Separating AC^0 from Depth-2 Majority Circuits

Alexander A. Sherstov

The University of Texas at Austin
Department of Computer Sciences
Austin, TX 78712 USA
sherstov@cs.utexas.edu

Abstract

We give the first proof that AC^0 cannot be simulated by MAJ ∘ MAJ circuits of size $2^{\mathrm{polylog}(n)}$. Namely, we construct an explicit AC^0 function that requires MAJ ∘ MAJ circuits of size $2^{\Omega(n^{1/5})}$. This solves an open problem arising in the work of Allender [1], and matches Allender's classic result that AC^0 can be simulated by MAJ ∘ MAJ ∘ MAJ circuits of size $2^{\mathrm{polylog}(n)}$. The hard function we construct is an AC^0 circuit of depth 3. This construction is best possible since all AC^0 circuits of depth *less than* 3 can be trivially simulated by polynomial-size MAJ ∘ MAJ circuits. This paper also complements work by Krause and Pudlák [10] on the hardness of computing AC^0 by threshold circuits.

Our proof is based on communication complexity. To obtain the above result, we develop a novel technique for communication lower bounds, the *Degree/Discrepancy Theorem.* This technique is a separate contribution of our paper. It allows one to translate lower bounds on the *threshold degree* of a Boolean function into upper bounds on the *discrepancy* of a closely related function. Upper bounds on the discrepancy, in turn, immediately imply communication lower bounds as well as lower bounds against threshold circuits.

As part of our proof, we use the Degree/Discrepancy Theorem to construct an explicit AC^0 circuit of depth 3 that has discrepancy $2^{-\Omega(n^{1/5})}$, under an explicit distribution. This yields the first known AC^0 function with exponentially small discrepancy: all previously known functions with low discrepancy feature PARITY or MAJORITY as a subfunction, and thus cannot be computed in AC^0.

Finally, we discuss applications of our work to computational learning theory, showing that polynomial-size DNF and CNF formulas have margin complexity $2^{\Omega(n^{1/5})}$.

1 Introduction

A classic result of complexity theory, due to Allender [1], states that every function in AC^0 can be computed by a depth-3 majority circuit of quasipolynomial size. The question immediately arises whether this result can be improved. In particular, can AC^0 be efficiently simulated by *depth-2* majority circuits? This question has remained unanswered since the publication of Allender's work in 1989. We solve this problem, giving an explicit AC^0 circuit that requires $\mathrm{MAJ} \circ \mathrm{MAJ}$ circuits of exponential size. Our result still holds if the bottom gates are replaced by arbitrary linear threshold functions (THRESH):

Theorem 1.1 ($\mathsf{AC}^{0,3}$ **requires large** $\mathrm{MAJ} \circ \mathrm{THRESH}$ **circuits).** *There is an (explicitly given) function $F : \{-1,1\}^N \times \{-1,1\}^N \to \{-1,1\}$ in $\mathsf{AC}^{0,3}$ (i.e., computable by an AND/OR/NOT circuit of size $\mathsf{poly}(N)$ and depth 3) such that any $\mathrm{MAJ} \circ \mathrm{THRESH}$ circuit for F requires $2^{\Omega(N^{1/5})}$ THRESH gates.*

The best previous lower bound was $2^{\mathsf{polylog}(N)}$ and is trivially obtained by recalling that AC^0 can compute *inner product mod 2* on $\log^c N$ variables, for any constant c. See [15] for details.

Our result establishes that Allender's elegant simulation is optimal. Yet $\mathrm{MAJ} \circ \mathrm{MAJ}$ circuits are known to be quite powerful; for example, the addition of N numbers, each N bits long, is computable by a polynomial-size $\mathrm{MAJ} \circ \mathrm{MAJ}$ circuit [23]. Moreover, every polynomial-size CNF or DNF formula (i.e., every $\mathsf{AC}^{0,2}$ circuit) *can* be trivially simulated by a $\mathrm{MAJ} \circ \mathrm{MAJ}$ circuit of polynomial size. Theorem 1.1 shows that for AC^0 circuits of depth 3 already such simulations do not exist. In this light, Theorem 1.1 complements Krause and Pudlák's construction [10] of an $\mathsf{AC}^{0,3}$ function that requires exponential-size $\mathrm{THRESH} \circ \mathrm{MOD}_r$ circuits, for all r. Note that the two results are incomparable.

A different and more revealing view of this paper is in terms of communication complexity [12]. The communication complexity of Boolean functions has long been an active area of research, due to its inherent appeal as a complexity subject as well as its varied applications in theoretical computer science. Despite this sustained interest, few general methods are known for communication lower bounds.

This paper contributes a new lower-bound technique based on the notion of threshold degree. For a Boolean function f, its *threshold degree* is the minimum d such that f can be computed by a majority vote of some fanin-d gates. Equivalent terminology includes "strong degree" [2], "voting polynomial degree" [10], and "PTF degree" [17]. This concept has an established role in the circuit complexity literature [2, 7, 9–11, 14]. In many cases [14], it is straightforward to obtain good lower bounds on the threshold degree. It is therefore natural to wonder whether lower bounds on the threshold degree can be used to obtain lower bounds on communication complexity. We answer this question in the affirmative. Given a Boolean function $f : \{-1,1\}^n \to \{-1,1\}$ with threshold degree d, we explicitly construct a related function $f^{[N]} : \{-1,1\}^N \times \{-1,1\}^N \to \{-1,1\}$ that has discrepancy at most $1/2^d$. Here $N = O(n^2/d)$ and

$$f^{[N]}(x,y) = f(\phi_1(x,y), \ldots, \phi_n(x,y)),$$

where each ϕ_i can be computed by a trivial DNF/CNF formula of size $O(N)$. Roughly speaking, $f^{[N]}$ applies the original function f to various subsets of the inputs. We defer the details to Section 3. This construction, the *Degree/Discrepancy Theorem*, is a separate contribution of our paper:

Theorem 1.2 (Degree/Discrepancy Theorem). *Let $f : \{-1,1\}^n \to \{-1,1\}$ have threshold degree $d \geqslant 1$. Then for any $N \geqslant n$,*

$$\operatorname{disc}\left(f^{[N]}\right) \leqslant \sqrt{\sum_{k=d}^{n} \binom{n}{k} \left(\frac{2n}{N}\right)^k}.$$

In particular,

$$\operatorname{disc}\left(f^{[N]}\right) < 2^{-d}$$

when $N \geqslant 10\mathrm{e}n^2/d$.

