank of symmetric functions. That approach is based on Forster's method [13] for proving lower bounds on the sign-rank in combination with the *pattern matrix method* [50].

In more detail, Forster [13] showed that a matrix with entries ±1 has high sign-rank if it has low spectral norm. Follow-up papers [14, 15] relaxed the assumptions on the entries of the matrix. We begin with a simple generalization of these results (Theorem 5.1), proving that in order to ensure high sign-rank, it suffices to require, along with low spectral norm, that *most* of the entries are not *too* small in absolute value.

In [50, 51] it was shown how to construct such matrices from a function $g : \{0,1\}^n \to \mathbb{R}$ with the following properties:

(1) low-order Fourier coefficients of g are zero, i.e., g is orthogonal to all low-degree polynomials;

(2) g is not too small in absolute value on most inputs $x \in \{0,1\}^n$.

The entries of the matrix are the values of g repeated in certain patterns; the resulting matrix is referred to as the *pattern matrix* for g.

If the original function g is Boolean, then (2) is immediate, but (1) rarely holds. A way to fix the problem is to find a probability distribution μ on $\{0,1\}^n$ which is *orthogonalizing* in the sense that the combined function $g'(x) = g(x)\mu(x)$ is orthogonal to all low-degree polynomials. But then care must be taken to ensure property (2) for the new function g'. In other words, μ must be nonnegligible on most inputs (*smooth*).

In [51], this program was carried out to obtain lower bounds on the sign-rank of symmetric functions. The existence of the required smooth orthogonalizing distribution was established using a linear-programming dual interpretation of Paturi's lower bounds [39] for the uniform approximation of symmetric functions.

In this paper we study the sign-rank of AC^0 functions, and specifically the sign-rank of the matrix derived from the communication version of the *Minsky-Papert function*:

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

Our proof relies on a linear-programming dual interpretation of the Minsky-Papert lower bound for the sign-representation of MP_m. The construction of the smooth orthogonalizing distribution in this paper, which is the crux of the program, is unrelated to the corresponding step in [51] and requires considerable new ideas.

Having described our proof at a high level, we will now examine it in more detail, from the bottom up. Figure 1.1 illustrates the main components of our proof. A starting point in our study is an elegant result due to Minsky and Papert [34], who constructed a linear-size DNF formula that cannot be sign-represented by polynomials of low degree.

Second, we revisit a fundamental technique from approximation theory, the *interpolation bound,* which bounds a degree-d univariate polynomial p 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 3.4). 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 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 3.6). 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. This completes step 1 of the above program, as desired.

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 analytic 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 analysis (see Section 3). To our knowledge, this

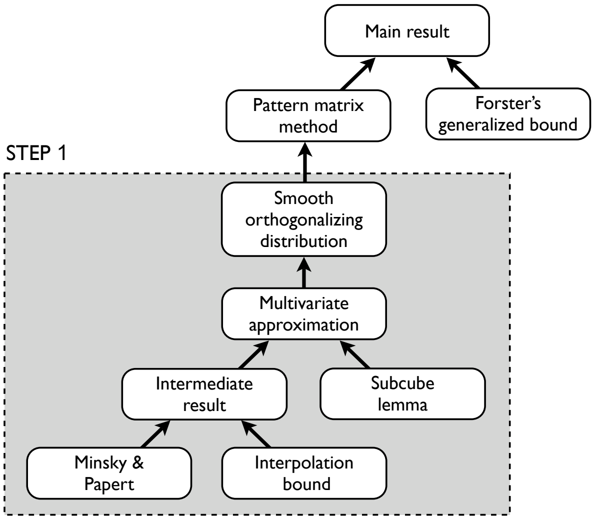

Fig. 1.1: Proof outline.

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 largest total degree of any monomial. The notation $[n]$ refers to the set $\{1,2,\dots,n\}$. 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 $\|A\|_\infty = \max_{i,j} |A_{ij}|$ 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 on 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-Schwarz inequality:

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

For $A, B \in \mathbb{R}^{m\times n}$, we write $\langle A, B\rangle = \sum_{i,j} A_{ij}B_{ij}$. 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 $A \circ B = [A_{ij}B_{ij}]$. Recall that

$$\operatorname{rank}(A\circ B) \leqslant \operatorname{rank}(A)\operatorname{rank}(B). \tag{2.3}$$

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. Fourier transform over $\mathbb{Z}_2^n$. Consider the vector space of functions $\{0,1\}^n \to \mathbb{R}$, equipped with the inner product

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

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 $\hat{f}(S) = \langle f, \chi_S \rangle$. 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 $(x_{\sigma(1)}, \dots, x_{\sigma(n)}) \in \{0,1\}^n$. 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

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

is symmetric. 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 [34]). *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*

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

We will need the following straightforward generalization.

