# SEPARATING AC$^0$ FROM DEPTH-2 MAJORITY CIRCUITS[*]

ALEXANDER A. SHERSTOV[†]

**Abstract.** We construct a function in $\mathsf{AC}^0$ that cannot be computed by a depth-2 majority circuit of size less than $\exp(\Theta(n^{1/5}))$. This solves an open problem due to Krause and Pudlák (1997) and matches Allender's classic result (1989) that $\mathsf{AC}^0$ can be efficiently simulated by depth-3 majority circuits. To obtain our result, we develop a novel technique for proving lower bounds on communication complexity. This technique, the *Degree/Discrepancy Theorem,* is of independent interest. It translates lower bounds on the threshold degree of any Boolean function into upper bounds on the discrepancy of a related function. Upper bounds on the discrepancy, in turn, immediately imply lower bounds on communication and circuit size. In particular, we exhibit the first known function in $\mathsf{AC}^0$ with exponentially small discrepancy, $\exp(-\Omega(n^{1/5}))$, thereby establishing the separations $\Sigma_2^{cc} \not\subseteq \mathsf{PP}^{cc}$ and $\Pi_2^{cc} \not\subseteq \mathsf{PP}^{cc}$ in communication complexity.

**Key words.** Majority circuits, constant-depth AND/OR/NOT circuits, communication complexity, discrepancy, threshold degree of Boolean functions, Degree/Discrepancy Theorem.

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

## 1. Introduction.

A natural and important computational model is that of a polynomial-size circuit of majority gates. This model has been extensively studied for the past two decades [14, 15, 26, 27, 38, 43–45]. Research has shown that majority circuits of depth 2 and 3 already possess surprising computational power. Indeed, it is a longstanding open problem [21] to exhibit a Boolean function that *cannot* be computed by a depth-3 majority circuit of polynomial size. To illustrate, the arithmetic operations of powering, multiplication, and division on $n$-bit integer arguments can all be computed by depth-3 majority circuits of polynomial size [45]. An even more striking example is the addition of $n$ $n$-bit integers, which is computable by a depth-2 majority circuit of polynomial size [45]. Depth-2 majority circuits of polynomial size can also compute every symmetric function (such as PARITY) and every DNF and CNF formula of polynomial size.

This chief goal of this paper is to relate the computational power of majority circuits to that of $\mathsf{AC}^0$, another extensively studied class, which consists of polynomial-size constant-depth circuits of AND, OR, NOT gates. A well-known result due to Allender [1] states that every function in $\mathsf{AC}^0$ can be computed by a depth-3 majority circuit of quasipolynomial size. For over ten years, it has been an open problem to determine whether Allender's simulation is optimal. Specifically, Krause and Pudlák [21, §6] ask whether every function in $\mathsf{AC}^0$ can be computed by a depth-2 majority circuit of quasipolynomial size. We solve this open problem completely, even in the more general setting of majority-of-threshold circuits (i.e., depth-2 circuits in which a majority gate receives inputs from arbitrary linear threshold gates):

THEOREM 1.1 (Main Result). *There is a function $f : \{-1,1\}^n \to \{-1,1\}$, explicitly given and computable by an $\mathsf{AC}^0$ circuit of depth 3, whose computation requires a majority vote of $\exp(\Omega(n^{1/5}))$ linear threshold gates.*

In other words, Allender's simulation is optimal in a strong sense. The lower bound in Theorem 1.1 is an exponential improvement over previous work. The best

[*] This paper appeared in preliminary form in *Proceedings of the 39th Annual ACM Symposium on Theory of Computing* (STOC), 2007, pp. 294–301.

[†] Department of Computer Sciences, The University of Texas at Austin, TX 78712 USA (sherstov@cs.utexas.edu).

previous lower bound [15, 27] was quasipolynomial and followed trivially from the observation that $\mathsf{AC}^0$ can compute INNER PRODUCT MODULO 2 on $\log^c n$ variables, for any constant $c > 1$.

**1.1. A communication-complexity perspective.** A different and perhaps more revealing view of this work is in terms of communication complexity [23]. The communication complexity of Boolean functions in different models has long been an active area of research, due to its inherent appeal as a complexity subject as well as its numerous applications in theoretical computer science. Our work contributes a novel and powerful technique for communication lower bounds, which is based on the representation of a Boolean function as the sign of a real-valued polynomial. Specifically, fix a Boolean function $f : \{-1,1\}^n \to \{-1,1\}$. Its *threshold degree,* $\deg_\pm(f)$, is defined as the minimum degree of a polynomial $p(x_1, x_2, \ldots, x_n)$ that represents $f$ in sign: $f(x) \equiv \operatorname{sign}(p(x))$. This concept has played a prominent role in the study of circuit complexity [2, 21, 22, 26] and has yielded valuable insights in other areas, including computational learning theory [17, 20]. In many cases [26], it is straightforward to obtain strong lower bounds on the threshold degree. Since the threshold degree is a measure of the complexity of a given Boolean function, it is natural to wonder whether it can yield lower bounds on communication in a suitable setting. Our work confirms this intuition for every $f$.

More precisely, fix a Boolean function $f : \{-1,1\}^n \to \{-1,1\}$ with threshold degree $d$. Let $N$ be a given integer, $N \geqslant n$. We introduce and study the two-party communication problem of computing

$$f(x|_S),$$

where the Boolean string $x \in \{-1,1\}^N$ is Alice's input and the set $S \subset \{1, 2, \ldots, N\}$ of size $|S| = n$ is Bob's input. The symbol $x|_S$ stands for the projection of $x$ onto the indices in $S$, in other words, $x|_S = (x_{i_1}, x_{i_2}, \ldots, x_{i_n}) \in \{-1,1\}^n$, where $i_1 < i_2 < \cdots < i_n$ are the elements of $S$. Intuitively, this problem models a situation when Alice and Bob's joint computation depends on only $n$ of the inputs $x_1, x_2, \ldots, x_N$. Alice knows the values of all the inputs $x_1, x_2, \ldots, x_N$ but does not know which $n$ of them are relevant. Bob, on the other hand, knows which $n$ inputs are relevant but does not know their values. As one would hope, we prove that $d$ gives a lower bound on the communication requirements of this problem. We phrase our result in terms of *discrepancy,* a central quantity in communication complexity that immediately yields lower bounds on communication in a variety of models (see §2.1):

THEOREM 1.2 (Degree/Discrepancy Theorem). *Let $f : \{-1,1\}^n \to \{-1,1\}$ be given with threshold degree $d \geqslant 1$. Then for $N \geqslant n$, the matrix $M = [f(x|_S)]_{x,S}$ has discrepancy*

$$\operatorname{disc}(M) \leqslant \left(\frac{4\mathrm{e}n^2}{Nd}\right)^{d/2},$$

*where* $\mathrm{e} = 2.718\ldots.$

Theorem 1.2, which we call the *Degree/Discrepancy Theorem* for obvious reasons, is a separate contribution of our work. Given a function $f$ with threshold degree $d$, it generates a communication problem with discrepancy at most $2^{-d}$ (by setting $N \geqslant 16\mathrm{e}n^2/d$). This exponentially small upper bound on the discrepancy immediately translates in an $\Omega(d)$ lower bound on communication in a variety of models (deterministic, nondeterministic, randomized, quantum with and without entanglement). Moreover, the resulting lower bounds on communication remain valid when Alice and Bob merely seek to predict the answer with nonnegligible advantage, a critical aspect for lower bounds against threshold circuits. This contrasts with other communication-complexity methods [32, 34], that only apply to high-accuracy computation. Finally, discrepancy arises prominently in contexts beyond communication complexity, such as estimation of margin complexity [25, 40] and approximate rank [19]. For all these reasons, we believe that the Degree/Discrepancy Theorem is of considerable interest in its own right. We prove it by a novel application of Gordan's Transposition Theorem [37, §7.8], which is a classical result from the theory of linear inequalities.

