OPTIMAL BOUNDS FOR SIGN-REPRESENTING THE INTERSECTION OF TWO HALFSPACES BY POLYNOMIALS*

ALEXANDER A. SHERSTOV[†]

ABSTRACT. The *threshold degree* of a function $f: \{0, 1\}^n \to \{-1, +1\}$ is the least degree of a real polynomial p with $f(x) \equiv \operatorname{sgn} p(x)$. We prove that the intersection of two halfspaces on $\{0, 1\}^n$ has threshold degree $\Omega(n)$, which matches the trivial upper bound and solves an open problem due to Klivans (2002). The best previous lower bound was $\Omega(\sqrt{n})$. Our result shows that the intersection of two halfspaces on $\{0, 1\}^n$ only admits a trivial $2^{\Theta(n)}$-time learning algorithm based on sign-representation by polynomials, unlike the advances achieved in PAC learning DNF formulas and read-once Boolean formulas.

1. INTRODUCTION

A well-studied notion in computational learning theory is that of a *perceptron*. This term stands for the representation of a given Boolean function $f: \{0, 1\}^n \to \{-1, +1\}$ in the form $f(x) \equiv \operatorname{sgn} p(x)$ for a real polynomial p of some degree d. The least degree d for which f admits such a representation is called the *threshold degree* of f, denoted $\deg_{\pm}(f)$. In other words, $\deg_{\pm}(f)$ is the least degree of a real polynomial that agrees with f in sign. Perceptrons are appealing from a learning standpoint because they immediately lead to efficient learning algorithms. In more detail, let $f: \{0, 1\}^n \to \{-1, +1\}$ be an unknown function of threshold degree d. Then by definition, f has a representation of the form

$$f(x) \equiv \operatorname{sgn} \left(\sum_{|S| \leqslant d} \lambda_S \prod_{i \in S} x_i \right)$$

for some reals λ_S and is thus a halfspace in $N = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d}$ dimensions. As a result, f can be PAC learned in time polynomial in N, using any of a variety of halfspace learning algorithms. (Throughout this paper, the term "PAC learning" refers to Valiant's standard model [40] of learning under arbitrary distributions.)

The study of perceptrons dates back forty years to the seminal monograph of Minsky and Papert [25], who examined the threshold degree of several common functions. Since then, the perceptron-based approach has yielded the fastest known PAC learning algorithms for a variety of concept classes. One such is the class of DNF formulas of polynomial size, posed a challenge in Valiant's original paper [40] and extensively studied over the past two decades. The fastest known algorithm for PAC learning DNF formulas runs in time $\exp\{\tilde{O}(n^{1/3})\}$ and is due to Klivans and Servedio [18]. Specifically, the authors of [18] prove an upper bound of $O(n^{1/3} \log n)$ on the threshold degree of polynomial-size DNF

formulas, which essentially matches a classical lower bound of $\Omega(n^{1/3})$ due to Minsky and Papert [25].

Another success story of the perceptron-based approach is the concept class of Boolean formulas, i.e., Boolean circuits with fan-out 1 at every gate. O'Donnell and Servedio [29] proved an upper bound of $\sqrt{s} \log^{O(d)} s$ on the threshold degree of Boolean formulas of size s and depth d, giving the first subexponential algorithm for a family of formulas of superconstant depth. This upper bound on the threshold degree was improved to $s^{0.5+o(1)}$ for any depth d by Ambainis et al. [2], building on a quantum query algorithm of Farhi et al. [10]. More recently, Lee [24] sharpened the upper bound to a tight $O(\sqrt{s})$. This line of research gives the fastest known algorithm for PAC learning Boolean formulas.

Another extensively studied problem in computational learning theory, and the subject of this paper, is the problem of learning *intersections of halfspaces*, i.e., learning conjunctions of functions of the form $f(x) = \text{sgn}(\sum \alpha_i x_i - \theta)$ for some reals $\alpha_1, \ldots, \alpha_n, \theta$. While solutions are known to several restrictions of this problem [7, 23, 41, 3, 17, 19, 16], no algorithm has been discovered for PAC learning the intersection of even two halfspaces in time faster than $2^{\Theta(n)}$. Progress on proving hardness results has also been scarce. Indeed, all previous hardness results [8, 1, 20, 14] either require polynomially many halfspaces or assume *proper* learning. In particular, we are not aware of any representation-independent hardness results for PAC learning the intersection of $O(1)$ halfspaces.

Our result. Since the perceptron-based approach yields the fastest known algorithms for PAC learning DNF formulas and read-once Boolean formulas, it is natural to wonder whether it can yield any nontrivial results for the intersection of two halfspaces. Letting $D(n)$ stand for the maximum threshold degree over all intersections of two halfspaces on $\{0, 1\}^n$, the question becomes whether $D(n)$ is a nontrivial (sublinear) function of the dimension n. This question has been studied by several authors, as summarized in Table 1. Forty years ago, Minsky and Papert [25] used a compactness argument to show that $D(n) = \omega(1)$, the function in question being the intersection of two majorities on disjoint sets of variables. O'Donnell and Servedio [29] studied the same function using a rather different approach and thereby proved that $D(n) = \Omega(\log n / \log \log n)$. No nontrivial upper bounds on $D(n)$ being known, Klivans [15, §7] formally posed the problem of proving a lower bound substantially better than $\Omega(\log n)$ or an upper bound of $o(n)$.