Proposition 2.3. *Let* $n_1, \dots, n_k$ *be positive integers,* $n = n_1 + \dots + n_k$*. 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* $x = (x^{(1)}, \dots, x^{(k)})$*, where* $x^{(i)} = (x_{n_1 + \dots + n_{i-1} + 1}, \dots, x_{n_1 + \dots + n_i})$*. Then there is a polynomial* p *on* $\mathbb{R}^k$ *of degree at most* r *such that*

$$\mathop{\mathbf{E}}_{\sigma_1 \in S_{n_1}, \dots, \sigma_k \in S_{n_k}} \Big[\phi(\sigma_1 x^{(1)}, \dots, \sigma_k x^{(k)})\Big] = p\Big(|x^{(1)}|, \dots, |x^{(k)}|\Big) \qquad \forall x \in \{0,1\}^n.$$

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 only applied to matrices A with nonzero entries.)

In general, the sign-rank of a matrix can be vastly smaller than its rank. For example, consider the following nonsingular matrices of order $n \geqslant 3$, representing well-known problems greater-than and equality in communication complexity:

$$\begin{bmatrix} 1 & & & & & \\ & 1 & & & \text{\Large 1} & \\ & & 1 & & & \\ & & & \ddots & & \\ & \text{\Large -1} & & & 1 & \\ & & & & & 1 \end{bmatrix}, \qquad \begin{bmatrix} -1 & & & & & \\ & -1 & & & \text{\Large 1} & \\ & & -1 & & & \\ & & & \ddots & & \\ & \text{\Large 1} & & & -1 & \\ & & & & & -1 \end{bmatrix}.$$

These matrices have sign-rank at most 2 and 3, respectively. Indeed, the first matrix has the same sign pattern as the matrix $[2(j-i)+1]_{i,j}$. The second has the same sign pattern as the matrix $[(1-\epsilon)-\langle v_i, v_j\rangle]_{i,j}$, where $v_1, v_2, \ldots, v_n \in \mathbb{R}^2$ are arbitrary pairwise distinct unit vectors and ϵ is a suitably small positive real, cf. [40, §5].

Bounding the sign-rank from below is a considerable challenge. In a breakthrough result, Forster [13] 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 [13], 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:*

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

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

2.5. Pattern matrices. *Pattern matrices* were introduced in [49, 50] and proved useful in obtaining strong lower bounds on communication complexity. Relevant definitions and results from [50] 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

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

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

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

The *pattern matrix method* [50, 51] combines Theorem 2.6 with a suitable analytic property of the given Boolean function $f\colon \{0,1\}^n \to \{-1,+1\}$ to obtain a communication lower bound for the (N, n, f)-pattern matrix, in the corresponding model of communication. Examples of analytic properties that have been previously used in this context include *high approximate degree* [50], *high threshold degree* [49, 50], and *smooth orthogonalizability* [51]. The intermediate object, denoted by ϕ in Theorem 2.6, is taken to be the real function that arises from linear programming duality and witnesses the corresponding analytic property. See [52] for a detailed guide to the pattern matrix method and follow-up work. Following [51], in this paper we will be working with *smooth orthogonalizability* as our analytic property of choice.

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 3.6). 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 points. For other uses of this technique, see Cheney [9, §7, Lem. 1] and Rivlin [45, 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*

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

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:

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

□

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

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

The following properties of g are immediate:

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

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

The degree of every monomial in g is between $k+1$ and n, so that

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

Furthermore,

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

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). □

Remark 3.3. One can use Lemma 3.2 to bound f on inputs other than 1^n. For example, it follows immediately that $|f(y)| \leqslant 2^k\binom{|y|}{k}\max_{|x|\leqslant k}|f(x)|$, where $y \in \{0,1\}^n$ is arbitrary with $|y| > k$. We will not need this observation, however.

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

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

Looking ahead, Z^+ and $Z \setminus Z^+$ correspond to the sets on which the Minsky-Papert function $\bigwedge_{i=1}^m \bigvee_{j=1}^{4m^2} x_{ij}$ is true and false, respectively. Accordingly, we 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 [34] of the threshold degree of CNF formulas.

Lemma 3.4. *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 $|Q(z)| \leqslant 4^{m+d}$ at every point $z \in Z^+$ with $\Delta(u,z) < m/3$, where $u = (1^2, 3^2, 5^2, \dots, (2m-1)^2) \in Z^+$.