We are now in a position to outline the proof of our main result, Theorem 1.1. We start with a well-known DNF formula $\Phi$, constructed by Minsky and Papert [26], that has high threshold degree. An application of Theorem 1.2 to $\Phi$ yields a communication problem with low discrepancy. By design, this communication problem can be viewed as an $\mathsf{AC}^0$ circuit of depth 3. Recalling that its discrepancy is exponentially small, we immediately conclude that it cannot be computed by a depth-2 majority circuit of subexponential size. This completes the proof.

**1.2. On the discrepancy of $\mathsf{AC}^0$ circuits.** Recall that a key component of our proof is the construction of an $\mathsf{AC}^0$ circuit with exponentially small discrepancy. Prior to this work, it was not known whether such a circuit existed. In particular, all previously known functions with exponentially small discrepancy (e.g., [14, 27]) contain PARITY or MAJORITY as a subfunction and therefore cannot be computed in $\mathsf{AC}^0$. In view of the intrinsic value of discrepancy as a complexity measure, we state this result on its own.

THEOREM 1.3 (Discrepancy of $\mathsf{AC}^0$ circuits). *There is a function $f : \{-1,1\}^n \times \{-1,1\}^n \to \{-1,1\}$, explicitly given and computable by an $\mathsf{AC}^0$ circuit of depth 3, that has discrepancy $\exp(-\Omega(n^{1/5}))$ with respect to an explicitly given distribution.*

Theorem 1.3 is best possible in that every $\mathsf{AC}^0$ circuit of depth 1 or 2 has discrepancy at least $n^{-O(1)}$ with respect to all distributions and all partitions of the variables (see §5). As a direct corollary to Theorem 1.3, we separate communication classes $\Sigma_2^{cc}$ and $\Pi_2^{cc}$ from $\mathsf{PP}^{cc}$:

COROLLARY 1.4. $\Sigma_2^{cc} \not\subseteq \mathsf{PP}^{cc}, \quad \Pi_2^{cc} \not\subseteq \mathsf{PP}^{cc}.$

Here $\Sigma_2^{cc}$ and $\Pi_2^{cc}$ are the second level of the polynomial hierarchy in communication complexity, whereas $\mathsf{PP}^{cc}$ is the class of all communication matrices with nonnegligible discrepancy. Prior to this work, it was entirely conceivable that $\mathsf{PP}^{cc}$ contained both $\Sigma_2^{cc}$ and $\Pi_2^{cc}$ and the rest of the polynomial hierarchy $\mathsf{PH}^{cc}$. See §6 for further details.

Another $\mathsf{AC}^0$ circuit of depth 3 with exponentially small discrepancy was constructed independently by Buhrman, Vereshchagin, and de Wolf [5, §3]. Their proof uses quite different techniques (approximation theory and quantum communication complexity). The circuit of Buhrman et al. has discrepancy $\exp(-\Omega(n^{1/3}))$, which is a stronger bound than Theorem 1.3. An advantage of the construction in this paper is that it is self-contained and from first principles, whereas the work of Buhrman et al. builds on a subtle result of Razborov [35]. In addition, our method is not restricted to $\mathsf{AC}^0$ but, rather, applies to *any* function with high threshold degree.

**1.3. Recent progress.** We are pleased to report that the Degree/Discrepancy Theorem has inspired important progress in communication complexity by several researchers, which we briefly survey. The first of these new results is concerned with bounded-error communication. By combining the method of Theorem 1.2 with techniques from matrix analysis and approximation theory, the author [41] obtained strong lower bounds on bounded-error communication for a broad new class of problems. These lower bounds remain valid in the quantum model (regardless of prior entanglement) and subsume Razborov's breakthrough lower bounds for symmetric functions [35].

In another development [42], we used the method of Theorem 1.2 as a starting point to derive essentially optimal lower bounds on the *unbounded-error* communication complexity of every symmetric function. The unbounded-error model is more powerful than all the familiar models of communication (both classical and quantum), and proving lower bounds in it is a substantial challenge. The only previous nontrivial lower bounds for this model appeared in the groundbreaking work of Forster [12] and its extensions.

Razborov and Sherstov [36] recently obtained the first exponential lower bound on the *sign-rank* of $\mathsf{AC}^0$, thereby solving a longstanding open problem due to Babai, Frankl, and Simon [3]. The results in [36] additionally give strong lower bounds for PAC learning polynomial-size DNF and CNF formulas (see Remark 8.2). The method of the Degree/Discrepancy Theorem is one of the starting points in that work.

The Degree/Discrepancy Theorem has also found applications to multiparty communication complexity. The first of these is work by Chattopadhyay [6], who observed that the method of Theorem 1.2 adapts in a straightforward manner to the multiparty model. Analogous to this adaptation, Lee and Shraibman [24] and Chattopadhyay and Ada [7] adapted to the multiparty model the author's recent work [41] on two-party bounded-error communication. They thereby obtained improved lower bounds on the bounded-error communication complexity of DISJOINTNESS for up to $\log\log n$ players. David and Pitassi [9] combined this line of work with the probabilistic method, establishing a separation of the communication classes $\mathsf{NP}_k^{cc}$ and $\mathsf{BPP}_k^{cc}$ for up to $k = (1-\epsilon)\log n$ players. Their construction was derandomized in a follow-up paper by David, Pitassi, and Viola [10], resulting in an explicit separation. See the survey article [39] for a unified guide to these results, complete with all the key proofs.

**1.4. Organization.** The remainder of this paper is organized as follows. Section 2 provides relevant background on communication complexity and threshold functions. Section 3 is devoted to the proof of the Degree/Discrepancy Theorem, our main technical tool. Section 4 studies a particular $\mathsf{AC}^0$ circuit $\Phi$ with high threshold degree. Section 5 applies the Degree/Discrepancy Theorem to $\Phi$, yielding an explicit $\mathsf{AC}^0$ circuit $f$ of depth 3 with exponentially small discrepancy. Section 6 uses this discrepancy result to separate the classes $\Sigma_2^{cc}$ and $\Pi_2^{cc}$ from $\mathsf{PP}^{cc}$ in communication complexity. In Section 7, we derive our main result, an exponential lower bound on the size of depth-2 majority circuits that compute $\mathsf{AC}^0$. Section 8 concludes with an application of our results to computational learning theory.

## 2. Preliminaries.

Throughout this work, we identify $-1$ and $1$ with "true" and "false," respectively. We view Boolean functions as mappings $X \to \{-1,1\}$, where $X$ is a finite set such as $X = \{-1,1\}^n$. The symbol $[n]$ stands for the set $\{1,2,\dots,n\}$. For integers $N, n$ with $N \geqslant n$, the symbol $\binom{[N]}{n}$ denotes the family of all size-$n$ subsets of $\{1,2,\dots,N\}$. For a string $x \in \{-1,1\}^N$ and a set $S \in \binom{[N]}{n}$, we define $x|_S = (x_{i_1}, x_{i_2}, \dots, x_{i_n}) \in \{-1,1\}^n$, where $i_1 < i_2 < \cdots < i_n$ are the elements of $S$. The notation $\mathbb{R}^{m\times n}$ refers to the family of all $m \times n$ matrices with real entries. We specify matrices by their generic entry, e.g., $A = [F(i,j)]_{i,j}$. As usual, we denote the base of the natural logarithm by $\mathrm{e} = 2.718\ldots$.