In words, Theorem 1.2 gives a method for obtaining discrepancy upper bounds from lower bounds on the threshold degree. Discrepancy upper bounds, in turn, immediately yield lower bounds on communication in every model (randomized, nondeterministic, deterministic). Discrepancy is particularly powerful in that it yields communication lower bounds even for computing the function to any *nonnegligible* advantage, a critical aspect for lower bounds against threshold circuits. This contrasts with other communication-complexity methods (e.g., the corruption bound of Razborov [21]) that only apply for computing the function to a small constant error. An additional advantage of discrepancy is its relationship to other complexity measures, e.g., its equivalence to the margin complexity of sign matrices (Linial and Shraibman [13]).

By applying the technique of Theorem 1.2 to an explicit function $f \in \mathsf{AC}^{0,2}$ with high threshold degree, we obtain an explicit function $f^{[N]} \in \mathsf{AC}^{0,3}$ that requires exponential-size MAJ $\circ$ THRESH circuits. This establishes Theorem 1.1.

An intermediate product of our proof is the explicitly given function $f^{[N]} \in \mathsf{AC}^{0,3}$ that has exponentially small discrepancy: $\operatorname{disc}_\lambda(f^{[N]}) = 2^{-\Omega(N^{1/5})}$, where λ is an explicitly given distribution. We find this fact interesting in its own right since all previously known functions with exponentially small discrepancy (e.g., [5, 15]) contain contain PARITY or MAJORITY as a subfunction, and thus cannot be computed in AC^0.

Theorem 1.3 (Discrepancy of $\mathsf{AC}^{0,3}$). *There is an (explicitly given) function $F : \{-1,1\}^N \times \{-1,1\}^N \to \{-1,1\}$ in $\mathsf{AC}^{0,3}$ (i.e., computable by an AND/OR/NOT circuit of size $\mathsf{poly}(N)$ and depth 3) and an explicit distribution λ over $\{-1,1\}^N \times \{-1,1\}^N$, such that* $\operatorname{disc}_\lambda(F) = 2^{-\Omega(N^{1/5})}$.

Theorem 1.3 is best possible in that every function $F : \{-1,1\}^N \times \{-1,1\}^N \to \{-1,1\}$ computable in $\mathsf{AC}^{0,2}$ has large discrepancy: $\operatorname{disc}_\lambda(F) = 1/N^{O(1)}$ under all distributions λ. See Section 5 for details.

The remainder of this paper is organized as follows. Section 2 provides necessary 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 function $f \in \mathsf{AC}^{0,2}$ with high threshold degree. Section 5 applies the Degree/Discrepancy Theorem to f, yielding an explicit function $f^{[N]} \in \mathsf{AC}^{0,3}$ with exponentially small discrepancy. Section 6 uses this discrepancy result to obtain exponential lower bounds on the size of MAJ $\circ$ THRESH circuits that implement $f^{[N]}$. Section 7 concludes with an application to computational learning theory.

2 Preliminaries

We view Boolean functions as mappings $\{-1,1\}^n \to \{-1,1\}$. The notation $[n]$ stands for the set $\{1,2,\ldots,n\}$. The expression

$$\binom{\{1,\ldots,N\}}{n}$$

denotes the family of all size-n subsets of $\{1,2,\ldots,N\}$.

The notation $\mathbb{R}^{m\times n}$ 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 that the (i,j)th entry of A is given by the expression $F(i,j)$.

For a constant integer $d \geqslant 1$, we denote the class of polynomial-size unbounded-fanin AND/OR/NOT circuits of depth d by $\mathsf{AC}^{0,d}$. As usual, AC^0 denotes the union $\bigcup_d \mathsf{AC}^{0,d}$, as d ranges over all constants. We denote a *majority* gate by MAJ. The abbreviation THRESH refers to a *linear threshold* gate, i.e., a function of the form $f = \operatorname{sign}(\sum_{i=1}^n a_i x_i - \theta)$ for some reals $a_1,\ldots,a_n,\theta$. We follow the general convention in denoting depth-2 majority circuits by MAJ $\circ$ MAJ; majority-of-threshold circuits by MAJ $\circ$ THRESH; majority-of-parity circuits by MAJ $\circ$ PARITY, and so on.

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 sets, possibly of unequal cardinality. We identify a function f with its *communication matrix* $M = [f(x,y)]_{x,y} \in \{-1,1\}^{|X|\times|Y|}$. 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 $\frac{1}{2}+\frac{\gamma}{2}$ (or, equivalently, with *advantage* γ) for each 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 $\frac{1}{2}-\frac{\gamma}{2}$ (or, equivalently, *advantage* γ) with respect to the distribution μ 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 μ 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 powerful technique that lower-bounds the randomized and distributional complexity in terms of the discrepancy:

Proposition 2.1 (Kushilevitz and Nisan [12, pp. 36–38]). *For every Boolean function $f(x,y)$, every distribution μ, 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 following fact is useful in analyzing the discrepancy. It arises as a special case in the work of Ford and Gál [3, Theorem 3.1] on multiparty communication complexity. It is also implicit in an article by Raz [20, Lemma 5.1].