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

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

where

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

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

$$p(2i-1) \geqslant -1 \qquad (i \in S), \tag{3.8}$$
$$p(2i) \leqslant 1 \qquad (i = 0,1,\dots,m). \tag{3.9}$$

Claim 3.5. *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 3.5 provides $|S| > 2m/3 \geqslant 2d \geqslant \deg(p)$ points in $[0, 2m]$, with pairwise distances at least 1, at which p is bounded in absolute value by 1. By Lemma 3.1,

$$\max_{0\leqslant t\leqslant 2m} |p(t)| \leqslant 2^{\deg(p)}\binom{2m}{\deg(p)} \leqslant 4^{m+d}.$$

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

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

Then

$$|Q(z)| \leqslant 16^m \qquad\qquad (z \in Z^+).$$

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 3.4 that $|f(x)| \leqslant 4^{m+d}$ for every $x \in \{0,1\}^m$ with $|x| < m/3$, and our goal is to show that $|f(1^m)| \leqslant 16^m$. Since Q has degree d, the Fourier transform of f is supported on characters of order up to d. As a result,

$$\begin{aligned} |f(1^m)| &\leqslant 2^d \binom{m}{d} \max_{|x| \leqslant d} |f(x)| && \text{by Lemma 3.2} \\ &\leqslant 2^{2m+3d} \binom{m}{d} && \text{by Lemma 3.4} \\ &\leqslant 16^m. \end{aligned}$$

□

4. A smooth orthogonalizing distribution. An important concept in our work is that of an *orthogonalizing distribution* [51]. Let $f : \{0,1\}^n \to \{-1,+1\}$ be given. A distribution μ on $\{0,1\}^n$ is *d-orthogonalizing* for f if

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

In words, a distribution μ is d-orthogonalizing for f if with respect to μ, the function f is orthogonal to every character of 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 [34] and has played an important role in later works [30, 38, 49, 50]. An explicit m-orthogonalizing distribution for MP_m is known [49]. 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 [49] 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 $\mu(x) \geqslant \frac{1}{2}16^{-m}\,2^{-n}$ 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.,

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

It suffices to show that the following linear program has optimum at least $\frac{1}{2}16^{-m}$:

$$\boxed{\begin{array}{rll} \text{variables:} & \epsilon \geqslant 0; \quad \mu(x) \geqslant 0 \text{ for } x \in \{0,1\}^n & \\ \text{maximize:} & \epsilon & \\ \text{subject to:} & \displaystyle\sum_{x \in \{0,1\}^n} \mu(x)\mathrm{MP}_m(x)\chi_S(x) = 0 & \text{for } |S| < m/3, \\ & \displaystyle\sum_{x \in \{0,1\}^n} \mu(x) \leqslant 1, & \\ & \mu(x) \geqslant \epsilon 2^{-n} & \text{for } x \in X. \end{array}} \tag{LP1}$$

The optimum being nonzero, it will follow by a scaling argument that any optimal solution has $\sum \mu(x) = 1$. As a result, μ will be the sought probability distribution.

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:

$$\boxed{\begin{array}{rll} \text{variables:} & \text{a polynomial } Q \text{ on } \mathbb{R}^m \text{ of degree} < m/3; & \\ & \eta \geqslant 0; \quad \delta_z \geqslant 0 \text{ for } z \in Z^+ & \\ \text{minimize:} & \eta & \\ \text{subject to:} & \displaystyle\sum_{x \in X} \delta_{z(x)} \geqslant 2^n, & \\ & F(z)Q(z) \geqslant -\eta & \text{for } z \in Z, \\ & F(z)Q(z) \geqslant -\eta + \delta_z & \text{for } z \in Z^+. \end{array}} \tag{LP2}$$

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 $\eta \leqslant \frac{1}{2}16^{-m}$. Then $|Q(z)| \leqslant \frac{1}{2}$ for each $z \in Z^+$, by Theorem 3.6. 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 $\frac{1}{2}16^{-m}$. ☐

5. A generalization of Forster's bound. Using Theorem 2.4, Forster gave a simple proof of the following fundamental result [13, 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. [14, 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 [15, §5] considered a different generalization, inspired by the notion of matrix rigidity (see, e.g., [43]). 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 (5.1) and (5.2).