Recall that $\mathsf{AC}^0$ is the family of polynomial-size unbounded-fanin constant-depth circuits with AND, OR, NOT gates. We adopt the standard definition of the sign function:

$$\operatorname{sign}(t) = \begin{cases} 1 & \text{if } t > 0, \\ 0 & \text{if } t = 0, \\ -1 & \text{if } t < 0. \end{cases}$$

A *linear threshold gate* is a Boolean function $f : \{-1,1\}^n \to \{-1,1\}$ of the form $f(x) \equiv \operatorname{sign}(\sum_{i=1}^n a_i x_i - \theta)$ for some fixed reals $a_1, \ldots, a_n, \theta$. Observe that a linear threshold gate generalizes the familiar majority gate.

**2.1. Communication complexity.** We consider Boolean functions $f : X \times Y \to \{-1,1\}$. Typically $X = Y = \{-1,1\}^n$, but we also allow $X$ and $Y$ to be arbitrary finite sets. We identify a function $f$ with its *communication matrix* $M = [f(x,y)]_{x\in X,\, y\in Y}$. In particular, we use the terms "communication complexity of $f$" and "communication complexity of $M$" interchangeably (and likewise for other complexity measures, such as discrepancy). The two communication models of interest to us are the randomized model and the deterministic model. The *randomized complexity* $\mathrm{R}_{1/2-\gamma/2}(f)$ of $f$ is the minimum cost of a randomized protocol for $f$ that computes $f(x,y)$ correctly with probability at least $\frac{1}{2} + \frac{1}{2}\gamma$ (equivalently, with *advantage* $\gamma$) on every input $(x,y)$. The *public-coin randomized complexity* $\mathrm{R}^{\mathrm{pub}}_{1/2-\gamma/2}(f)$ is defined analogously with the only difference that the communicating parties now have a source of shared random bits, i.e., they can observe tosses of a common coin without communicating. The *distributional complexity* $\mathrm{D}^{\mu}_{1/2-\gamma/2}(f)$ is the minimum cost of a deterministic protocol for $f$ that has error at most $\frac{1}{2} - \frac{1}{2}\gamma$ (equivalently, *advantage* $\gamma$) with respect to the distribution $\mu$ over the inputs.

A *rectangle* of $X \times Y$ is any set $R = A \times B$ with $A \subseteq X$ and $B \subseteq Y$. For a fixed distribution $\mu$ over $X \times Y$, the *discrepancy* of $f$ is defined as

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

where the maximum is taken over all rectangles $R$. We define $\operatorname{disc}(f) = \min_\mu \operatorname{disc}_\mu(f)$. The *discrepancy method* is a fundamental technique that places a lower bound on the randomized and distributional complexity in terms of the discrepancy:

Proposition 2.1 (Kushilevitz and Nisan [23, pp. 36–38]). *For every Boolean function $f : X \times Y \to \{-1,1\}$, every distribution $\mu$ on $X \times Y$, and every $\gamma > 0$,*

$$\mathrm{R}_{1/2-\gamma/2}(f) \;\geqslant\; \mathrm{R}^{\mathrm{pub}}_{1/2-\gamma/2}(f) \;\geqslant\; \mathrm{D}^{\mu}_{1/2-\gamma/2}(f) \;\geqslant\; \log_2 \frac{\gamma}{\operatorname{disc}_\mu(f)}.$$

The above definition of discrepancy is not convenient to work with. The following well-known lemma bounds the discrepancy in terms of a more analytically pleasing quantity. For completeness, we include a proof.

Lemma 2.2 (Discrepancy bound; cf. [4, 8, 11, 33]). *Let $f : X \times Y \to \{-1,1\}$ be given, and let $\mu$ be a probability distribution over $X \times Y$. Then*

$$\operatorname{disc}_\mu(f)^2 \leqslant |X| \sum_{y,y' \in Y} \left| \sum_{x \in X} \mu(x,y)\mu(x,y')f(x,y)f(x,y') \right| .$$

*Proof.* (Adapted from Raz [33].) Let $R = A \times B$ be a rectangle over which the discrepancy is achieved. Define $\alpha_x = 1$ for all $x \in A$, and likewise $\beta_y = 1$ for all $y \in B$. For all remaining $x$ and $y$, let $\alpha_x$ and $\beta_y$ be independent random variables distributed uniformly over $\{-1,1\}$. Passing to expectations,

$$\begin{aligned}
&\left| \mathbf{E}\left[ \sum_{x,y} \alpha_x \beta_y \mu(x,y) f(x,y) \right] \right| \\
&\qquad = \left| \sum_{(x,y)\in R} \underbrace{\mathbf{E}[\alpha_x\beta_y]}_{=1} \mu(x,y)f(x,y) + \sum_{(x,y)\notin R} \underbrace{\mathbf{E}[\alpha_x\beta_y]}_{=0} \mu(x,y)f(x,y) \right| \\
&\qquad = \operatorname{disc}_\mu(M).
\end{aligned}$$

In particular, there exists a fixed assignment $\alpha_x, \beta_y \in \{-1,1\}$ for all $x, y$ such that

$$\operatorname{disc}_\mu(f) \leqslant \left| \sum_{x,y} \alpha_x \beta_y \mu(x,y) f(x,y) \right| .$$

Squaring both sides and applying the Cauchy-Schwarz inequality,

$$\begin{aligned}
\operatorname{disc}_\mu(f)^2 &\leqslant |X| \sum_x \left( \alpha_x \sum_y \beta_y \mu(x,y) f(x,y) \right)^2 \\
&= |X| \sum_{y,y'} \beta_y \beta_{y'} \sum_x \mu(x,y)\mu(x,y')f(x,y)f(x,y') \\
&\leqslant |X| \sum_{y,y'} \left| \sum_x \mu(x,y)\mu(x,y')f(x,y)f(x,y') \right| ,
\end{aligned}$$

as desired. □

A definitive resource for further details is the book of Kushilevitz and Nisan [23].

**2.2. Threshold degree.** Let $f : \{-1,1\}^n \to \{-1,1\}$ be a given Boolean function. The *threshold degree* of $f$, denoted $\deg_\pm(f)$, is the least degree of a polynomial $p(x_1, x_2, \ldots, x_n)$ such that $f(x) \equiv \operatorname{sign}(p(x))$. In view of the domain of $f$, any such polynomial $p$ can be assumed to be multilinear. Note that any function $f$ that depends on $k$ or fewer of the $n$ variables has threshold degree at most $k$. For a set $S \subseteq [n]$, we write $\chi_S = \prod_{i\in S} x_i$. In this notation, the threshold degree of $f$ is the smallest $d$ such that $f(x) \equiv \operatorname{sign}(\sum_{|S|\leqslant d} a_S \chi_S(x))$ for some fixed reals $a_S$. Threshold degree is also known in the literature as "strong degree" [2], "voting polynomial degree" [21], "PTF degree" [30], and "sign degree" [5].

Crucial to understanding the threshold degree is the following result from the theory of linear inequalities, which follows in a straightforward manner from linear-programming duality.