Lemma 2.2 (Ford and Gál [3], Raz [20]). *Let $M \in \{-1,1\}^{|X|\times|Y|}$, and let μ be a probability distribution over $X \times Y$. Then there is a choice of signs $\alpha_x, \beta_y \in \{-1,1\}$ for all $x \in X$, $y \in Y$ such that*

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

Proof (adapted from Raz [20]): Let $R = A \times B$ be the rectangle over which the discrepancy is achieved. Fix $\alpha_x = 1$ for all $x \in A$, and likewise $\beta_y = 1$ for all $y \in B$. Choose the remaining signs α_x, β_y independently and at random. Passing to expectations,

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

In particular, there exists a setting $\alpha_x, \beta_y \in \{-1,1\}$ for all x,y with the desired property. □

A definitive resource for further details is the book of Kushilevitz and Nisan [12].

2.2 Threshold Degree

Let $\chi_S \coloneqq \prod_{i\in S} x_i$. Since we view Boolean functions as mappings $\{-1,1\}^n \to \{-1,1\}$, the function χ_S is the parity of the bits in the set S. Every function $f : \{-1,1\}^n \to \{-1,1\}$ can be represented as

$$f(x) \equiv \operatorname{sign}\left(\sum_{S\in\mathscr{S}} a_S \chi_S(x)\right),$$

for a suitable choice of $\mathscr{S} \subset \mathscr{P}([n])$ and real coefficients a_S. For example, the majority function can be represented as $\mathrm{MAJ}(x) = \operatorname{sign}(x_1 + x_2 + \cdots + x_n)$, and the parity function can be written as $\mathrm{PARITY}(x) = \operatorname{sign}(x_1 x_2 \ldots x_n) = \operatorname{sign}(\chi_{[n]})$.

The *degree* of a particular representation, $\operatorname{sign}(\sum_{S\in\mathscr{S}} a_S\chi_S)$, is defined as $\max_{S\in\mathscr{S}} |S|$, the largest fanin of a parity gate χ_S. The *threshold degree* of a function f is the minimum degree over all the representations of f. We denote the threshold degree of f by $\deg(f)$. It is clear that functions that depend on k out of the n variables have threshold degree at most k. Another simple observation is that a function and its negation have the same threshold degree: $\deg(f) = \deg(-f)$ for all f. Threshold degree is also known in the literature as "strong degree" [2], "voting polynomial degree" [10], and "PTF degree" [17].

A relevant result is the following theorem from the theory of linear inequalities:

Theorem 2.3 (Gordan's Transposition Theorem [22, Sec. 7.8]). *Let $A \in \mathbb{R}^{m\times n}$. Then exactly one of the following holds:*

(1) $u^t A > 0$ for some vector u;

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

The notation $u^tA > 0$ and $v \geqslant 0$ above for vectors u^tA and v is to be understood entry-wise, as usual. A straightforward consequence of Gordan's Transposition Theorem is the following well-known result regarding threshold representations (see also [2, 17]).

Theorem 2.4 (Existence of a threshold representation; cf. [2, 17]). *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 holds:*

(1) f can be represented as $f(x) \equiv \operatorname{sign}(\sum_{i=1}^k a_i\phi_i(x))$ for some real coefficients $a_1, a_2, \ldots, a_k$;

(2) there is a distribution μ over $\{-1,1\}^n$ such that $\mathop{\mathbf{E}}_{x\sim\mu}[f(x)\phi_i(x)] = 0$ for each $i = 1, 2, \ldots, k$.

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 2.4.1. *Let $f : \{-1,1\}^n \to \{-1,1\}$ be arbitrary. Then exactly one of the following holds:*

(1) $\deg(f) \leqslant d$;

(2) there is a distribution μ over $\{-1,1\}^n$ such that $\mathop{\mathbf{E}}_{x\sim\mu}[f(x)\chi_S(x)] = 0$ for all χ_S with $|S| \leqslant d$.

3 The Degree/Discrepancy Theorem

This section establishes a technique that allows one to start with a function f with high threshold degree and obtain a closely related function $f^{[N]}$ with low discrepancy. We formalize $f^{[N]}$ in the following definition.

Definition 3.1. Given $f:\{-1,1\}^n \to \{-1,1\}$, integer $N \geqslant n$, define $f^{[N]}:\{-1,1\}^N \times \binom{\{1,\dots,N\}}{n} \to \{-1,1\}$ as

$$f^{[N]}(x,S) \leftrightharpoons f(x_{i_1},x_{i_2},\dots,x_{i_n}),$$

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

The relationship between the threshold degree of f and the discrepancy of $f^{[N]}$, the *Degree/Discrepancy Theorem*, is the main result of this section.

Theorem 1.2 (Restated from p. 1). *Let $f:\{-1,1\}^n \to \{-1,1\}$ have threshold degree $d \geqslant 1$. Then for any $N \geqslant n$,*

$$\operatorname{disc}\left(f^{[N]}\right) \leqslant \sqrt{\sum_{k=d}^{n} \binom{n}{k} \left(\frac{2n}{N}\right)^k}.$$

In particular,

$$\operatorname{disc}\left(f^{[N]}\right) < 2^{-d}$$

when $N \geqslant 10\mathrm{e}n^2/d$.