It was recently shown in [34] that $D(n) = \Omega(\sqrt{n})$, solving Klivans' problem and ruling out an $n^{o(\sqrt{n})}$-time PAC learning algorithm based on perceptrons. It is clear, however, that a PAC learning algorithm for the intersection of two halfspaces in time $n^{\Theta(\sqrt{n})}$ would still be a breakthrough in computational learning theory, comparable to the advances in the study of DNF formulas and read-once Boolean formulas. The main contribution of this paper is to prove that $D(n) = \Omega(n)$, which matches the trivial upper bound and definitively

Result	Reference
$D(n) = \omega(1)$	[25]
$D(n) = \Omega(\log n / \log \log n)$	[29]
$D(n) = \Omega(\sqrt{n})$	[34]
$D(n) = \Theta(n)$	this paper

TABLE 1. Lower bounds for the intersection of two halfspaces.

rules out the perceptron-based approach for learning the intersection of two halfspaces in nontrivial time.

THEOREM 1 (Main result). *For $n = 1, 2, 3, \ldots$, let $D(n)$ denote the maximum threshold degree of a function of the form $f(x) \wedge g(x)$, where $f, g: \{0, 1\}^n \rightarrow \{-1, +1\}$ are halfspaces. Then*

$$D(n) = \Theta(n).$$

To be more precise, we give a randomized algorithm which with probability at least $1 - e^{-n/12}$ constructs two halfspaces on $\{0, 1\}^n$ whose intersection has threshold degree $\Theta(n)$. In Section 6, we develop several refinements of Theorem 1. For example, we show that the intersection of two halfspaces on $\{0, 1\}^n$ requires a perceptron with $\exp\{\Theta(n)\}$ monomials, i.e., does not have a sparse sign-representation. We also give an essentially tight lower bound on the threshold degree of the intersection of a halfspace and a majority function, improving quadratically on the previous bound in [34].

In summary, unlike DNF formulas and read-once Boolean formulas, the intersection of two halfspaces does not admit a nontrivial sign-representation. Apart from computational learning theory, lower bounds on the threshold degree have played a key role in several works on circuit complexity [30, 39, 21, 22, 35], Turing complexity classes [4, 6, 5], and communication complexity [35, 36, 37, 31]. For this reason, we consider Theorem 1 and the techniques used to obtain it to be of interest outside of computational learning.

Theorem 1 and much previous work suggest that the nature of a PAC learning problem changes significantly when, instead of Valiant's original arbitrary-distribution setting, one considers learning with respect to restricted distributions. For example, the uniform distribution on the sphere $\mathbb{S}^{n-1}$ or hypercube $\{0, 1\}^n$ allows the use of tools other than sign-representing polynomials, such as Fourier analysis. In particular, polynomial-time algorithms are known for the uniform-distribution learning of intersections of a constant number of halfspaces on the sphere [7, 41] and hypercube [17]. Furthermore, if membership queries are allowed, DNF formulas are known to be learnable in polynomial time with respect to the uniform distribution on the hypercube [12].

Our techniques. Let $f \wedge f$ denote the conjunction of two copies of a given Boolean function f, each on an independent set of variables. It was shown in [34] that the threshold degree of $f \wedge f$ equals, up to a small multiplicative constant, the least degree of a rational function R with $\|f - R\|_\infty \leqslant 1/3$. With this characterization in hand, the equality $\deg_\pm(f \wedge f) = \Theta(\sqrt{n})$ was derived in [34] by solving the rational approximation problem for the halfspace

$$f(x) = \operatorname{sgn}\left(1 + \sum_{i=1}^{\sqrt{n}} \sum_{j=1}^{\sqrt{n}} 2^i x_{ij}\right).$$

Unfortunately, the $\Theta(\sqrt{n})$ barrier is fundamental to the analysis in [34]. To prove that in fact $D(n) = \Theta(n)$, we pursue a rather different approach.

The intuition behind our work is as follows. Let $\alpha_1, \alpha_2, \ldots, \alpha_n$ be given nonzero integers, and let $f: \{0, 1\}^n \rightarrow \{-1, +1\}$ be a given Boolean function such that $f(x)$ is completely determined by the sum $\sum \alpha_i x_i$. When approximating f pointwise by polynomials and rational functions of a given degree, can one restrict attention to those approximants that are, like f, functions of the sum $\sum \alpha_i x_i$ alone rather than the individual bits

$x_1, x_2, \ldots, x_n$? If true, this claim would dramatically simplify the analysis of the threshold degree of f by reducing it to a univariate question. Minsky and Papert [25] showed that the claim is indeed true in the highly special case $\alpha_1 = \alpha_2 = \cdots = \alpha_n$. For the purposes of this paper, however, the nonzero coefficients $\alpha_1, \alpha_2, \ldots, \alpha_n$ must be of increasing orders of magnitude and in particular must satisfy

$$\max_{i,j} \left| \frac{\alpha_i}{\alpha_j} \right| > \exp\{\Omega(n)\}.$$

Minsky and Papert's argument breaks down completely in this setting, and with good reason: coefficients $\alpha_1, \alpha_2, \ldots, \alpha_n$ are easily constructed [5] for which the passage to univariate approximation increases the degree requirement from 1 to n.

To overcome this difficulty, we use techniques from Fourier analysis and matrix perturbation theory. Specifically, we define an appropriate distribution on n-tuples $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ and study the behavior of the sum $\sum \alpha_i x_i$ as the vector x ranges over $\{0, 1\}^n$. We prove that for a typical n-tuple $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ and any collection of sums $S \subset \mathbb{Z}$ of interest, the subset $X_S \subset \{0, 1\}^n$ that induces the sums in S is highly random in that membership in X_S is uncorrelated with any polynomial of degree up to $\Theta(n)$. With some additional work, this allows the sought passage to a univariate question. In particular, we are able to prove the existence of a halfspace $f \colon \{0, 1\}^n \to \{-1, +1\}$ such that any multivariate rational approximant for f gives a univariate rational approximant for the sign function on $\{\pm 1, \pm 2, \pm 3, \ldots, \pm 2^{\Theta(n)}\}$ with the same degree and error. The univariate question being well-understood, we infer that f requires a rational function of degree $\Omega(n)$ for pointwise approximation within $1/3$ and hence $\deg_\pm(f \wedge f) \geqslant \Omega(n)$ by the characterization from [34].