THEOREM 2.3 (Gordan's Transposition Theorem [37, §7.8]). *Let $A \in \mathbb{R}^{m\times n}$. Then exactly one of the following statements holds:*

(i) *$u^{\mathsf{T}}A > 0$ for some vector $u$;*

(ii) *$Av = 0$ for some nonzero vector $v \geqslant 0$.*

The vector notation $u^{\mathsf{T}}A > 0$ and $v \geqslant 0$ above is to be understood entrywise, as usual. A consequence of Gordan's Transposition Theorem is the following well-known result regarding threshold representations:

THEOREM 2.4 (see [2, 30]). *Let $\phi_1, \phi_2, \ldots, \phi_k : \{-1,1\}^n \to \mathbb{R}$ be arbitrary real functions, and let $f : \{-1,1\}^n \to \{-1,1\}$ be a given Boolean function. Then exactly one of the following statements holds:*

(i) *$f(x) \equiv \operatorname{sign}(\sum_{i=1}^k a_i\phi_i(x))$ for some reals $a_1, a_2, \ldots, a_k$;*

(ii) *there is a distribution $\mu$ over $\{-1,1\}^n$ such that*

$$\mathop{\mathbf{E}}_{x\sim\mu}[f(x)\phi_i(x)] = 0, \qquad i = 1, 2, \ldots, k.$$

*Proof.* Consider the $k \times 2^n$ matrix $A = [f(x)\phi_i(x)]_{i,x}$. The claim follows from Theorem 2.3, with $u$ playing the role of a set of coefficients $(a_1, a_2, \ldots, a_k) \in \mathbb{R}^k$, and $v$ playing the role of a probability distribution. □

COROLLARY 2.5. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be arbitrary, $d$ a nonnegative integer. Then exactly one of the following holds:*

(i) *$\deg_\pm(f) \leqslant d$;*

(ii) *there is a distribution $\mu$ over $\{-1,1\}^n$ such that*

$$\mathop{\mathbf{E}}_{x\sim\mu}[f(x)\chi_S(x)] = 0, \qquad |S| = 0, 1, \ldots, d.$$

## 3. The Degree/Discrepancy Theorem.

This section marks the beginning of our proof. Its purpose is to establish Theorem 1.2 (the Degree/Discrepancy Theorem), which plays a central role in the development to follow.

THEOREM 1.2 (Restated from p. 2). *Let $f : \{-1,1\}^n \to \{-1,1\}$ be given with threshold degree $d \geqslant 1$. Let $N$ be a given integer, $N \geqslant n$. Define $M = [f(x|_S)]_{x,S}$, where the indices range as follows: $x \in \{-1,1\}^N$, $S \in \binom{[N]}{n}$. Then*

$$\operatorname{disc}(M) \leqslant \left(\frac{4en^2}{Nd}\right)^{d/2}. \tag{3.1}$$

*Proof.* Let $\mu$ be a probability distribution over $\{-1,1\}^n$ with respect to which $\mathbf{E}_{z\sim\mu}[f(z)p(z)] = 0$ for every real-valued function $p$ of $d-1$ or fewer of the variables $z_1, \ldots, z_n$. The existence of $\mu$ is assured by Corollary 2.5. Throughout this proof, the symbol $\mathcal{U}$ shall stand for the uniform distribution over the relevant domain. We will analyze the discrepancy of $M$ with respect to the distribution

$$\lambda(x, S) = 2^{-N+n}\binom{N}{n}^{-1}\mu(x|_S).$$

By Lemma 2.2,

$$\operatorname{disc}_\lambda(M)^2 \leqslant 4^n \mathop{\mathbf{E}}_{(S,T)\sim\mathcal{U}}|\Gamma(S,T)|, \tag{3.2}$$

where we let

$$\Gamma(S,T) = \mathop{\mathbf{E}}_{x\sim\mathcal{U}}\Big[\mu(x|_S)\mu(x|_T)f(x|_S)f(x|_T)\Big].$$

To analyze this expression, we prove two key claims.

Claim 3.2. *Assume that $|S\cap T| \leqslant d-1$. Then $\Gamma(S,T)=0$.*

*Proof.* For notational convenience, assume that $S=\{1,2,\dots,n\}$. Then

$$\begin{aligned}\Gamma(S,T) &= \mathop{\mathbf{E}}_{x\sim\mathcal{U}}\Big[\mu(x_1,\dots,x_n)\mu(x|_T)f(x_1,\dots,x_n)f(x|_T)\Big]\\ &= \frac{1}{2^N}\sum_{x_1,\dots,x_n}\mu(x_1,\dots,x_n)f(x_1,\dots,x_n)\sum_{x_{n+1},\dots,x_N}\mu(x|_T)f(x|_T)\\ &= \frac{1}{2^N}\mathop{\mathbf{E}}_{(x_1,\dots,x_n)\sim\mu}\left[f(x_1,\dots,x_n)\cdot\underbrace{\left(\sum_{x_{n+1},\dots,x_N}\mu(x|_T)f(x|_T)\right)}_{*}\right].\end{aligned}$$

Since $|S\cap T|\leqslant d-1$, the starred expression is a real-valued function of at most $d-1$ variables. The claim follows by the definition of $\mu$. □

Claim 3.3. *Assume that $|S\cap T|=k$. Then $|\Gamma(S,T)|\leqslant 2^{k-2n}$.*

*Proof.* For notational convenience, assume that

$$\begin{aligned}&S=\{1,2,\dots,n\},\\ &T=\{1,2,\dots,k\}\cup\{n+1,n+2,\dots,n+(n-k)\}.\end{aligned}$$

We have:

$$\begin{aligned}|\Gamma(S,T)| &\leqslant \mathop{\mathbf{E}}_{x\sim\mathcal{U}}\Big|\mu(x|_S)\mu(x|_T)f(x|_S)f(x|_T)\Big|\\ &= \mathop{\mathbf{E}}_{x_1,\dots,x_{2n-k}}[\mu(x_1,\dots,x_n)\mu(x_1,\dots,x_k,x_{n+1},\dots,x_{2n-k})]\\ &\leqslant \underbrace{\mathop{\mathbf{E}}_{x_1,\dots,x_n}[\mu(x_1,\dots,x_n)]}_{=2^{-n}}\cdot\max_{x_1,\dots,x_k}\underbrace{\mathop{\mathbf{E}}_{x_{n+1},\dots,x_{2n-k}}[\mu(x_1,\dots,x_k,x_{n+1},\dots,x_{2n-k})]}_{\leqslant 2^{-(n-k)}}.\end{aligned}$$

The bounds $2^{-n}$ and $2^{-(n-k)}$ follow because $\mu$ is a probability distribution. □

In view of Claims 3.2 and 3.3, inequality (3.2) simplifies to

$$\operatorname{disc}_\lambda(M)^2 \leqslant \sum_{k=d}^{n} 2^k\,\mathbf{P}[|S\cap T|=k].$$

Since

$$\mathbf{P}[|S\cap T|=k] = \binom{n}{k}\binom{N-n}{n-k}\binom{N}{n}^{-1} \leqslant \binom{n}{k}\left(\frac{n}{N}\right)^k \leqslant \left(\frac{\mathrm{e}n^2}{Nk}\right)^k,$$

and since the discrepancy cannot exceed 1, we conclude that

$$\operatorname{disc}_\lambda(M)^2 \leqslant \min\left\{1,\ \sum_{k=d}^{n}\left(\frac{2\mathrm{e}n^2}{Nk}\right)^k\right\} \leqslant \left(\frac{4\mathrm{e}n^2}{Nd}\right)^d.$$

This completes the proof of Theorem 1.2. □

REMARK 3.4. The proof above analyzes the discrepancy of $M = [f(x|_S)]_{x,S}$ with respect to a certain distribution $\lambda$, derived directly from a distribution $\mu$ under which $f$ is uncorrelated with every function of fewer than $\deg_\pm(f)$ variables. Therefore, the discrepancy bound (3.1) is achieved with respect to an explicitly given distribution whenever $\mu$ is explicitly given.