Proof of Theorem 1.2

To reduce notational clutter, we put $F(x,S) \leftrightharpoons f^{[N]}(x,S)$. Let μ be a probability distribution over $\{-1,1\}^n$ under which $\mathop{\mathbf{E}}_{y\sim\mu}[f(y)\phi(y)] = 0$ for any real-valued function ϕ of $d-1$ or fewer of the variables $y_1,\dots,y_n$. The existence of μ is assured by Corollary 2.4.1. We will denote the uniform distribution by $\mathscr{U}$; the corresponding domain will be clear from the context. Let $\mu_S(x) \leftrightharpoons \mu(x_{i_1},x_{i_2},\dots,x_{i_n})$, where $i_1 < i_2 < \cdots < i_n$ are the elements of S. We will analyze the discrepancy of F under the distribution

$$\lambda(x,S) \leftrightharpoons \frac{\mu_S(x)}{2^{N-n}\cdot\binom{N}{n}}.$$

By Lemma 2.2, there is a choice of values $\alpha_x, \beta_S \in \{-1,1\}$ for all x and S such that

$$\operatorname{disc}_\lambda(F) \leqslant \left|\sum_{x,S} \alpha_x \beta_S \lambda(x,S) F(x,S)\right|.$$

As a result,

$$\begin{aligned}
\operatorname{disc}_\lambda(F)^2 &\leqslant \left(\sum_{x,S} \alpha_x \beta_S \lambda(x,S) F(x,S)\right)^2 \\
&= 4^n \left(\mathop{\mathbf{E}}_{x\sim\mathscr{U}} \mathop{\mathbf{E}}_{S\sim\mathscr{U}} [\alpha_x \beta_S \mu_S(x) F(x,S)]\right)^2 && \text{by definition of } \lambda \\
&\leqslant 4^n \mathop{\mathbf{E}}_{x\sim\mathscr{U}} \left[\left(\mathop{\mathbf{E}}_{S\sim\mathscr{U}} [\beta_S \mu_S(x) F(x,S)]\right)^2\right] && \text{since } \mathbf{E}[Z]^2 \leqslant \mathbf{E}[Z^2] \text{ for all } Z \\
&\leqslant 4^n \mathop{\mathbf{E}}_{(S,T)\sim\mathscr{U}} \left|\mathop{\mathbf{E}}_{x\sim\mathscr{U}} [\mu_S(x)\mu_T(x) F(x,S) F(x,T)]\right|.
\end{aligned}$$

The last equation implies that

$$\operatorname{disc}_\lambda(F)^2 \;\leqslant\; 4^n \sum_{k=0}^{n} \Pr[|S\cap T| = k]\cdot \max_{S,T:|S\cap T|=k} \left| \mathop{\mathbf{E}}_{x\sim\mathscr{U}} [\mu_S(x)\mu_T(x)F(x,S)F(x,T)] \right|. \tag{3.1}$$

To analyze this expression, we prove two key claims.

Claim 3.1.1. *Let* $|S\cap T| \leqslant d-1$. *Then* $\mathop{\mathbf{E}}_{x\sim\mathscr{U}} [\mu_S(x)\mu_T(x)F(x,S)F(x,T)] = 0$.

Proof of Claim 3.1.1. For notational convenience, assume that $S = \{1,2,\ldots,n\}$. Then

$$\begin{aligned}
\mathop{\mathbf{E}}_{x\sim\mathscr{U}} [\mu_S(x)\mu_T(x)F(x,S)F(x,T)] &= \mathop{\mathbf{E}}_{x\sim\mathscr{U}} [\mu(x_1,\ldots,x_n)\mu_T(x)f(x_1,\ldots,x_n)F(x,T)] \\
&= \frac{1}{2^N} \sum_{x_1,\ldots,x_n} \mu(x_1,\ldots,x_n)f(x_1,\ldots,x_n) \sum_{x_{n+1},\ldots,x_N} \mu_T(x)F(x,T) \\
&= \frac{1}{2^N} \mathop{\mathbf{E}}_{(x_1,\ldots,x_n)\sim\mu} \left[f(x_1,\ldots,x_n)\cdot \underbrace{\left(\sum_{x_{n+1},\ldots,x_N} \mu_T(x)F(x,T) \right)}_{*} \right].
\end{aligned}$$

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

Claim 3.1.2. *Let* $|S\cap T| = k$. *Then* $\left| \mathop{\mathbf{E}}_{x\sim\mathscr{U}} [\mu_S(x)\mu_T(x)F(x,S)F(x,T)] \right| \leqslant 2^{k-2n}$.

Proof of Claim 3.1.2. For notational convenience, let

$$S = \{1,2,\ldots,n\} \quad \text{and} \quad T = \{1,2,\ldots,k\}\cup\{n+1,n+2,\ldots,n+(n-k)\}.$$

We have:

$$\begin{aligned}
\left| \mathop{\mathbf{E}}_{x\sim\mathscr{U}} [\mu_S(x)\mu_T(x)F(x,S)F(x,T)] \right| &\leqslant \mathop{\mathbf{E}}_{x\sim\mathscr{U}} [|\mu_S(x)\mu_T(x)F(x,S)F(x,T)|] \\
&= \mathop{\mathbf{E}}_{x\sim\mathscr{U}} [\mu_S(x)\mu_T(x)] \\
&= \mathop{\mathbf{E}}_{x_1,\ldots,x_{2n-k}} [\mu(x_1,\ldots,x_n)\mu(x_1,\ldots,x_k,x_{n+1},\ldots,x_{2n-k})] \\
&\leqslant \underbrace{\mathop{\mathbf{E}}_{x_1,\ldots,x_n} [\mu(x_1,\ldots,x_n)]}_{=2^{-n}} \cdot \max_{x_1,\ldots,x_k} \underbrace{\mathop{\mathbf{E}}_{x_{n+1},\ldots,x_{2n-k}} [\mu(x_1,\ldots,x_k,x_{n+1},\ldots,x_{2n-k})]}_{\leqslant 2^{-(n-k)}}.
\end{aligned}$$