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

$$\operatorname{rank}(R) = r, \tag{5.3}$$
$$A \circ R \geqslant 0, \tag{5.4}$$
$$\|R\|_\infty \leqslant 1, \tag{5.5}$$
$$\|R\|_{\mathrm{F}} = \sqrt{s/r}. \tag{5.6}$$

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

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

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 $r = \text{sign-rank}(A)$. □

6. Main result. At last, we are in a position to prove our main result.

THEOREM 1.1 (Restated from p. 2). *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 $n = 4m^3$ and $N = 17^6 n$. 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 $\mu(x) \geqslant \frac{1}{2} 16^{-m}\, 2^{-n}$. As a result, all but a $2^{-\Omega(m^2)}$ fraction of the entries of $M \circ P$ are at least $\frac{1}{2} 16^{-m}\, 2^{-n}$ in absolute value. Theorem 5.1 at once implies that

$$\text{sign-rank}(M) \geqslant \text{sign-rank}(M \circ P) \geqslant \min\left\{ \frac{16^{-m}\, 2^{-n}\sqrt{s}}{4\,\|M \circ P\|},\ 2^{\Omega(m^2)} \right\}, \tag{6.1}$$

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

We now bound the spectral norm of $M \circ P$ precisely as in [51, §6]. 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

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

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

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

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

$$\|M \circ P\| \leqslant \sqrt{s}\cdot 2^{-n}\left(\frac{N}{n}\right)^{-m/6} = 17^{-m}\, 2^{-n}\sqrt{s}.$$

Along with (6.1), this estimate shows that M has sign-rank at least $2^{\Omega(m)}$.

It remains to verify that M is a submatrix of $[f_{cm}(x,y)]_{x,y}$, where $c = \lceil\sqrt{8N/n}\rceil = \Theta(1)$. The set $V \in \mathcal{V}(N,n)$ in Definition 2.5 is naturally represented by a tuple $(\dots, v_{ij}, \dots) \in [N/n]^n$. Then for all $x \in \{0,1\}^N$ and $(V, w) \in \mathcal{V}(N,n) \times \{0,1\}^n$,

$$\bigwedge_{i=1}^{m} \bigvee_{j=1}^{4m^2} (x|_V \oplus w)_{ij} = \bigwedge_{i=1}^{m} \bigvee_{j=1}^{4m^2} \bigvee_{k=1}^{N/n} \bigvee_{\epsilon\in\{0,1\}} ((x_{ijk} = \epsilon) \wedge (w_{ij} \neq \epsilon) \wedge (v_{ij} = k)).$$

Merging the three groups of OR gates gives bottom fan-in $8m^2N/n \leqslant (cm)^2$. □

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

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

By property (2.3) of the Hadamard product, the sign-rank of $[f_m(x,y)]_{x,y}$ does not exceed $m^{2m} + 1 = 2^{O(m\log m)}$.

7. On communication complexity classes. We proceed to explore the consequences of our main result in the study of communication complexity classes. Throughout this section, the symbol $\{f_n\}$ shall stand for a family of functions $f_1, f_2, \dots, f_n, \dots$, where $f_n : \{0,1\}^n \times \{0,1\}^n \to \{-1,+1\}$.

The first class that we consider, UPP^{cc}, corresponds to functions with efficient unbounded-error protocols.

Definition 7.1 (Babai et al. [3, §4]). *A family $\{f_n\}$ is in UPP^{cc} iff for some constant c and all natural $n > c$, there exists a probabilistic communication protocol with private coins such that:*

(1) *for every input* (x, y), *the protocol outputs the correct value* $f_n(x, y)$ *with probability greater than* $1/2$*;*

(2) *the number of bits exchanged is at most* $\log^c n$.

Note that in this model the number of coin flips is not included in the complexity measure. Requiring the same bound of $\log^c n$ on the number of coin flips results in another extensively studied class [3], called PP^{cc}.

For our purposes, however, an equivalent matrix-analytic characterization [40] is more convenient. By the *sign-rank* of a function $f : X \times Y \to \{-1, +1\}$, where X, Y are finite sets, we shall mean the sign-rank of the matrix $[f(x, y)]_{x\in X,\, y\in Y}$.

Theorem 7.2 (Paturi & Simon [40]). *A function family* $\{f_n\}$ *is in the class* UPP^{cc} *iff for some constant* $c > 1$ *and all* $n > c$, *the function* f_n *has sign-rank at most* $2^{\log^c n}$.