REMARK 3.5. The discrepancy bound (3.1) holds not only for $M$ but also for any sign matrix that contains $M$. This observation is immediate from the definition of discrepancy. It allows a considerable degree of flexibility in applying Theorem 1.2, as will become apparent in §5.

REMARK 3.6. The discrepancy bound in Theorem 1.2 is not tight. In subsequent work [41, Thm. 7.3], the author strengthened it to: $\operatorname{disc}(M) \leqslant (4n/N)^{d/2}$. Moreover, this new bound continues to hold when Bob's input $S$ is restricted to have a particularly simple form. This stronger Degree/Discrepancy Theorem leads to quantitative improvements on this paper's Theorems 1.1 and 1.3; see [41, §7] for details. These improvements are built around a matrix-analytic approach to estimating the discrepancy, as opposed to the combinatorial derivation above. However, the proof of the Degree/Discrepancy Theorem in this paper has the advantage that it easily adapts to the multiparty model and is the foundation of the recent multiparty results [6, 7, 9, 10, 24]. The matrix-analytic approach does not seem to extend to three or more communicating parties.

**4. A function with high threshold degree.** Consider the Boolean function $\mathrm{MP}_m$ on $n = 4m^3$ variables, given by

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

A moment's reflection shows that the threshold degree of $\mathrm{MP}_m$ is at most $m$. Indeed,

$$\mathrm{MP}_m(x) = \operatorname{sign}\left\{ -\frac{1}{2} + \prod_{i=1}^{m} (4m^2 + x_{i,1} + x_{i,2} + \cdots + x_{i,4m^2}) \right\}.$$

(Recall that $x_{i,j} \in \{-1, 1\}$, where $-1$ corresponds to "true.") Minsky and Papert [26], who originally defined this function, proved that this upper bound is tight.

THEOREM 4.1 (Minsky and Papert [26]). *$\mathrm{MP}_m$ has threshold degree $m$.*

Minsky and Papert's proof, while short and elegant, does not yield an explicit distribution over $\{-1, 1\}^{4m^3}$ with respect to which $\mathrm{MP}_m$ is orthogonal to all functions of fewer than $m$ variables. The existence of such a distribution is assured by Corollary 2.5. The purpose of this section is to construct it. While this construction is not needed for our circuit lower bound (Theorem 1.1), it yields additional insight into the discrepancy of $\mathsf{AC}^0$ (Theorem 1.3).

We shall construct the desired distribution by extending an earlier argument, due to O'Donnell and Servedio [29], that makes the crux of the Minsky-Papert construction explicit. A starting point in our discussion is the following fact.

PROPOSITION 4.2 (O'Donnell and Servedio [29]). *Let $\nu$ be the binomial distribution over $\{0, 1, \ldots, 2m\}$, i.e., $\nu(t) = 2^{-2m}\binom{2m}{t}$. Then for every polynomial $p$ of degree at most $2m - 1$,*

$$\mathop{\mathbf{E}}_{t\sim\nu}[(-1)^t p(t)] = 0.$$

*Proof.* We present the proof from [29]. The claim holds for the monomials $p = 1, t, t^2, \ldots, t^{2m-1}$ in view of the combinatorial identity

$$\sum_{t=0}^{2m} \binom{2m}{t} (-1)^t t^d = 0, \qquad\qquad d = 0, 1, \ldots, 2m-1.$$

By linearity of expectation, this completes the proof. $\square$

O'Donnell and Servedio used Proposition 4.2 to obtain an explicit distribution over $\{0, 1, \ldots, 2m\}$ under which every low-degree symmetric polynomial has zero correlation with $\mathrm{MP}_m$. However, what we seek is an explicit distribution over $\{-1, 1\}^{4m^3}$. To this end, we take the argument of O'Donnell and Servedio a step further. The technical exposition follows.

For $t = 0, 1, \ldots, 2m$, define

$$X_t = \left\{ x : \quad \sum_{j=1}^{4m^2} \frac{1 - x_{i,j}}{2} = 4m^2 - (t - (2i-1))^2 \quad \text{for } i = 1, 2, \ldots, m \right\}. \tag{4.1}$$

Thus, $X_0, X_1, \ldots, X_{2m}$ are disjoint sets of inputs. The same sets of inputs figure in previous analyses [26, 29]. It is easy to verify that for $t = 0, 1, \ldots, 2m$,

$$x \in X_t \qquad \Longrightarrow \qquad \mathrm{MP}_m(x) = (-1)^t. \tag{4.2}$$

Let $\nu$ be the distribution from Proposition 4.2. We will work with the following distribution $\mu$ over $\{-1, 1\}^{4m^3}$:

$$\mu(x) = \begin{cases} \nu(0)/|X_0| & \text{if } x \in X_0, \\ \nu(1)/|X_1| & \text{if } x \in X_1, \\ \quad\vdots & \\ \nu(2m)/|X_{2m}| & \text{if } x \in X_{2m}, \\ 0 & \text{otherwise.} \end{cases} \tag{4.3}$$

We are now in a position to prove the main result of this section.

THEOREM 4.3 (Explicit distribution for $\mathrm{MP}_m$). *Let $\mu$ be given by* (4.3). *Then*

$$\mathop{\mathbf{E}}_{\mu}[\mathrm{MP}_m \cdot \chi_S] = 0, \qquad\qquad |S| = 0, 1, \ldots, m-1.$$

*Proof.* Let $\chi_S$ be arbitrary with $|S| \leqslant m-1$. Call the variables $x_{i,1}, x_{i,2}, \ldots, x_{i,4m^2}$ the *ith block* of $x$. Let $\sigma_1, \sigma_2, \ldots, \sigma_m$ be fixed permutations for blocks $1, 2, \ldots, m$, respectively. The theorem follows immediately from the following two claims.

CLAIM 4.4. $\mathbf{E}_\mu[\mathrm{MP}_m \cdot (\chi_S \circ (\sigma_1, \ldots, \sigma_m))] = \mathbf{E}_\mu[\mathrm{MP}_m \cdot \chi_S]$ *for all* $\sigma_1, \ldots, \sigma_m$.

CLAIM 4.5. $\sum_{\sigma_1, \ldots, \sigma_m} \mathbf{E}_\mu[\mathrm{MP}_m \cdot (\chi_S \circ (\sigma_1, \ldots, \sigma_m))] = 0.$

We prove these claims below. This completes the proof of the theorem. $\square$

*Proof of Claim* 4.4. The functions $\mathrm{MP}_m(x)$ and $\mu(x)$ depend only on the sum of the bits in each block. Formally, $\mathrm{MP}_m \equiv \mathrm{MP}_m \circ (\sigma_1, \ldots, \sigma_m)$ and $\mu \equiv \mu \circ (\sigma_1, \ldots, \sigma_m)$. The claim follows. $\square$