The bounds 2^{-n} and $2^{-(n-k)}$ above simply use the fact that μ is a probability distribution. □

We now apply Claims 3.1.1 and 3.1.2 to simplify (3.1):

$$\begin{aligned}
\operatorname{disc}_\lambda(F)^2 &\leqslant 4^n \sum_{k=0}^{n} \Pr[|S\cap T| = k]\cdot \max_{S,T:|S\cap T|=k} \left| \mathop{\mathbf{E}}_{x\sim\mathscr{U}} [\mu_S(x)\mu_T(x)F(x,S)F(x,T)] \right| \\
&= 4^n \sum_{k=d}^{n} \Pr[|S\cap T| = k]\cdot \max_{S,T:|S\cap T|=k} \left| \mathop{\mathbf{E}}_{x\sim\mathscr{U}} [\mu_S(x)\mu_T(x)F(x,S)F(x,T)] \right| && \text{by Claim 3.1.1} \\
&\leqslant 4^n \sum_{k=d}^{n} \Pr[|S\cap T| = k]\cdot 2^{k-2n} && \text{by Claim 3.1.2.}
\end{aligned}$$

Since

$$\Pr[|S \cap T| = k] = \frac{\binom{n}{k}\binom{N-n}{n-k}}{\binom{N}{n}} \leqslant \binom{n}{k}\left(\frac{n}{N}\right)^k,$$

we obtain:

$$\mathrm{disc}_\lambda(F)^2 \leqslant \sum_{k=d}^{n} \binom{n}{k}\left(\frac{2n}{N}\right)^k.$$

Using the inequality $\binom{n}{k} < (\frac{\mathrm{e}n}{k})^k$, we see that

$$\mathrm{disc}_\lambda(F)^2 < \frac{1}{4^d}$$

as soon as $N \geqslant 10\mathrm{e}n^2/d$. This completes the proof of the theorem. □

Remark 3.2. The proof of Theorem 1.2 analyzes the discrepancy of $f^{[N]}$ under a certain distribution λ, defined in terms of the distribution μ under which f is uncorrelated with every function of at most $\deg(f)-1$ variables. Therefore, the distribution λ in the statement of Theorem 1.2 is explicitly specified whenever μ is explicitly specified.

Remark 3.3. Another helpful observation concerns the circuit complexity of $f^{[N]}$. As stated, $f^{[N]}$ is a function of (x,S) pairs, where $x \in \{-1,1\}^N$ and $S \subset [N]$. The point is that in implementing $f^{[N]}$ as a circuit, one is free to choose whatever representation of the set S makes for the most compact circuit. The discrepancy of $f^{[N]}$ is an intrinsic property of its communication matrix (Section 2.1) and is therefore not affected by the *syntactic* representation of the inputs to $f^{[N]}$. We will take advantage of this fact below in the proof of Theorem 1.3, where we construct a small $\mathsf{AC}^{0,3}$ circuit for $f^{[N]}$ by representing S as a string in $[N]^n$.

4 A Function in $\mathsf{AC}^{0,2}$ with High Threshold Degree

Consider the function $\mathrm{MP} : \{-1,1\}^{4m^3} \to \{-1,1\}$, defined as

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

Observe that $\deg(\mathrm{MP}) \leqslant m$ since by the distributivity law, MP is computable by the conjunction of OR gates with fanin m. Minsky and Papert [14], who originally defined this function, proved that this simple upper bound on deg(MP) is in fact tight.

Theorem 4.1 (Minsky-Papert [14]). *The function* MP *on* $4m^3$ *variables has* $\deg(\mathrm{MP}) = m$.

Minsky and Papert's proof, while short and elegant, does not yield an explicit distribution μ over $\{-1,1\}^{4m^3}$ such that $\mathbf{E}_\mu[\mathrm{MP}\cdot\chi_S] = 0$ for all χ_S with $|S| \leqslant \deg(\mathrm{MP})-1$. The existence of such a distribution is assured by Corollary 2.4.1. The purpose of this section is to construct the distribution μ explicitly. In Section 5, we will exploit this μ to obtain an explicit function f in $\mathsf{AC}^{0,3}$ and an *explicit* distribution λ such that $\mathrm{disc}_\lambda(f) = 2^{-\Omega(n^{1/5})}$. We do not need the explicitness of μ, however, for the circuit lower bounds (Theorem 1.1).

Theorem 4.2 (Explicit distribution for MP**).** *There is an explicit distribution over* $\{-1,1\}^{4m^3}$ *such that* $\mathbf{E}_\mu[\mathrm{MP}\cdot\chi_S] = 0$ *for all* χ_S *with* $|S| \leqslant m-1$.

We obtain Theorem 4.2 by extending a beautiful argument, due to O'Donnell and Servedio [18], that makes the crux of the Minsky-Papert construction explicit. See Appendix A for the proof.

5 Discrepancy of $\mathsf{AC}^{0,3}$

This section proves an exponentially small upper bound on the discrepancy of an explicit function in $\mathsf{AC}^{0,3}$. In the next section, we will apply this result to threshold circuits.