2. Preliminaries

Notation. We will view Boolean functions as mappings $X \to \{0, 1\}$ or $X \to \{-1, +1\}$ for some finite set X, where the output value 1 corresponds to "true" in the former case and "false" in the latter. We adopt the following standard definition of the sign function:

$$\operatorname{sgn} x = \begin{cases} -1, & x < 0, \\ 0, & x = 0, \\ 1, & x > 0. \end{cases}$$

The complement of a set S is denoted $\overline{S}$. We denote the symmetric difference of sets S and T by $S \oplus T = (S \cap \overline{T}) \cup (\overline{S} \cap T)$. For a finite set X, the symbol $\mathscr{P}(X)$ denotes the family of all $2^{|X|}$ subsets of X. For functions $f, g \colon X \to \mathbb{R}$ on a finite set X, we use the notation

$$\langle f, g \rangle = \frac{1}{|X|} \sum_{x \in X} f(x) g(x).$$

We let $\log x$ stand for the logarithm of x to the base 2. The binary entropy function $H \colon [0, 1] \to [0, 1]$ is given by $H(p) = -p \log p - (1 - p) \log(1 - p)$ and is strictly increasing on $[0, 1/2]$. The following bound is well known [13, p. 283]:

$$(2.1) \qquad \sum_{i=0}^{k} \binom{n}{i} \leqslant 2^{H(k/n)n}, \qquad\qquad k = 0, 1, 2, \ldots, \lfloor n/2 \rfloor.$$

For elements x, y of a given set, we use the Kronecker delta

$$\delta_{x,y} = \begin{cases} 1 & \text{if } x = y, \\ 0 & \text{otherwise.} \end{cases}$$

The symbol P_k stands for the family of all univariate real polynomials of degree up to k. The majority function $\mathrm{MAJ}_n \colon \{0, 1\}^n \to \{-1, +1\}$ has the usual definition:

$$\mathrm{MAJ}_n(x) = \begin{cases} -1 & \text{if } x_1 + x_2 + \cdots + x_n > n/2, \\ 1 & \text{otherwise.} \end{cases}$$

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

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

For $S \subseteq \{1, 2, \ldots, n\}$, define $\chi_S \colon \{0, 1\}^n \to \{-1, +1\}$ by $\chi_S(x) = (-1)^{\sum_{i \in S} x_i}$. Then $\{\chi_S\}_{S \subseteq \{1,2,\ldots,n\}}$ is an orthonormal basis for the inner product space in question. As a result, every function $f \colon \{0, 1\}^n \to \mathbb{R}$ has a unique representation of the form

$$f = \sum_{S \subseteq \{1,2,\ldots,n\}} \hat{f}(S) \chi_S,$$

where $\hat{f}(S) = \langle f, \chi_S \rangle$. The reals $\hat{f}(S)$ are called the *Fourier coefficients of f*. The orthonormality of $\{\chi_S\}$ immediately yields *Parseval's identity*:

$$(2.2) \qquad \sum_{S \subseteq \{1,2,\ldots,n\}} \hat{f}(S)^2 = \langle f, f \rangle = \underset{x \in \{0,1\}^n}{\mathbf{E}} [f(x)^2].$$

Matrices. The symbol $\mathbb{R}^{m \times n}$ refers to the family of all $m \times n$ matrices with real entries. A matrix $A \in \mathbb{R}^{n \times n}$ is called *strictly diagonally dominant* if

$$|A_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |A_{ij}|, \qquad\qquad i = 1, 2, \ldots, n.$$

A well-known result in matrix perturbation theory, due to Gershgorin [11], states that the eigenvalues of a matrix lie in the union of certain disks in the complex plane centered around the diagonal entries of the matrix. We will need the following very special case, which corresponds to showing that the eigenvalues are all nonzero.

THEOREM 2.1 (Gershgorin). *Let $A \in \mathbb{R}^{n \times n}$ be strictly diagonally dominant. Then A is nonsingular.*

Proof (Gershgorin). Fix a nonzero vector $x \in \mathbb{R}^n$ and choose i such that $|x_i| = \|x\|_\infty$. Then by strict diagonal dominance,

$$|(Ax)_i| = \left| \sum_{j=1}^{n} A_{ij} x_j \right| \geq |A_{ii}| \|x\|_\infty - \sum_{\substack{j=1 \\ j \neq i}}^{n} |A_{ij}| \|x\|_\infty > 0,$$

so that $Ax \neq 0$. □

Rational approximation. The degree of a rational function $p(x)/q(x)$, where p and q are polynomials on $\mathbb{R}^n$, is the maximum of the degrees of p and q. Consider a function $f\colon X \to \{-1, +1\}$, where $X \subseteq \mathbb{R}^n$. For $d \geqslant 0$, define

$$R(f, d) \;=\; \inf_{p,q} \sup_{x \in X} \left| f(x) - \frac{p(x)}{q(x)} \right|,$$

where the infimum is over multivariate polynomials p and q of degree up to d such that q does not vanish on X. In words, $R(f, d)$ is the least error in an approximation of f by a multivariate rational function of degree up to d. A closely related quantity is

$$R^{+}(f, d) \;=\; \inf_{p,q} \sup_{x \in X} \left| f(x) - \frac{p(x)}{q(x)} \right|,$$

where the infimum is over multivariate polynomials p and q of degree up to d such that q is positive on X. These two quantities are related in a straightforward way:

$$R^{+}(f, 2d) \;\leqslant\; R(f, d) \;\leqslant\; R^{+}(f, d).$$

The second inequality here is trivial. The first follows from the fact that every rational approximant $p(x)/q(x)$ of degree d gives rise to a degree-$2d$ rational approximant with the same error and a positive denominator, namely, $\{p(x)q(x)\}/q(x)^2$.