*Proof of Claim* 4.5. Write $\chi_S = \chi_{S_1}\chi_{S_2}\cdots\chi_{S_m}$, where

$$S_i = S \cap \{(i,1),\ldots,(i,4m^2)\}, \qquad i = 1,2,\ldots,m.$$

Then,

$$\begin{aligned}\sum_{\sigma_1,\ldots,\sigma_m} \mathop{\mathbf{E}}_{\mu}[\mathrm{MP}_m \cdot (\chi_S \circ (\sigma_1,\ldots,\sigma_m))] &= \sum_{\sigma_1,\ldots,\sigma_m} \mathop{\mathbf{E}}_{\mu}\left[\mathrm{MP}_m \cdot \prod_{i=1}^{m}(\chi_{S_i}\circ\sigma_i)\right] \\ &= \mathop{\mathbf{E}}_{\mu}\left[\mathrm{MP}_m \cdot \prod_{i=1}^{m}\left(\sum_{\sigma_i}\chi_{S_i}\circ\sigma_i\right)\right] \\ &= \mathop{\mathbf{E}}_{\mu}\left[\mathrm{MP}_m \cdot \prod_{i=1}^{m} p_i(x_{i,1}+x_{i,2}+\cdots+x_{i,4m^2})\right],\end{aligned}$$

where $p_1,p_2,\ldots,p_m$ are polynomials of degree at most $|S_1|,|S_2|,\ldots,|S_m|$, respectively. We now use the definition of $\mu$ to simplify the last equation.

$$\begin{aligned}&\mathop{\mathbf{E}}_{\mu}\left[\mathrm{MP}_m \cdot \prod_{i=1}^{m} p_i(x_{i,1}+x_{i,2}+\cdots+x_{i,4m^2})\right] \\ &\quad= \sum_{x}\mu(x)\mathrm{MP}_m(x)\prod_{i=1}^{m} p_i(x_{i,1}+x_{i,2}+\cdots+x_{i,4m^2}) \\ &\quad= \sum_{t=0}^{2m}\sum_{x\in X_t}\frac{\nu(t)}{|X_t|}\mathrm{MP}_m(x)\prod_{i=1}^{m} p_i(x_{i,1}+x_{i,2}+\cdots+x_{i,4m^2}) \\ &\quad= \sum_{t=0}^{2m}\sum_{x\in X_t}\frac{\nu(t)}{|X_t|}(-1)^t\underbrace{\prod_{i=1}^{m} p_i(2[t-(2i-1)]^2-4m^2)}_{\text{call this } p(t)} \qquad \text{by (4.1), (4.2)} \\ &\quad= \sum_{t=0}^{2m}\nu(t)(-1)^t p(t) \\ &\quad= 0,\end{aligned}$$

where the last equality follows by Proposition 4.2 since $p(t)$ has degree at most $2\sum_i |S_i| = 2|S| \leqslant 2m-2$. □

**5. Discrepancy of $\mathsf{AC}^0$ circuits.** This section proves an exponentially small upper bound on the discrepancy of an explicit function in $\mathsf{AC}^0$.

THEOREM 1.3 (Rephrased from p. 3). *There exists a function $f : \{-1,1\}^N \times \{-1,1\}^N \to \{-1,1\}$, explicitly given and computable by an $\mathsf{AC}^0$ circuit of depth 3, that has discrepancy $\exp(-\Omega(N^{1/5}))$ with respect to an explicitly given distribution.*

*Proof.* Consider the function $\mathrm{MP}_m$ on $n = 4m^3$ variables. Theorem 4.1 states that $\deg_\pm(\mathrm{MP}_m) = m$. Put $N = \lceil 16\mathrm{e}n^2/m\rceil = \lceil 256\mathrm{e}m^5\rceil$ and define the matrix $M = [\mathrm{MP}_m(x|_S)]_{x,S}$, where $x \in \{-1,1\}^N$ and $S \in \binom{[N]}{n}$. By Theorem 1.2,

$$\mathrm{disc}_\lambda(M) \leqslant 2^{-m} = \mathrm{e}^{-\Theta(N^{1/5})}$$

for a certain distribution $\lambda$. By Remark 3.4 and Theorem 4.3, the distribution $\lambda$ is given explicitly in terms of (4.3).

Represent a set $S \subset [N]$ with elements $i_1 < i_2 < \cdots < i_n$ by the Boolean string $(y^{(1)}, y^{(2)}, \dots, y^{(n)}) \in (\{-1,1\}^{\log N})^n$, where $y^{(k)}$ is the binary encoding of the integer $i_k$. We define $F : \{-1,1\}^N \times (\{-1,1\}^{\log N})^n \to \{-1,1\}$ by

$$F(x, y^{(1)}, y^{(2)}, \dots, y^{(n)}) = \mathrm{MP}_m(x|_S),$$

where $S$ is the set corresponding to $y^{(1)}, y^{(2)}, \dots, y^{(n)}$. In the event that the strings $y^{(1)}, y^{(2)}, \dots, y^{(n)}$ do not specify a legal set $S$ (e.g., they are not all distinct or ordered), the value of $F$ is irrelevant. By construction,

$$\mathrm{disc}_\lambda(F) = \mathrm{disc}_\lambda(M) \leqslant \mathrm{e}^{-\Theta(N^{1/5})}.$$

It remains to show that $F$ is computable by an $\mathsf{AC}^0$ circuit of depth 3. For this, observe that

$$F(x, y) = \mathrm{MP}_m(\phi(x, y^{(1)}), \dots, \phi(x, y^{(n)})),$$

where $\phi(x, y^{(i)})$ computes $x_{\mathrm{decimal}(y^{(i)})}$, i.e., computes $x_a$ with $a$ being the decimal integer whose binary representation is $y^{(i)}$. Each $\phi(x, y^{(i)})$ is clearly computable by a CNF formula of size $O(N)$. Hence, $F$ is computable by an $\mathsf{AC}^0$ circuit of depth 3 (by collapsing the two middle layers of AND gates). □

REMARK 5.1. The function $F$ in Theorem 1.3 can be viewed as a communication problem in which Alice is given an input $x \in \{-1,1\}^N$, Bob is given a polynomial-size DNF formula $f : \{-1,1\}^N \to \{-1,1\}$ (from a restricted set), and their objective is to evaluate $f(x)$. The proof of Theorem 1.3 shows that the communication matrix of this problem has discrepancy $\exp(-\Omega(N^{1/5}))$. We will revisit this observation in §8.

Theorem 1.3 exhibits an $\mathsf{AC}^0$ circuit of depth 3 with exponentially small discrepancy. At the same time, the discrepancy of every $\mathsf{AC}^0$ circuit of depth 2 is at least $n^{-O(1)}$. To our knowledge, this fact has not been noted down in the literature, and we present its proof below.

PROPOSITION 5.2. *Let $f : \{-1,1\}^n \times \{-1,1\}^n \to \{-1,1\}$ be an $\mathsf{AC}^0$ circuit of depth 1 or 2. Then $\mathrm{disc}_\mu(f) \geqslant n^{-O(1)}$ for every distribution $\mu$.*

*Proof.* By assumption, $f$ is a polynomial-size DNF or CNF formula. Without loss of generality, assume the former, i.e., $f = T_1 \vee T_2 \vee \cdots \vee T_s$, where $s = n^{O(1)}$ and each of $T_1, T_2, \dots, T_s$ is a conjunction of literals. Observe that

$$f = \text{MAJORITY}(T_1, \dots, T_s, T_{s+1}, \dots, T_{2s-1}),$$

where we define $T_{s+1} = T_{s+2} = \cdots = T_{2s-1} = -1$ (identically true). Consider the public-coin randomized protocol in which the parties pick $i \in \{1, 2, \dots, 2s-1\}$ uniformly at random, evaluate $T_i$ using constant communication, and output the result. This protocol evaluates $f$ correctly with probability at least $\frac{1}{2} + \Omega\left(\frac{1}{s}\right)$. Thus,

$$\mathrm{R}^{\mathrm{pub}}_{1/2-\Omega(1/s)}(f) = O(1).$$

Proposition 2.1 now implies that $\mathrm{disc}_\mu(f) \geqslant \Omega(1/s) \geqslant n^{-O(1)}$ for all $\mu$. □

**6. Discrepancy and the polynomial hierarchy.** In this section, we will briefly digress from the main development and explore the consequences of Theorem 1.3 in the study of communication complexity classes $\mathsf{PP}^{cc}$, $\Sigma_2^{cc}$, $\Pi_2^{cc}$.