We now turn to the communication-complexity analogue of the polynomial hierarchy, also defined by Babai et al. [3]. A function $f_n : \{0, 1\}^n \times \{0, 1\}^n \to \{-1, +1\}$ is called a *rectangle* if there exist subsets $A, B \subseteq \{0, 1\}^n$ such that

$$f_n(x, y) = -1 \quad \Leftrightarrow \quad x \in A,\ y \in B.$$

We call f_n the *complement of a rectangle* if the negated function $\neg f_n = -f_n$ is a rectangle.

Definition 7.3 (Babai et al. [3, §4]).

(1) *A family* $\{f_n\}$ *is in* Π_0^{cc} *iff each* f_n *is a rectangle. A family* $\{f_n\}$ *is in* Σ_0^{cc} *iff* $\{\neg f_n\}$ *is in* Π_0^{cc}.

(2) *Fix an integer* $k = 1, 2, 3, 4, \ldots$ *. A family* $\{f_n\}$ *is in* Σ_k^{cc} *iff for some constant* $c > 1$ *and all* $n > c$,

$$f_n = \bigvee_{i_1=1}^{2^{\log^c n}} \bigwedge_{i_2=1}^{2^{\log^c n}} \bigvee_{i_3=1}^{2^{\log^c n}} \cdots \bigodot_{i_k=1}^{2^{\log^c n}} g_n^{i_1,i_2,\ldots,i_k},$$

where $\bigodot = \bigvee$ (*resp.*, $\bigodot = \bigwedge$) *for* k *odd* (*resp., even*); *and each* $g_n^{i_1,i_2,\ldots,i_k}$ *is a rectangle* (*resp., the complement of a rectangle*) *for* k *odd* (*resp., even*). *A family* $\{f_n\}$ *is in* Π_k^{cc} *iff* $\{\neg f_n\}$ *is in* Σ_k^{cc}.

(3) *The polynomial hierarchy is given by* $\mathsf{PH}^{cc} = \bigcup_k \Sigma_k^{cc} = \bigcup_k \Pi_k^{cc}$, *where* $k = 0, 1, 2, 3, \ldots$ *ranges over all constants.*

(4) *A family* $\{f_n\}$ *is in* PSPACE^{cc} *iff for some constant* $c > 1$ *and all* $n > c$,

$$f_n = \bigvee_{i_1=1}^{2^{\log^c n}} \bigwedge_{i_2=1}^{2^{\log^c n}} \bigvee_{i_3=1}^{2^{\log^c n}} \cdots \bigvee_{i_k=1}^{2^{\log^c n}} g_n^{i_1,i_2,\ldots,i_k},$$

where $k < \log^c n$ *is odd and each* $g_n^{i_1,i_2,\ldots,i_k}$ *is a rectangle.*

Thus, the zeroth level (Σ_0^{cc} and Π_0^{cc}) of the polynomial hierarchy consists of rectangles and complements of rectangles, the simplest functions in communication complexity. The first level is easily seen to correspond to functions with efficient nondeterministic or co-nondeterministic protocols: $\Sigma_1^{cc} = \mathsf{NP}^{cc}$ and $\Pi_1^{cc} = \mathsf{coNP}^{cc}$. This level is well understood, and there exist powerful methods to show that $\{f_n\} \notin \Sigma_1^{cc}$ for a host of explicit functions $\{f_n\}$. Finding an explicit sequence $\{f_n\} \notin \Sigma_2^{cc}$, on the other hand, is a longstanding open problem.

The circuit class AC^0 is related to the polynomial hierarchy PH^{cc} in communication complexity in the obvious way. Specifically, if $f_n : \{0, 1\}^n \times \{0, 1\}^n \to \{-1, +1\}$,

$n = 1, 2, 3, 4, \dots$, is an AC^0 (or even quasi-AC^0) circuit family of depth k with an OR gate at the top (resp., AND gate), then $\{f_n\} \in \Sigma_{k-1}^{cc}$ (resp., $\{f_n\} \in \Pi_{k-1}^{cc}$). In particular, the depth-3 circuit family $\{f_n\}$ in Theorem 1.1 is in Π_2^{cc}, whereas $\{\neg f_n\}$ is in Σ_2^{cc}. In this light, Theorem 1.1 establishes the following separations:

THEOREM 7.4. $\Sigma_2^{cc} \not\subseteq \mathsf{UPP}^{cc}$, $\Pi_2^{cc} \not\subseteq \mathsf{UPP}^{cc}$.