The infimum in the definitions of $R(f, d)$ and $R^{+}(f, d)$ cannot in general be replaced by a minimum [32], even when X is finite subset of $\mathbb{R}$. This contrasts with the more familiar setting of a finite-dimensional normed linear space, where least-error approximants are guaranteed to exist.

For $S \subseteq \mathbb{R}$, we let

$$R^{+}(S, d) = \inf_{p,q} \sup_{x \in S} \left| \operatorname{sgn} x - \frac{p(x)}{q(x)} \right|,$$

where the infimum ranges over $p, q \in P_d$ such that q is positive on S. The study of the rational approximation of the sign function dates back to seminal work by Zolotarev [42] in the late nineteenth century. A much later result due to Newman [28] gives accurate estimates of $R^{+}([-n, -1] \cup [1, n], d)$ for all n and d. Newman's work in particular provides upper bounds on $R^{+}(\{\pm 1, \pm 2, \dots, \pm n\}, d)$, which in [34] were sharpened and complemented with matching lower bounds to the following effect:

THEOREM 2.2 (Sherstov). *Let n, d be positive integers, $R = R^{+}(\{\pm 1, \pm 2, \dots, \pm n\}, d)$. For $1 \leqslant d \leqslant \log n$,*

$$\exp\left\{ -\Theta\!\left(\frac{1}{n^{1/(2d)}} \right) \right\} \;\leqslant\; R \;<\; \exp\left\{ -\frac{1}{n^{1/d}} \right\}.$$

For $\log n < d < n$,

$$R = \exp\left\{ -\Theta\!\left(\frac{d}{\log(2n/d)} \right) \right\}.$$

For $d \geqslant n$,

$$R = 0.$$

Theorem 2.2 has the following corollary [34, Thm. 1.7], in which we adopt the notation $\operatorname{rdeg}_{\epsilon}(f) = \min\{d : R^{+}(f, d) \leqslant \epsilon\}$.

THEOREM 2.3 (Sherstov). *Let* $\mathrm{MAJ}_n\colon \{0,1\}^n \to \{-1,+1\}$ *denote the majority function. Then*

$$\mathrm{rdeg}_\epsilon(\mathrm{MAJ}_n) = \begin{cases} \Theta\left(\log\left\{\dfrac{2n}{\log(1/\epsilon)}\right\} \cdot \log\dfrac{1}{\epsilon}\right), & 2^{-n} < \epsilon < 1/3, \\[3mm] \Theta\left(1 + \dfrac{\log n}{\log\{1/(1-\epsilon)\}}\right), & 1/3 \leqslant \epsilon < 1. \end{cases}$$

Threshold degree. Let $f\colon X \to \{-1,+1\}$ be a given Boolean function, where $X \subset \mathbb{R}^n$ is finite. The *threshold degree* of f, denoted $\deg_\pm(f)$, is the least degree of a polynomial $p(x)$ such that $f(x) \equiv \mathrm{sgn}\, p(x)$. The term "threshold degree" appears to be due to Saks [33]. Equivalent terms in the literature include "strong degree" [4], "voting polynomial degree" [21], "polynomial threshold function degree" [29], and "sign degree" [9].

Given functions $f\colon X \to \{-1,+1\}$ and $g\colon Y \to \{-1,+1\}$, we let the symbol $f \wedge g$ stand for the function $X \times Y \to \{-1,+1\}$ given by $(f \wedge g)(x,y) = f(x) \wedge g(y)$. Note that in this notation, f and $f \wedge f$ are completely different functions, the former having domain X and the latter $X \times X$. The following ingenious observation, due to Beigel et al. [6], relates the notions of sign-representation and rational approximation for conjunctions of Boolean functions.

THEOREM 2.4 (Beigel, Reingold, and Spielman). *Let* $f\colon X \to \{-1,+1\}$ *and* $g\colon Y \to \{-1,+1\}$ *be given functions, where* $X, Y \subseteq \mathbb{R}^n$. *Let* d *be an integer with* $R^+(f,d) + R^+(g,d) < 1$. *Then*

$$\deg_\pm(f \wedge g) \leqslant 2d.$$

Proof (Beigel, Reingold, and Spielman). Consider rational functions $p_1(x)/q_1(x)$ and $p_2(y)/q_2(y)$ of degree at most d such that q_1 and q_2 are positive on X and Y, respectively, and

$$\sup_X \left| f(x) - \frac{p_1(x)}{q_1(x)} \right| + \sup_Y \left| g(y) - \frac{p_2(y)}{q_2(y)} \right| < 1.$$

Then

$$f(x) \wedge g(y) \equiv \mathrm{sgn}\{1 + f(x) + g(y)\} \equiv \mathrm{sgn}\left\{1 + \frac{p_1(x)}{q_1(x)} + \frac{p_2(y)}{q_2(y)}\right\}.$$

Multiplying the last expression by the positive quantity $q_1(x)q_2(y)$ gives $f(x) \wedge g(y) \equiv \mathrm{sgn}\{q_1(x)q_2(y) + p_1(x)q_2(y) + p_2(y)q_1(x)\}$. □

We will also need a converse to Theorem 2.4, proved in [34, Thm. 3.9].