Throughout this section, the symbol $\{f_n\}$ shall stand for a family of functions $f_1, f_2, \dots, f_n, \dots$, where $f_n : \{-1,1\}^n \times \{-1,1\}^n \to \{-1,1\}$.

Babai, Frankl, and Simon [3] originally defined the class $\mathsf{PP}^{cc}$ as the class of communication problems that have an efficient protocol with nonnegligible bias. For our purposes, it will be more convenient to use an equivalent characterization of $\mathsf{PP}^{cc}$ in terms of discrepancy, obtained by Klauck [16].

Theorem 6.1 (Klauck [16]). *A family $\{f_n\}$ is in $\mathsf{PP}^{cc}$ if and only if for some constant $c > 1$ and all $n$,*

$$\operatorname{disc}(f_n) > 2^{-\log^c n}.$$

We now define classes $\Sigma_2^{cc}$ and $\Pi_2^{cc}$, which represent the second level of the polynomial hierarchy in communication complexity. A function $f_n : \{-1,1\}^n \times \{-1,1\}^n \to \{-1,1\}$ is called a *rectangle* if there exist subsets $A, B \subseteq \{-1,1\}^n$ 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 6.2 (Babai, Frankl, and Simon [3, §4]).

(1) *A family $\{f_n\}$ is in $\Pi_0^{cc}$ if each $f_n$ is a rectangle. A family $\{f_n\}$ is in $\Sigma_0^{cc}$ if $\{\neg f_n\}$ is in $\Pi_0^{cc}$.*

(2) *Fix an integer $k = 1, 2, \dots$. A family $\{f_n\}$ is in $\Sigma_k^{cc}$ if for some constant $c > 1$ and all $n$,*

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

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

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

Thus, the zeroth level ($\Sigma_0^{cc}$ and $\Pi_0^{cc}$) 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 $\Pi_1^{cc} = \mathsf{coNP}^{cc}$.

The circuit class $\mathsf{AC}^0$ is related to the polynomial hierarchy $\mathsf{PH}^{cc}$ in communication complexity in the obvious way. Specifically, if $f_n : \{-1,1\}^n \times \{-1,1\}^n \to \{-1,1\}$, $n = 1, 2, 3, 4, \dots$, is an $\mathsf{AC}^0$ 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 1.3 is in $\Sigma_2^{cc}$, whereas $\{\neg f_n\}$ is in $\Pi_2^{cc}$. In this light, Theorems 1.3 and 6.1 have the following corollary:

Corollary 1.4 (Restated from p. 3). $\Sigma_2^{cc} \not\subseteq \mathsf{PP}^{cc}$, $\Pi_2^{cc} \not\subseteq \mathsf{PP}^{cc}$.

Observe that the separations in Corollary 1.4 are achieved for explicit functions, constructed in Theorem 1.3. Corollary 1.4 is tight in that $\mathsf{PP}^{cc}$ trivially contains $\Sigma_0^{cc}$, $\Sigma_1^{cc}$, $\Pi_0^{cc}$, $\Pi_1^{cc}$.

**7. Lower bounds for majority-of-threshold circuits.** At last, we are in a position to prove the main result of this paper. We will follow an established argument, due to Nisan [27], that relates discrepancy to the size of majority-of-threshold circuits. The key piece of the argument is the following statement.

Theorem 7.1 (Nisan [27]). *Let $f : \{-1,1\}^n \to \{-1,1\}$ be a linear threshold function. Then $\mathrm{R}^{\mathrm{pub}}_\epsilon(f) = O(\log n + \log \frac{1}{\epsilon})$, for any partition of the variables and any $\epsilon = \epsilon(n)$.*

We have:

Theorem 1.1 (Rephrased from p. 1). *There exists a function $f : \{-1,1\}^N \times \{-1,1\}^N \to \{-1,1\}$, explicitly given and computable by an $\mathsf{AC}^0$ circuit of depth 3, whose computation requires a majority vote of $\exp(\Omega(N^{1/5}))$ linear threshold gates.*

*Proof.* Nisan [27, Thm. 4] proved an analogous statement for the function inner product modulo 2, and we merely adapt his argument to our setting. Let $F$ be the function in the statement of Theorem 1.3, with $\mathrm{disc}(F) = \exp(-\Omega(N^{1/5}))$. Proposition 2.1 implies that for any $\gamma > 0$,

$$\mathrm{R}^{\mathrm{pub}}_{1/2-\gamma/2}(F) = \Omega(N^{1/5}) - \log \frac{1}{\gamma}. \tag{7.1}$$

On the other hand, suppose that $F = \text{majority}(h_1, h_2, \ldots, h_s)$, where each $h_i$ is a linear threshold function. Then the parties can randomly pick $i \in \{1, 2, \ldots, s\}$, evaluate $h_i$ correctly with probability $1 - 1/(4s)$ using Theorem 7.1, and output the result. This protocol would have communication cost $O(\log N + \log s)$ and would predict $F$ correctly with probability at least $(\frac{1}{2} + \frac{1}{2s}) - \frac{1}{4s} = \frac{1}{2} + \frac{1}{4s}$ on every input. Thus,

$$\mathrm{R}^{\mathrm{pub}}_{1/2-1/4s}(F) = O(\log N + \log s). \tag{7.2}$$

Comparing (7.1) and (7.2), we see that $s = \exp(\Omega(N^{1/5}))$. □

**8. An application to learning DNF formulas.** We conclude with an application of our results to computational learning theory. Let $\mathcal{C}$ be an arbitrary set of Boolean functions $\{-1,1\}^n \to \{-1,1\}$. Suppose it is possible to fix polynomial-time computable Boolean functions $h_1, \ldots, h_d : \{-1,1\}^n \to \{-1,1\}$ such that every function $f \in \mathcal{C}$ can be represented as

$$f(x) \equiv \mathrm{sign}\left(\sum_{i=1}^{d} a_i h_i(x)\right)$$

for some integers $a_1, \ldots, a_d$ with $|a_1| + \cdots + |a_d| \leqslant W$. The obvious complexity measures of this representation are $d$ and $W$. If $d$ and $W$ are polynomial in $n$, simple and efficient algorithms exist for learning $\mathcal{C}$ from random examples under every distribution, e.g., the classic Perceptron algorithm [26, 28]. Such classes $\mathcal{C}$ admit learning with *large margin* and therefore possess a variety of desirable characteristics [18].

Given $\mathcal{C}$, it is thus natural to ask whether it is possible to choose $h_1, \ldots, h_d$ such that $d = \mathsf{poly}(n)$ and $W = \mathsf{poly}(n)$. The question is particularly intriguing for polynomial-size DNF and CNF formulas, a concept class that has eluded every attempt at an efficient, distribution-free learning algorithm. Our machinery yields a strong negative answer to this question. We restrict our attention to DNF formulas, the CNF case being closely analogous.