Theorem 1.3 (Restated from p. 2). *There is an (explicitly given) function $F : \{-1,1\}^N \times \{-1,1\}^N \to \{-1,1\}$ in $\mathsf{AC}^{0,3}$ (i.e., computable by an AND/OR/NOT circuit of size $\mathsf{poly}(N)$ and depth 3) and an explicit distribution λ over $\{-1,1\}^N \times \{-1,1\}^N$, such that $\operatorname{disc}_\lambda(F) = 2^{-\Omega(N^{1/5})}$.*

Proof. Consider the function MP on $n = 4m^3$ variables. Theorem 4.1 states that $\deg(\mathrm{MP}) = m$. Consider now the function $\mathrm{MP}^{[N]}$, where $N = 10\mathrm{e}(4m^3)^2/m = 160\mathrm{e}m^5$. By Theorem 1.2,

$$\operatorname{disc}_\lambda\left(\mathrm{MP}^{[N]}\right) < \frac{1}{2^m} \tag{5.1}$$

under a suitable distribution λ. Theorem 4.2 gives an explicit distribution μ such that $\mathbf{E}_\mu\left[\mathrm{MP}\cdot\chi_S\right] = 0$ for all χ_S with $|S| \leqslant m-1$. Therefore, the distribution λ in (5.1) is explicitly given by Remark 3.2.

Represent a set $S \subset [N]$ with elements $i_1 < i_2 < \cdots < i_n$ by a Boolean string $(y_1, y_2, \ldots, y_n) \in (\{-1,1\}^{\log N})^n$, where y_k is the binary representation of the integer i_k. We define $F : \{-1,1\}^N \times (\{-1,1\}^{\log N})^n \to \{-1,1\}$ as follows:

$$F(x, y_1, y_2, \ldots, y_n) \rightleftharpoons \mathrm{MP}^{[N]}(x, S),$$

where S is the set whose elements, when written in binary, are $y_1 < y_2 < \cdots < y_n$. In the event that $y_1, y_2, \ldots, y_n$ do not specify a legal set S (e.g., they are not all distinct or ordered), the value of F is immaterial. By construction,

$$\operatorname{disc}_\lambda(F) = \operatorname{disc}_\lambda\left(\mathrm{MP}^{[N]}\right). \tag{5.2}$$

Equations (5.1) and (5.2) show that $\operatorname{disc}_\lambda(F) < 2^{-m} = 2^{-\Omega(N^{1/5})}$. Furthermore, F is function on at most $2N$ variables. It remains to show that F is computable by a polynomial-size AND/OR/NOT circuit of depth 3. For this, observe that

$$F(x,y) = \mathrm{MP}(\phi_1(x,y_1), \ldots, \phi_n(x,y_n)),$$

where $\phi_i(x,y_i)$ computes $x_{\mathrm{decimal}(y_i)}$, i.e., computes x_a with a being the decimal integer whose binary representation is y_i. Each ϕ_i is clearly computable by a decision tree of size $O(N)$, and thus also by a CNF formula of size $O(N)$. Hence, f is computable in $\mathsf{AC}^{0,3}$ (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 the objective is to evaluate $f(x)$. The proof of Theorem 1.3 shows that the communication matrix of this problem has discrepancy $2^{-\Omega(N^{1/5})}$. We will use this view in a later section.

Theorem 1.3 shows that $\mathsf{AC}^{0,3}$ has functions with exponentially small discrepancy. At the same time, the discrepancy of every function in $\mathsf{AC}^{0,2}$ is at least $1/\mathsf{poly}(n)$. This interesting fact may have been previously discovered; for completeness, we document its proof below.

Proposition 5.2 (Discrepancy of $\mathsf{AC}^{0,2}$). *Let $f : \{-1,1\}^n \times \{-1,1\}^n \to \{-1,1\}$ be a function in $\mathsf{AC}^{0,2}$ (i.e., a polynomial-size DNF or CNF formula). Then $\operatorname{disc}_\mu(f) = 1/n^{O(1)}$ for every distribution μ.*

Proof. Assume f is a polynomial-size DNF formula; the CNF case is analogous. Then in particular f can be represented as $\mathrm{MAJ}(T_1, T_2, \ldots, T_s)$, where $s = n^{O(1)}$ and each T_i is a term. Consider a public-coin randomized protocol in which the parties randomly pick $i \in \{1, 2, \ldots, s\}$, evaluate T_i using $O(1)$ communication, and output the result. This protocol evaluates f correctly with probability at least $\frac{1}{2} + \frac{1}{s}$. Thus,

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

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

6 Lower Bounds for MAJ ∘ THRESH Circuits

As a consequence of Theorem 1.3, we obtain an explicit function in $\mathsf{AC}^{0,3}$ that requires MAJ ∘ THRESH circuits of exponential size. We apply an established argument, due to Nisan [15], to prove that a low-discrepancy function requires large MAJ ∘ THRESH circuits. The key piece of the argument is the following statement.

Theorem 6.1 (Nisan [15]). *Let* $f = \mathrm{sign}(\sum_{i=1}^n a_i x_i)$ *be a linear threshold function. Then* $\mathrm{R}^{\mathrm{pub}}_\varepsilon(f) = O(\log n + \log \frac{1}{\varepsilon})$, *for any partition of the variables and any* $\varepsilon = \varepsilon(n)$.

We are now in a position to finish our task.