Several years prior to our work, Forster [13] proved that the familiar inner product function $\text{IP}_n(x,y) = (-1)^{\sum x_i y_i}$ has sign-rank $2^{\Theta(n)}$. Since $\{\text{IP}_n\} \in \mathsf{PSPACE}^{cc}$, Forster's result yields the separation $\mathsf{PSPACE}^{cc} \not\subseteq \mathsf{UPP}^{cc}$. Theorem 7.4 in this paper substantially strengthens it, showing that even the second level $(\Sigma_2^{cc}, \Pi_2^{cc})$ of the polynomial hierarchy is not contained in UPP^{cc}. This settles the open problem due to Babai et al. [3, p. 345], who asked whether $\Sigma_2^{cc} \subseteq \mathsf{UPP}^{cc}$. Observe that Theorem 7.4 is best possible in that UPP^{cc} trivially contains Σ_0^{cc}, Σ_1^{cc}, Π_0^{cc}, and Π_1^{cc}.

The classes PP^{cc} and UPP^{cc} both correspond to small-bias communication and, in fact, were both inspired by the class PP in computational complexity. It is well-known and straightforward to show that $\mathsf{PP}^{cc} \subseteq \mathsf{UPP}^{cc}$. It turns out that UPP^{cc} is strictly more powerful than PP^{cc}, as shown by Buhrman et al. [8] and independently in [48]. In this light, Theorem 7.4 in this paper substantially strengthens earlier separations of Σ_2^{cc} and Π_2^{cc} from PP^{cc}, obtained independently in [8] and [49]:

THEOREM 7.5 (Buhrman et al. [8], Sherstov [49]). $\Sigma_2^{cc} \not\subseteq \mathsf{PP}^{cc}$, $\Pi_2^{cc} \not\subseteq \mathsf{PP}^{cc}$.

Sign-rank is a much more challenging quantity to analyze than *discrepancy*, a combinatorial complexity measure that is known [25] to characterize membership in PP^{cc}. Indeed, exponentially small upper bounds on the discrepancy were known more than twenty years ago [56, 10], whereas the first exponential lower bound on the sign-rank for an explicit function was only obtained recently by Forster [13]. It is not surprising, then, that this paper has required ideas that are quite different from the methods of both [8] and [48, 49].

8. 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' *statistical query model* [21]. A suitable *lower* bound on the spectral norm of every such matrix, on the other hand, would result in a breakthrough separation of PH^{cc} and PSPACE^{cc}. See [3, 43, 33, 48] 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* $\deg_\pm(f)$ is the least degree of a multivariate polynomial $p(x_1, \dots, x_n)$ such that $f(x) \equiv \operatorname{sgn} p(x)$. 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 $n^{O(\deg_\pm(f))}$. Is that upper bound close to optimal? Specifically, does M_f have sign-rank $\exp(\deg_\pm(f)^{\Omega(1)})$ 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, $f = \text{MP}$. For $f = \text{PARITY}$ the hypothesis follows immediately the seminal work of Forster [13]. More generally, it was proven in [51] that the hypothesis holds for all symmetric functions.

In the field of communication complexity, we were able to resolve the main question left open by Babai et al. [3], but only in one direction: $\mathsf{PH}^{cc} \not\subseteq \mathsf{UPP}^{cc}$. As we already noted, the other direction remains wide open despite much research: no lower bounds are known for PH^{cc} or even Σ_2^{cc}. The latter question is in turn equivalent to lower bounds for bounded-depth circuits in the context of *graph complexity* (e.g., see [41, 42, 19] and the literature cited therein). It is also known [43, Thm. 1] that sufficiently rigid matrices do not belong to PH^{cc}, which provides further motivation to be looking for lower bounds on matrix rigidity.

Acknowledgments. The authors would like to thank Scott Aaronson, Adam Klivans, Dieter van Melkebeek, Yaoyun Shi, and Ronald de Wolf for helpful feedback on an earlier version of this manuscript. We are also grateful to our anonymous reviewers for their useful comments.

REFERENCES

[1] Michael Alekhnovich, Mark Braverman, Vitaly Feldman, Adam R. Klivans, and Toniann Pitassi, *Learnability and automatizability*, in Proc. of the 45th Symposium on Foundations of Computer Science (FOCS), 2004, pp. 621–630.

[2] Noga Alon, Peter Frankl, and Vojtech Rödl, *Geometrical realization of set systems and probabilistic communication complexity*, in Proc. of the 26th Symposium on Foundations of Computer Science (FOCS), 1985, pp. 277–280.

[3] László Babai, Peter Frankl, and Janos Simon, *Complexity classes in communication complexity theory*, in Proc. of the 27th Symposium on Foundations of Computer Science (FOCS), 1986, pp. 337–347.