THEOREM 2.5 (Sherstov). *Let* $f\colon X \to \{-1,+1\}$ *and* $g\colon Y \to \{-1,+1\}$ *be given functions, where* $X, Y \subset \mathbb{R}^n$ *are arbitrary finite sets. Assume that* f, g *are not identically false. Let* $d = \deg_\pm(f \wedge g)$. *Then*

$$R^+(f,4d) + R^+(g,2d) < 1.$$

Symmetric functions. Let S_n denote the symmetric group on n elements. For $\sigma \in S_n$ and $x \in \{0,1\}^n$, we denote $\sigma x = (x_{\sigma(1)}, \ldots, x_{\sigma(n)}) \in \{0,1\}^n$. For $x \in \{0,1\}^n$, we define $|x| = x_1 + x_2 + \cdots + x_n$. A function $\phi\colon \{0,1\}^n \to \mathbb{R}$ is called *symmetric* if $\phi(x) = \phi(\sigma x)$ for every $x \in \{0,1\}^n$ and every $\sigma \in S_n$. Equivalently, ϕ is symmetric if $\phi(x)$ is uniquely

determined by $|x|$. Symmetric functions on $\{0, 1\}^n$ are intimately related to univariate polynomials, as borne out by Minsky and Papert's *symmetrization argument* [25]:

PROPOSITION 2.6 (Minsky and Papert). *Let $\phi: \{0, 1\}^n \to \mathbb{R}$ be a polynomial of degree d. Then there is a polynomial $p \in P_d$ such that*

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

We will need the following consequence of Minsky and Papert's technique for rational functions, pointed out in [34, Prop. 2.7].

PROPOSITION 2.7. *Let $n_1, \ldots, n_k$ be positive integers. Consider a function $F: \{0, 1\}^{n_1} \times \cdots \times \{0, 1\}^{n_k} \to \{-1, +1\}$ such that $F(x_1, \ldots, x_k) \equiv f(|x_1|, \ldots, |x_k|)$ for some $f: \{0, 1, \ldots, n_1\} \times \cdots \times \{0, 1, \ldots, n_k\} \to \{-1, +1\}$. Then for all d,*

$$R^+(F, d) = R^+(f, d).$$

3. ANALYSIS OF RANDOM HALFSPACES

In this section, we prove a certain structural property of random halfspaces. Specifically, we will fix integers $w_1, w_2, \ldots, w_n$ at random from a suitable range and analyze the sum

$$\sum_{i=1}^{n} w_i x_i$$

as x ranges over $\{0, 1\}^n$. Our objective will be to show that, for a typical choice of the weights $w_1, w_2, \ldots, w_n$, the distribution of this sum modulo $2^{\Theta(n)}$ is highly random. More precisely, we will show that the subset $X_s \subset \{0, 1\}^n$ that induces any particular sum s modulo $2^{\Theta(n)}$ is relatively large and that membership in X_s is almost uncorrelated with any polynomial of low degree. We start with a technical lemma.

LEMMA 3.1. *Let $f, g: \{0, 1\}^n \to \{0, 1\}$ be given functions. Fix an integer k with $0 \leqslant k \leqslant n/2$. For a set $S \subseteq \{1, 2, \ldots, n\}$, define $F_S: \{0, 1\}^n \to \{0, 1\}$ by*

$$F_S(x) = f(x) \wedge \left(g(x) \oplus \bigoplus_{i \in S} x_i \right).$$

Fix a real $\zeta > 0$. Then with probability at least $1 - 2^{-n + H(k/n)n + 2\zeta n}$ over a uniformly random choice of $S \in \mathscr{P}(\{1, 2, \ldots, n\})$, one has

$$(3.1) \qquad \left| \hat{F}_S(T) - \frac{1}{2} \hat{f}(T) \right| \leqslant 2^{-\zeta n - 1}, \qquad\qquad |T| \leqslant k.$$

Proof. Define $\phi: \{0, 1\}^n \to [-1/2, 1/2]$ by $\phi(x) = f(x)g(x) - \frac{1}{2} f(x)$. Define $\mathscr{S} \subseteq \mathscr{P}(\{1, 2, \ldots, n\})$ by $\mathscr{S} = \{S : |\hat{\phi}(S)| \geqslant 2^{-\zeta n - 1}\}$. By Parseval's identity (2.2),

$$(3.2) \qquad\qquad\qquad |\mathscr{S}| \leqslant 4^{\zeta n}.$$

Since $F_S(x) = \frac{1}{2} f(x) + (-1)^{\sum_{i \in S} x_i} \phi(x)$, we have

$$(3.3) \qquad \left| \hat{F}_S(T) - \frac{1}{2} \hat{f}(T) \right| = |\hat{\phi}(S \oplus T)|, \qquad S, T \subseteq \{1, 2, \ldots, n\}.$$

For a uniformly random $S \in \mathscr{P}(\{1, 2, \ldots, n\})$, the set $\{S \oplus T : |T| \leqslant k\}$ contains any fixed element of $\mathscr{P}(\{1, 2, \ldots, n\})$ with probability $2^{-n} \sum_{i=0}^{k} \binom{n}{i}$. By the union bound, we infer that

$$\mathbf{P}_S[\{S \oplus T : |T| \leqslant k\} \cap \mathscr{S} \neq \varnothing] \leqslant |\mathscr{S}| 2^{-n} \sum_{i=0}^{k} \binom{n}{i},$$

which in view of (2.1) and (3.2) is bounded from above by $2^{-n+H(k/n)n+2\zeta n}$. This observation, along with (3.3), completes the proof. $\qquad\square$

Using Lemma 3.1 and induction, we now obtain a key intermediate result.