Theorem 1.1 (Restated from p. 1). *There is an (explicitly given) function* $F : \{-1,1\}^N \times \{-1,1\}^N \to \{-1,1\}$ *in* $\mathsf{AC}^{0,3}$ *(i.e., computable by an AND/OR/NOT circuit of size* $\mathsf{poly}(N)$ *and depth* 3*) such that any* MAJ ∘ THRESH *circuit for F requires* $2^{\Omega(N^{1/5})}$ THRESH *gates.*

Proof. Let F be the function in the statement of Theorem 1.3, with $\mathrm{disc}(F) = 2^{-\Omega(N^{1/5})}$. Proposition 2.1 implies that for any $\gamma > 0$,

$$\mathrm{R}^{\mathrm{pub}}_{1/2-\gamma/2}(F) = \Omega(N^{1/5}) - \log \frac{1}{\gamma}. \tag{6.1}$$

On the other hand, suppose F is computed by $\mathrm{MAJ}(h_1, h_2, \ldots, h_s)$, where each h_i is a linear threshold function. Then the parties can randomly pick $i \in \{1, 2, \ldots, s\}$ and use Theorem 6.1 to evaluate h_i correctly with probability $1 - 1/(2s)$, with only $O(\log N + \log s)$ bits of communication. The proposed protocol would have advantage at least $1/(2s)$ in predicting F. Thus,

$$\mathrm{R}^{\mathrm{pub}}_{1/2-1/(2s)}(F) = O(\log N + \log s). \tag{6.2}$$

Comparing (6.1) and (6.2), we see that $s = 2^{\Omega(N^{1/5})}$. □

Remark 6.2. It has been shown [5, 6] that polynomial-size MAJ ∘ THRESH circuits are exactly the same complexity class as polynomial-size MAJ ∘ MAJ circuits. Therefore, we could replace references to MAJ ∘ THRESH circuits by MAJ ∘ MAJ, in Theorem 1.1 above and elsewhere in this paper. We prefer the keep the MAJ ∘ THRESH notation, however, as the more descriptive one in the context of circuit lower bounds.

7 Representing DNF formulas as a Threshold of Features

We conclude with an application of our results to learning theory. Let $\mathscr{C}$ be an arbitrary set of Boolean functions. 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 $f \in \mathscr{C}$ can be represented as

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

where $a_1,\ldots,a_d$ are integers specific to f, 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, elegant and efficient algorithms exist for learning $\mathscr{C}$ from random examples under every distribution, e.g., the classic Perceptron algorithm [14, 16]. Such classes $\mathscr{C}$ possess a variety of other desirable characteristics [8]. (The simplicity of this learning task is due to the fact that the target functions are halfspaces with a large margin, $\frac{1}{dW}$, in terms of the feature set $h_1,\ldots,h_d$.) Given $\mathscr{C}$, a natural question to ask is 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 possibility. We confine our attention below to DNF formulas; the CNF case is closely analogous.

Theorem 7.1. *Let $\mathscr{C}$ denote the concept class of polynomial-size DNF formulas. Let $h_1,\ldots,h_d\colon\{-1,1\}^n \to \{-1,1\}$ be arbitrary Boolean functions such that every $f \in \mathscr{C}$ can be expressed as $f(x) \equiv \mathrm{sign}(\sum_{i=1}^d a_i h_i(x))$ for some integers $a_1,\ldots,a_d$ with $|a_1|+\cdots+|a_d| \leqslant W$. Then $dW \geqslant 2^{\Omega(n^{1/5})}$.*

Proof. Consider the communication problem F in which Alice is given an input $x \in \{-1,1\}^n$, Bob is given a function $f \in \mathscr{C}$, and the objective is to compute $f(x)$. By Remark 5.1, the communication matrix of this problem has discrepancy $2^{-\Omega(n^{1/5})}$. We will construct a cost-2 randomized protocol for the problem, with advantage $1/(dW)$ on every input. Then we will have

$$\frac{1}{dW} \overset{\text{Prop. 2.1}}{\leqslant} 4\,\mathrm{disc}(F) \overset{\text{Rem. 5.1}}{\leqslant} \frac{1}{2^{\Omega(n^{1/5})}},$$

and the proof will be complete.

We now describe the protocol. The idea behind this construction is not new; see [4, 13, 19] for similar work. First, the parties pick $i \in \{1,\ldots,d\}$ uniformly at random. Then Alice sends $h_i(x)$ to Bob. Bob retrieves the representation of f as $f(x) \equiv \mathrm{sign}(\sum_{i=1}^d a_i h_i(x))$ for some integers $a_1,\ldots,a_d$. With probability $\frac{1}{2}+\frac{1}{2}\cdot\frac{|a_i|}{|a_1|+\cdots+|a_d|}$, Bob announces $h_i(x)\cdot\mathrm{sign}(a_i)$ as the output. With the remaining probability, he announces $-h_i(x)\cdot\mathrm{sign}(a_i)$. 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:

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

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 AC^0 from $\mathsf{MAJ}\circ\mathsf{MAJ}$ circuits of quasipolynomial size.

References

[1] E. Allender. A note on the power of threshold circuits. In *FOCS*, pages 580–584, 1989.

[2] J. Aspnes, R. Beigel, M. Furst, and S. Rudich. The expressive power of voting polynomials. In *STOC '91: Proceedings of the twenty-third annual ACM symposium on Theory of computing*, pages 402–409, New York, NY, USA, 1991. ACM Press.

[3] J. Ford and A. Gál. Hadamard tensors and lower bounds on multiparty communication complexity. In *ICALP*, pages 1163–1175, 2005.

[4] J. Forster, M. Krause, S. V. Lokam, R. Mubarakzjanov, N. Schmitt, and H.-U. Simon. Relations between communication complexity, linear arrangements, and computational complexity. In *FST TCS '01: Proceedings of the 21st Conference on Foundations of Software Technology and Theoretical Computer Science*, pages 171–182, London, UK, 2001. Springer-Verlag.

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

[6] M. Goldmann and M. Karpinski. Simulating threshold circuits by majority circuits. *SIAM J. Comput.*, 27(1):230–246, 1998.

[7] A. R. Klivans and R. Servedio. Learning DNF in time $2^{\tilde{O}(n^{1/3})}$. In *STOC '01: Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 258–265, New York, NY, USA, 2001. ACM Press.