[4] Shai Ben-David, Nadav Eiron, and Hans Ulrich Simon, *Limitations of learning via embeddings in Euclidean half spaces*, J. Mach. Learn. Res., 3 (2003), pp. 441–461.

[5] Avrim Blum, Alan M. Frieze, Ravi Kannan, and Santosh Vempala, *A polynomial-time algorithm for learning noisy linear threshold functions*, Algorithmica, 22 (1998), pp. 35–52.

[6] Avrim Blum, Adam Kalai, and Hal Wasserman, *Noise-tolerant learning, the parity problem, and the statistical query model*, J. ACM, 50 (2003), pp. 506–519.

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

[8] Harry Buhrman, Nikolai K. Vereshchagin, and Ronald de Wolf, *On computation and communication with small bias*, in Proc. of the 22nd Conf. on Computational Complexity (CCC), 2007, pp. 24–32.

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

[10] Benny Chor and Oded Goldreich, *Unbiased bits from sources of weak randomness and probabilistic communication complexity*, SIAM J. Comput., 17 (1988), pp. 230–261.

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

[12] Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami, *New results for learning noisy parities and halfspaces*, in Proceedings of the 47th Annual Symposium on Foundations of Computer Science (FOCS), 2006, pp. 563–574.

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

[14] Jürgen Forster, Matthias Krause, Satyanarayana V. Lokam, Rustam Mubarakzjanov, Niels Schmitt, and Hans-Ulrich 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), 2001, pp. 171–182.

[15] Jürgen Forster and Hans Ulrich Simon, *On the smallest possible dimension and the largest possible margin of linear arrangements representing given concept classes*, Theor. Comput. Sci., 350 (2006), pp. 40–48.

[16] Mikael Goldmann, Johan Håstad, and Alexander A. Razborov, *Majority gates vs. general weighted threshold gates*, Computational Complexity, 2 (1992), pp. 277–300.

[17] András Hajnal, Wolfgang Maass, Pavel Pudlák, Mario Szegedy, and György Turán, *Threshold circuits of bounded depth*, J. Comput. Syst. Sci., 46 (1993), pp. 129–154.

[18] David P. Helmbold, Robert H. Sloan, and Manfred K. Warmuth, *Learning integer lattices*, SIAM J. Comput., 21 (1992), pp. 240–266.

[19] Stasys Jukna, *On graph complexity*, Combinatorics, Probability and Computing, 15 (2006), pp. 1–22.

[20] Bala Kalyanasundaram and Georg Schnitger, *The probabilistic communication complexity of set intersection*, SIAM J. Discrete Math., 5 (1992), pp. 545–557.

[21] Michael Kearns, *Efficient noise-tolerant learning from statistical queries*, in Proc. of the 25th Symposium on Theory of Computing (STOC), 1993, pp. 392–401.

[22] Michael Kearns and Leslie Valiant, *Cryptographic limitations on learning Boolean formulae and finite automata*, J. ACM, 41 (1994), pp. 67–95.

[23] Michael J. Kearns and Umesh V. Vazirani, *An Introduction to Computational Learning Theory*, MIT Press, Cambridge, 1994.

[24] Michael Kharitonov, *Cryptographic hardness of distribution-specific learning*, in Proc. of the 25th Symposium on Theory of Computing, 1993, pp. 372–381.

[25] Hartmut Klauck, *Lower bounds for quantum communication complexity*, SIAM J. Comput., 37 (2007), pp. 20–46.

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

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

[28] ———, *Unconditional lower bounds for learning intersections of halfspaces*, Machine Learning, 69 (2007), pp. 97–114.

[29] ———, *Cryptographic hardness for learning intersections of halfspaces*, J. Comput. Syst. Sci., 75 (2009), pp. 2–12.

[30] Matthias Krause and Pavel Pudlák, *On the computational power of depth-*2 *circuits with threshold and modulo gates*, Theor. Comput. Sci., 174 (1997), pp. 137–156.

[31] Eyal Kushilevitz and Noam Nisan, *Communication complexity*, Cambridge University Press, New York, 1997.

[32] Nati Linial, Shahar Mendelson, Gideon Schechtman, and Adi Shraibman, *Complexity measures of sign matrices*, Combinatorica, 27 (2007), pp. 439–463.

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

[34] Marvin L. Minsky and Seymour A. Papert, *Perceptrons: Expanded edition*, MIT Press, Cambridge, Mass., 1988.

[35] Ilan Newman, *Private vs. common random bits in communication complexity*, Inf. Process. Lett., 39 (1991), pp. 67–71.