LEMMA 3.2. *Fix an integer $k \geqslant 0$ and reals $\epsilon, \zeta \in (0, 1/2)$. Choose sets $S_0, S_1, \ldots, S_k \in \mathscr{P}(\{1, 2, \ldots, n\})$ uniformly at random. Fix any integer s and define $f \colon \{0, 1\}^n \to \{0, 1\}$ by*

$$(3.4) \qquad f(x) = 1 \qquad \Leftrightarrow \qquad \sum_{i=0}^{k} 2^i \sum_{j \in S_i} x_j \equiv s \pmod{2^{k+1}}.$$

Then with probability at least $1 - (k+1)2^{-n+H(\epsilon)n+2\zeta n}$ over the choice of $S_0, S_1, \ldots, S_k$, one has

$$(3.5) \qquad \left| \hat{f}(T) - \frac{\delta_{T,\varnothing}}{2^{k+1}} \right| \leqslant 2^{-\zeta n}, \qquad\qquad |T| \leqslant \epsilon n.$$

Proof. In view of the modular counting in (3.4), one may assume that $0 \leqslant s < 2^{k+1}$ and therefore $s = \sum_{i=0}^{k} 2^i b_i$ for some $b_0, b_1, \ldots, b_k \in \{0, 1\}$. The proof of the lemma is by induction on k for a fixed s.

The base case $k = 0$ corresponds to $f(x) = \frac{1}{2} + \frac{1}{2}(-1)^{b_0} \chi_{S_0}(x)$. One obtains (3.5) by conditioning on the event that $|S_0| > \epsilon n$, which in view of (2.1) occurs with probability no smaller than $1 - 2^{-n+H(\epsilon)n}$.

We now consider the inductive step. Define $f' \colon \{0, 1\}^n \to \{0, 1\}$ by

$$f'(x) = 1 \qquad \Leftrightarrow \qquad \sum_{i=0}^{k-1} 2^i \sum_{j \in S_i} x_j \equiv \sum_{i=0}^{k-1} 2^i b_i \pmod{2^k}.$$

Let E_1 be the event, over the choice of $S_0, \ldots, S_{k-1}$, that $|\widehat{f'}(T) - 2^{-k}\delta_{T,\varnothing}| \leqslant 2^{-\zeta n}$ for $|T| \leqslant \epsilon n$. By the inductive hypothesis,

$$(3.6) \qquad \mathbf{P}[E_1] \geqslant 1 - k2^{-n+H(\epsilon)n+2\zeta n}.$$

Let E_2 be the event, over the choice of $S_0, \ldots, S_k$, that $|\hat{f}(T) - \frac{1}{2}\widehat{f'}(T)| \leqslant 2^{-\zeta n-1}$ for $|T| \leqslant \epsilon n$. In this terminology, it suffices to show that

$$(3.7) \qquad \mathbf{P}[E_1 \wedge E_2] \geqslant 1 - (k+1)2^{-n+H(\epsilon)n+2\zeta n}.$$

Observe that

$$f(x) = f'(x) \wedge \left(g(x) \oplus \bigoplus_{i \in S_k} x_i \right),$$

where $g \colon \{0, 1\}^n \to \{0, 1\}$ is the function such that $g(x) = 1$ if and only if b_k is the $(k+1)$st least significant bit of the integer $\sum_{i=0}^{k-1} 2^i \sum_{j \in S_i} x_j$. As a result, Lemma 3.1

shows that $\mathbf{P}[E_2] \geqslant 1 - 2^{-n+H(\epsilon)n+2\zeta n}$. This bound, along with (3.6), settles (3.7) and thereby completes the induction. □

We have reached the main result of this section.

THEOREM 3.3 (Key property of random halfspaces). *Fix an integer $k \geqslant 0$ and reals $\epsilon, \zeta \in (0, 1/2)$. Choose integers $w_1, w_2, \ldots, w_n$ uniformly at random from $\{0, 1, \ldots, 2^{k+1} - 1\}$. For $s \in \mathbb{Z}$, define $f_s \colon \{0, 1\}^n \to \{0, 1\}$ by*

$$(3.8) \qquad f_s(x) = 1 \qquad \Leftrightarrow \qquad \sum_{i=1}^{n} w_i x_i \equiv s \pmod{2^{k+1}}.$$

Then with probability at least $1 - (k + 1)2^{-n+H(\epsilon)n+2\zeta n+k+1}$ over the choice of $w_1, w_2, \ldots, w_n$, one has

$$\left| \hat{f}_s(T) - \frac{\delta_{T,\varnothing}}{2^{k+1}} \right| \leqslant 2^{-\zeta n}, \qquad\qquad |T| \leqslant \epsilon n, \quad s \in \mathbb{Z}.$$

Proof. In view of the modular counting in (3.8), it suffices to prove the theorem for $s \in \{0, 1, \ldots, 2^{k+1} - 1\}$. The functions f_s have the following equivalent definition: pick sets $S_0, S_1, \ldots, S_k \in \mathscr{P}(\{1, 2, \ldots, n\})$ uniformly at random and define

$$f_s(x) = 1 \qquad \Leftrightarrow \qquad \sum_{i=0}^{k} 2^i \sum_{j \in S_i} x_j \equiv s \pmod{2^{k+1}}.$$

The proof is now complete by Lemma 3.2 and the union bound over s. □

4. ZEROING OUT CORRELATIONS BY A CHANGE OF DISTRIBUTION

Recall the setting of the previous section, where we fixed integers $w_1, w_2, \ldots, w_n$ at random from a suitable range and analyzed the sum $\sum_{i=1}^{n} w_i x_i$ as x ranged over $\{0, 1\}^n$. We showed that the subset $X_s \subset \{0, 1\}^n$ that induces any particular sum s modulo $2^{\Theta(n)}$ is relatively large and that membership in X_s has *almost* zero correlation with any given polynomial of low degree. For the purposes of this paper, the correlations with low-degree polynomials need to be *exactly* zero. In this section we show that, with respect to a suitable distribution μ_s on each X_s, membership in X_s will indeed have zero correlation with any low-degree polynomial.