Theorem 8.1. *Let $\mathcal{C}$ denote the concept class of polynomial-size DNF formulas. Let $h_1, \dots, h_d : \{-1,1\}^n \to \{-1,1\}$ be arbitrary Boolean functions such that every $f \in \mathcal{C}$ can be expressed as $f(x) \equiv \operatorname{sign}(\sum_{i=1}^d a_i h_i(x))$ for some integers $a_1, \dots, a_d$ with $|a_1| + \cdots + |a_d| \leqslant W$. Then*

$$dW \geqslant \mathrm{e}^{\Omega(n^{1/5})}.$$

*Proof.* Consider the communication problem $F$ in which Alice is given an input $x \in \{-1,1\}^n$, Bob is given a function $f \in \mathcal{C}$, and their objective is to compute $f(x)$. By Remark 5.1, the communication matrix of this problem has discrepancy

$$\operatorname{disc}(F) \leqslant \mathrm{e}^{-\Omega(n^{1/5})}.$$

We will construct a cost-2 public-coin randomized protocol for the problem, with advantage $1/(dW)$ on every input. Proposition 2.1 will then imply that

$$\frac{1}{dW} \leqslant 4 \operatorname{disc}(F),$$

and the proof will be complete.

The idea behind the protocol is not original; see [13, 14, 25, 31] for similar work. First, the parties pick an index $i \in \{1, \dots, d\}$ uniformly at random. Then Alice sends $h_i(x)$ to Bob. Bob retrieves the representation of $f$ as $f(x) \equiv \operatorname{sign}(\sum_{i=1}^d a_i h_i(x))$ for some integers $a_1, \dots, a_d$. With probability $\frac{1}{2} + \frac{1}{2} \cdot \frac{|a_i|}{|a_1|+\cdots+|a_d|}$, Bob announces $h_i(x) \cdot \operatorname{sign}(a_i)$ as the output. With the remaining probability, he announces $-h_i(x) \cdot \operatorname{sign}(a_i)$. Thus, Bob's expected output is $\frac{a_i h_i(x)}{|a_1|+\cdots+|a_d|}$. As a result, the protocol achieves the desired advantage:

$$f(x) \cdot \sum_{i=1}^{d} \frac{1}{d} \cdot \frac{a_i h_i(x)}{|a_1| + \cdots + |a_d|} = \frac{1}{d} \cdot \frac{|a_1 h_1(x) + \cdots + a_d h_d(x)|}{|a_1| + \cdots + |a_d|} \geqslant \frac{1}{dW}.$$

□

Remark 8.2. Using subtle techniques, Razborov and Sherstov [36] have recently proved the following substantially stronger result. Let $h_1, \dots, h_d : \{-1,1\}^n \to \mathbb{R}$ be arbitrary real functions such that every DNF formula $f$ of linear size is representable as $f(x) \equiv \operatorname{sign}(\sum_{i=1}^d a_i h_i(x))$ for some reals $a_1, \dots, a_d$. Then $d \geqslant \exp(\Omega(n^{1/3}))$. This lower bound on $d$ is essentially optimal [17] and rules out the possibility of PAC learning DNF formulas in the important *dimension complexity* framework; see [36] for details.

**Acknowledgments.** I would like to thank Anna Gál, Adam Klivans, and Sasha Razborov for helpful discussions and feedback on an earlier version of this manuscript. I am thankful to Adam Klivans for pointing me to the problem of separating $\mathsf{AC}^0$ from depth-2 majority circuits.

## REFERENCES

[1] Eric Allender, *A note on the power of threshold circuits*, in Proc. of the 30th Symposium on Foundations of Computer Science (FOCS), 1989, pp. 580–584.

[2] James Aspnes, Richard Beigel, Merrick L. Furst, and Steven Rudich, *The expressive power of voting polynomials*, Combinatorica, 14 (1994), pp. 135–148.

[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] László Babai, Noam Nisan, and Mario Szegedy, *Multiparty protocols, pseudorandom generators for logspace, and time-space trade-offs*, J. Comput. Syst. Sci., 45 (1992), pp. 204–232.

[5] 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.

[6] Arkadev Chattopadhyay, *Discrepancy and the power of bottom fan-in in depth-three circuits*, in Proc. of the 48th Symposium on Foundations of Computer Science (FOCS), 2007, pp. 449–458.

[7] Arkadev Chattopadhyay and Anil Ada, *Multiparty communication complexity of disjointness*, in Electronic Colloquium on Computational Complexity (ECCC), January 2008. Report TR08-002.

[8] Fan R. K. Chung and Prasad Tetali, *Communication complexity and quasi randomness*, SIAM J. Discrete Math., 6 (1993), pp. 110–123.

[9] Matei David and Toniann Pitassi, *Separating NOF communication complexity classes* $\mathsf{RP}$ *and* $\mathsf{NP}$, in Electronic Colloquium on Computational Complexity (ECCC), February 2008. Report TR08-014.

[10] Matei David, Toniann Pitassi, and Emanuele Viola, *Improved separations between nondeterministic and randomized multiparty communication*, in Proc. of the 12th Intl. Workshop on Randomization and Computation (RANDOM), 2008, pp. 371–384.

[11] Jeff Ford and Anna Gál, *Hadamard tensors and lower bounds on multiparty communication complexity*, in Proc. of the 32nd International Colloquium on Automata, Languages and Programming (ICALP), 2005, pp. 1163–1175.

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

[13] 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.

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

[15] 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.

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

[17] 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.

[18] ———, *Learning intersections of halfspaces with a margin*, in Proc. of the 17th Conf. on Learning Theory (COLT), 2004, pp. 348–362.

[19] 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.

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

[21] 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.

[22] ———, *Computing Boolean functions by polynomials and threshold circuits*, Comput. Complex., 7 (1998), pp. 346–370.

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

[24] Troy Lee and Adi Shraibman, *Disjointness is hard in the multi-party number-on-the-forehead model*, in Proc. of the 23rd Conf. on Computational Complexity (CCC), 2008, pp. 81–91.

[25] Nathan Linial and Adi Shraibman, *Learning complexity vs. communication complexity*, in Proc. of the 23rd Conf. on Computational Complexity (CCC), 2008, pp. 53–63.

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

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

[28] 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.

[29] 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.

[30] ———, *Extremal properties of polynomial threshold functions*, J. Comput. Syst. Sci., 74 (2008), pp. 298–312.

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

[32] Ran Raz, *Fourier analysis for probabilistic communication complexity*, Comput. Complex., 5 (1995), pp. 205–221.

[33] ———, *The BNS-Chung criterion for multi-party communication complexity*, Comput. Complex., 9 (2000), pp. 113–122.

[34] Alexander A. Razborov, *On the distributional complexity of disjointness*, Theor. Comput. Sci., 106 (1992), pp. 385–390.

[35] ———, *Quantum communication complexity of symmetric predicates*, Izvestiya: Mathematics, 67 (2003), pp. 145–159.

[36] Alexander A. Razborov and Alexander A. Sherstov, *The sign-rank of* $\mathsf{AC}^0$, in Proc. of the 49th Symposium on Foundations of Computer Science (FOCS), 2008, pp. 57–66.

[37] Alexander Schrijver, *Theory of linear and integer programming*, John Wiley & Sons, Inc., New York, 1998.

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

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

[40] ———, *Halfspace matrices*, Comput. Complex., 17 (2008), pp. 149–178.

[41] ———, *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.

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

[43] Kai-Yeung Siu and Jehoshua Bruck, *On the power of threshold circuits with small weights*, SIAM J. Discrete Math., 4 (1991), pp. 423–435.

[44] Kai-Yeung Siu, Jehoshua Bruck, Thomas Kailath, and Thomas Hofmeister, *Depth efficient neural networks for division and related problems*, IEEE Transactions on Information Theory, 39 (1993), pp. 946–956.

[45] Kai-Yeung Siu and Vwani P. Roychowdhury, *On optimal depth threshold circuits for multiplication and related problems.*, SIAM J. Discrete Math., 7 (1994), pp. 284–292.