[8] A. R. Klivans and R. A. Servedio. Learning intersections of halfspaces with a margin. In *COLT*, pages 348–362, 2004.

[9] A. R. Klivans and A. A. Sherstov. Improved lower bounds for learning intersections of halfspaces. In *Proceedings of the 19th Annual Conference on Learning Theory (COLT)*, Pittsburg, USA, June 2006.

[10] M. Krause and P. Pudlák. On the computational power of depth 2 circuits with threshold and modulo gates. In *STOC '94: Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pages 48–57, New York, NY, USA, 1994. ACM Press.

[11] M. Krause and P. Pudlák. Computing boolean functions by polynomials and threshold circuits. *Comput. Complex.*, 7(4):346–370, 1998.

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

[13] N. Linial and A. Shraibman. Lower bounds in communication complexity based on factorization norms. Manuscript at `http://www.cs.huji.ac.il/~nati/PAPERS/quant_cc.pdf`, June 2006.

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

[15] N. Nisan. The communication complexity of threshold gates. In *Proceedings of "Combinatorics, Paul Erdos is Eighty"*, pages 301–315, 1993.

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

[17] R. O'Donnell and R. A. Servedio. Extremal properties of polynomial threshold functions. In *IEEE Conference on Computational Complexity*, pages 3–12, 2003.

[18] R. O'Donnell and R. A. Servedio. New degree bounds for polynomial threshold functions. In *STOC '03: Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 325–334, New York, NY, USA, 2003. ACM Press.

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

[20] R. Raz. The BNS-Chung criterion for multi-party communication complexity. *Comput. Complex.*, 9(2):113–122, 2000.

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

[22] A. Schrijver. *Theory of linear and integer programming*. John Wiley & Sons, Inc., New York, NY, USA, 1998.

[23] K.-Y. Siu and V. P. Roychowdhury. On optimal depth threshold circuits for multiplication and related problems. *SIAM J. Discrete Math.*, 7(2):284–292, 1994.

A An Explicit Distribution for the MP Function

Recall the function $\mathrm{MP} : \{-1,1\}^{4m^3} \to \{-1,1\}$, defined as

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

The objective of this section is to obtain an explicit distribution μ over $\{-1,1\}^{4m^3}$ such that $\mathbf{E}_{x\sim\mu}[\mathrm{MP}\cdot\chi_S] = 0$ for all χ_S with $|S| \leqslant \deg(\mathrm{MP})-1$. The *existence* of such a distribution is assured by Corollary 2.4.1. Our analysis works by extending a beautiful argument, due to O'Donnell and Servedio [18], that makes the crux of the Minsky-Papert construction explicit.

Proposition A.1 (O'Donnell and Servedio [18]). *Let $T = \{0,1,\ldots,2m\}$. Let $\nu(t) = 2^{-2m}\binom{2m}{t}$, a probability distribution over T. Then $\mathbf{E}_\nu[(-1)^t p(t)] = 0$ for every polynomial $p(t)$ of degree at most $2m-1$.*

Proof (O'Donnell and Servedio [18]). It suffices to prove the claim for $p(t) = t^d$, where $d \leqslant 2m-1$. The latter follows from the combinatorial identity $\sum_{t=0}^{2m}\binom{2m}{t}(-1)^t t^d = 0$, for all $d = 0,1,\ldots,2m-1$. □

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

For $t \in \{0,1,\ldots,2m\}$, define

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

Thus, $X_0, X_1, \ldots, X_{2m}$ are disjoint sets of inputs. The same sets of inputs figure in the analysis of Minsky and Papert [14] and O'Donnell and Servedio [18]. It is easy to check that for $t = 0,1,\ldots,2m$,

$$x \in X_t \qquad \Longrightarrow \qquad \mathrm{MP}(x) = (-1)^t. \tag{A.2}$$

Let ν be the distribution over $\{0,1,\dots,2m\}$ as defined in Proposition A.1. We will work with the following distribution μ over $\{-1,1\}^{4m^3}$:

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

Theorem 4.2 (Restated from p. 7). *There is an explicit distribution (namely, μ above) over $\{-1,1\}^{4m^3}$ such that $\mathbf{E}_\mu[\mathrm{MP}\cdot\chi_S] = 0$ for all χ_S with $|S| \leqslant m-1$.*

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

Claim A.1.1. $\mathbf{E}_\mu[\mathrm{MP}\cdot(\chi_S \circ (\sigma_1,\dots,\sigma_m))] = \mathbf{E}_\mu[\mathrm{MP}\cdot\chi_S]$ *for all* $\sigma_1,\dots,\sigma_m$.

Claim A.1.2. $\sum_{\sigma_1,\dots,\sigma_m} \mathbf{E}_\mu[\mathrm{MP}\cdot(\chi_S \circ (\sigma_1,\dots,\sigma_m))] = 0.$

We prove these claims below. This completes the proof of the theorem. □

Proof of Claim A.1.1. The functions $\mathrm{MP}(x)$ and $\mu(x)$ depend only on the sum of the bits in each block. Formally, $\mathrm{MP} \equiv \mathrm{MP} \circ (\sigma_1,\dots,\sigma_m)$ and $\mu \equiv \mu \circ (\sigma_1,\dots,\sigma_m)$. The claim follows. □

Proof of Claim A.1.2. Write $\chi_S = \chi_{S_1}\chi_{S_2}\dots\chi_{S_m}$, where $S_i = S \cap \{(i,1),\dots,(i,4m^2)\}$. Then,

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

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

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

where the last line follows by Proposition A.1 since $p(t)$ has degree at most $2\sum_i |S_i| = 2|S| \leqslant 2m-2$. □