A starting point in our discussion is a general statement on zeroing out the correlations of given Boolean functions $\chi_1, \chi_2, \ldots, \chi_k$ with another Boolean function f. Recall that for functions $f, g \colon X \to \mathbb{R}$ on a finite set X, we use the notation

$$\langle f, g \rangle = \frac{1}{|X|} \sum_{x \in X} f(x) g(x).$$

THEOREM 4.1. *Let $f, \chi_1, \ldots, \chi_k \colon X \to \{-1, +1\}$ be given functions on a finite set X. Suppose that*

$$(4.1) \qquad \sum_{i=1}^{k} |\langle f, \chi_i \rangle| < \frac{1}{2},$$

$$(4.2) \qquad \sum_{\substack{j=1 \\ j \neq i}}^{k} |\langle \chi_i, \chi_j \rangle| \leqslant \frac{1}{2}, \qquad\qquad i = 1, 2, \ldots, k.$$

Then there exists a probability distribution μ on X such that

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

REMARK 4.2. A comment is in order on the hypothesis of Theorem 4.1. The theorem states that if $\chi_1, \chi_2, \ldots, \chi_k$ each have a small correlation with f and, in addition, have small pairwise correlations, then a distribution exists with respect to which f is completely uncorrelated with $\chi_1, \chi_2, \ldots, \chi_k$. The latter part of the hypothesis, namely the requirement (4.2) of small pairwise correlations for $\chi_1, \chi_2, \ldots, \chi_k$, may seem unnecessary at first. In actuality, it is vital. Exponential lower bounds on the weights of linear perceptrons [27, 38] imply, by linear programming duality, the existence of functions $f, \chi_1, \chi_2, \ldots, \chi_k \colon X \to \{-1, +1\}$ such that $|\langle f, \chi_i \rangle| = \exp\{-\Theta(k)\}$, $i = 1, 2, \ldots, k$, and yet

$$(4.3) \qquad f(x) \equiv \operatorname{sgn}\left(\sum_{i=1}^{k} \alpha_i \chi_i(x) \right)$$

for some fixed reals $\alpha_1, \ldots, \alpha_k$. In this construction, the correlation of f with each χ_i is small, in fact exponentially smaller than what is assumed in Theorem 4.1; nevertheless, the representation (4.3) rules out a distribution μ with respect to which f could have zero correlation with each χ_i, because such a distribution μ would lead to the contradiction

$$0 < \mathop{\mathbf{E}}_{\mu}\left[\left| \sum_{i=1}^{k} \alpha_i \chi_i(x) \right|\right] = \mathop{\mathbf{E}}_{\mu}\left[f(x) \sum_{i=1}^{k} \alpha_i \chi_i(x) \right] = \sum_{i=1}^{k} \alpha_i \mathop{\mathbf{E}}_{\mu}[f(x)\chi_i(x)] = 0.$$

Proof of Theorem 4.1. Consider the linear system

$$(4.4) \qquad\qquad\qquad\qquad M\alpha = \gamma$$

in the unknown $\alpha \in \mathbb{R}^k$, where $M = [\langle \chi_i, \chi_j \rangle]_{i,j}$ is a matrix of order k and $\gamma = (\langle f, \chi_1 \rangle, \ldots, \langle f, \chi_k \rangle) \in \mathbb{R}^k$. Then (4.2) shows that M is strictly diagonally dominant and hence nonsingular by Theorem 2.1. Fix the unique solution α to the system (4.4). Then $2|\alpha_i| - \sum_{j=1}^{k} |\alpha_j \langle \chi_i, \chi_j \rangle| \leqslant |\langle f, \chi_i \rangle|$ for $i = 1, 2, \ldots, k$. Summing these k inequalities, we obtain

$$2\sum_{i=1}^{k} |\alpha_i| - \sum_{j=1}^{k} |\alpha_j| \sum_{i=1}^{k} |\langle \chi_i, \chi_j \rangle| \leqslant \sum_{i=1}^{k} |\langle f, \chi_i \rangle|,$$

which in view of (4.1) and (4.2) shows that $\sum_{i=1}^{k}|\alpha_i| < 1$. Therefore, the function $\mu\colon X \to \mathbb{R}$ given by

$$\mu(x) = \epsilon\left(1 - f(x)\sum_{i=1}^{k}\alpha_i\,\chi_i(x)\right)$$

is a probability distribution on X for a suitable normalizing factor $\epsilon > 0$. At last,

$$\underset{\mu}{\mathbf{E}}\left[f(x)\chi_i(x)\right] = \epsilon|X|\left(\langle f, \chi_i\rangle - \sum_{j=1}^{k}\alpha_j\,\langle \chi_i, \chi_j\rangle\right) = 0,$$

where the final equality holds by (4.4).   □

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

THEOREM 4.3. *Let $\alpha > 0$ be a sufficiently small absolute constant. Choose integers $w_1, w_2, \ldots, w_n$ uniformly at random from $\{0, 1, \ldots, 2^{\lfloor \alpha n\rfloor + 1} - 1\}$. For $s \in \mathbb{Z}$, define*