[36] Noam Nisan, *The communication complexity of threshold gates*, in Combinatorics, Paul Erdős is Eighty, 1993, pp. 301–315.

[37] A. B. J. Novikoff, *On convergence proofs on perceptrons*, in Proc. of the Symposium on the Mathematical Theory of Automata, vol. XII, 1962, pp. 615–622.

[38] Ryan O'Donnell and Rocco A. Servedio, *New degree bounds for polynomial threshold functions*, in Proc. of the 35th Symposium on Theory of Computing (STOC), 2003, pp. 325–334.

[39] Ramamohan Paturi, *On the degree of polynomials that approximate symmetric Boolean functions*, in Proc. of the 24th Symposium on Theory of Computing, 1992, pp. 468–474.

[40] Ramamohan Paturi and Janos Simon, *Probabilistic communication complexity*, J. Comput. Syst. Sci., 33 (1986), pp. 106–123.

[41] Pavel Pudlák, Vojtech Rödl, and Petr Savický, *Graph complexity*, Acta Inf., 25 (1988), pp. 515–535.

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

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

[44] ———, *On the distributional complexity of disjointness*, Theor. Comput. Sci., 106 (1992), pp. 385–390.

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

[46] Frank Rosenblatt, *The perceptron: A probabilistic model for information storage and organization in the brain*, Psychological Review, 65 (1958), pp. 386–408.

[47] Alexander A. Sherstov, *Powering requires threshold depth* 3, Inf. Process. Lett., 102 (2007), pp. 104–107.

[48] ———, *Halfspace matrices*, Comput. Complex., 17 (2008), pp. 149–178. Preliminary version in 22nd CCC, 2007.

[49] ———, *Separating* AC^0 *from depth-2 majority circuits*, SIAM J. Comput., 38 (2009), pp. 2113–2129. Preliminary version in 39th STOC, 2007.

[50] ———, *The pattern matrix method for lower bounds on quantum communication*, in Proc. of the 40th Symposium on Theory of Computing (STOC), 2008, pp. 85–94.

[51] ———, *The unbounded-error communication complexity of symmetric functions*, in Proc. of the 49th Symposium on Foundations of Computer Science (FOCS), 2008, pp. 384–393.

[52] ———, *Communication lower bounds using dual polynomials*, Bulletin of the EATCS, 95 (2008), pp. 59–93.

[53] Nathan Srebro and Adi Shraibman, *Rank, trace-norm and max-norm.*, in Proc. of the 18th Conf. on Learning Theory (COLT), 2005, pp. 545–560.

[54] Jun Tarui and Tatsuie Tsukiji, *Learning DNF by approximating inclusion-exclusion formulae*, in Proc. of the 14th Conf. on Computational Complexity (CCC), 1999, pp. 215–221.

[55] Leslie G. Valiant, *A theory of the learnable*, Commun. ACM, 27 (1984), pp. 1134–1142.

[56] Umesh V. Vazirani, *Strong communication complexity or generating quasirandom sequences form two communicating semi-random sources*, Combinatorica, 7 (1987), pp. 375–392.

Appendix 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* [31, 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 [35].

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 DISJOINTNESS function on n-bit strings has complexity $O(\log n)$ in the unbounded-error model and $\Omega(n)$ in the randomized model [20, 44]. Furthermore, explicit functions are known [8, 48] 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 $f : X \times Y \to \{-1,+1\}$ 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 [31, 11] for rigorous definitions of these various models; our sole intention was to point out that the unbounded-error model is at least as powerful.

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

THEOREM B.1 (Forster [13, 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 proof of this result is rather elaborate and uses compactness arguments in an essential way.

The vector norm $\|\cdot\|$ above and throughout this section is the Euclidean norm $\|\cdot\|_2$. The development that follows is closely analogous to Forster's derivation [13, p. 617].

THEOREM 2.4 (Restated from p. 9). *Let X, Y be finite sets, $M = [M_{xy}]_{x\in X, y\in Y}$ 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:*

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

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

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

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

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, v_y are not already in general position, we can replace them with their slightly perturbed versions $\tilde{u}_x, \tilde{v}_y$ that *are* in general position. Provided that the perturbations are small enough, property (B.5) will still hold, i.e., we will have $\langle \tilde{u}_x, \tilde{v}_y\rangle M_{xy} > 0$ whenever $M_{xy} \neq 0$. As a result, we can assume w.l.o.g. that $\{u_x\}, \{v_y\}$ are in general position and in particular that all $\{v_y\}$ are nonzero.

Since sign-rank$(M) \leqslant$ 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:

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

where the last step follows from (B.6). □