(4.5)
$$X_s = \left\{x \in \{0, 1\}^n : \sum_{i=1}^{n} w_i x_i \equiv s \pmod{2^{\lfloor \alpha n\rfloor + 1}}\right\}.$$

Then with probability at least $1 - \mathrm{e}^{-n/3}$ over the choice of $w_1, w_2, \ldots, w_n$, there is a distribution μ_s on X_s (for each s) such that

(4.6)
$$\underset{\mu_s}{\mathbf{E}}\left[p(x)\right] = \underset{\mu_t}{\mathbf{E}}\left[p(x)\right]$$

for any $s, t \in \mathbb{Z}$ and any polynomial p of degree at most αn.

Proof. Let $\alpha > 0$ be sufficiently small. We will assume throughout the proof that $n \geqslant 1/\alpha$, the theorem being trivial otherwise. Set $\epsilon = 2\alpha$, $\zeta = 1/5$, and $k = \lfloor \alpha n\rfloor$ in Theorem 3.3. Then with probability at least $1 - \mathrm{e}^{-n/3}$ over the choice of $w_1, w_2, \ldots, w_n$, one has

(4.7)
$$\left|\hat{f}_s(T) - \frac{\delta_{T,\varnothing}}{2^{\lfloor \alpha n\rfloor + 1}}\right| \leqslant 2^{-n/5}, \qquad\qquad |T| \leqslant 2\alpha n, \quad s \in \mathbb{Z},$$

where $f_s\colon \{0, 1\}^n \to \{0, 1\}$ is given by $f_s(x) = 1 \Leftrightarrow x \in X_s$. It follows that for each s,

(4.8)
$$|X_s| = 2^n\,\hat{f}_s(\varnothing) \geqslant 2^n\,(2^{-\lfloor \alpha n\rfloor - 1} - 2^{-n/5}).$$

For $f, g\colon \{0, 1\}^n \to \mathbb{R}$, we will write $\langle f, g\rangle_{X_s} = |X_s|^{-1}\sum_{x \in X_s} f(x)g(x)$. Let $\mathscr{S} \subset \mathscr{P}(\{1, 2, \ldots, n\})$ be the system of nonempty subsets of at most αn elements. Fix any $T \in \mathscr{S}$. Then for each s,

(4.9)
$$\sum_{\substack{S \in \mathscr{S} \\ S \neq T}}|\langle \chi_s, \chi_T\rangle_{X_s}| = \frac{2^n}{|X_s|}\sum_{\substack{S \in \mathscr{S} \\ S \neq T}}|\hat{f}_s(S \oplus T)| \leqslant \frac{2^n}{|X_s|}\cdot|\mathscr{S}|\,2^{-n/5} < \frac{1}{2},$$

where the final two inequalities follow from (2.1), (4.7), and (4.8). Similarly, for each s,

(4.10)
$$\sum_{S \in \mathscr{S}}|\langle f_s, \chi_s\rangle_{X_s}| = \frac{2^n}{|X_s|}\sum_{S \in \mathscr{S}}|\hat{f}_s(S)| \leqslant \frac{2^n}{|X_s|}\cdot|\mathscr{S}|\,2^{-n/5} < \frac{1}{2}.$$

In view of (4.9) and (4.10), Theorem 4.1 provides a distribution μ_s on $\{0, 1\}^n$ that is supported on X_s and obeys $\hat{\mu}_s(S) = 0$ for $S \in \mathscr{S}$. Since μ_s is a probability distribution, we additionally have $\hat{\mu}_s(\varnothing) = 2^{-n}$ for all s. In particular, the distributions μ_s have identical Fourier spectra up to coefficients of order αn, which is another way of stating (4.6).   □

5. REDUCTION TO A UNIVARIATE PROBLEM

Recall from the Introduction that the crux of our proof is to establish the existence of a halfspace $f \colon \{0, 1\}^n \to \{-1, +1\}$ that requires a rational function of degree $\Theta(n)$ for pointwise approximation within $1/3$. The purpose of this section is to reduce this task, for a suitably chosen random halfspace, to a univariate problem. The univariate problem pertains to the uniform approximation of the sign function on the set $\{\pm 1, \pm 2, \pm 3, \ldots, \pm 2^{\Theta(n)}\}$ and has been solved in previous work. Key to this univariate reduction will be the construction of probability distributions in the previous two sections.

THEOREM 5.1 (Reduction to a univariate problem). *Put $k = \lfloor \alpha n \rfloor$, where $\alpha > 0$ is the absolute constant from Theorem 4.3. Choose $w_1, w_2, \ldots, w_n$ uniformly at random from $\{0, 1, \ldots, 2^{k+1} - 1\}$. Define $f \colon \{0, 1\}^n \times \{0, 1, 2, \ldots, n\} \to \{-1, +1\}$ by*

$$f(x, t) = \operatorname{sgn}\left(\frac{1}{2} + \sum_{i=1}^{n} w_i x_i - 2^{k+1} t \right).$$

Then with probability at least $1 - \mathrm{e}^{-n/3}$ over the choice of $w_1, w_2, \ldots, w_n$, one has

$$(5.1) \qquad R^+(f, d) \geqslant R^+(\{\pm 1, \pm 2, \pm 3, \ldots, \pm 2^k\}, d), \qquad d = 0, 1, \ldots, k.$$

Proof. For $s = \pm 1, \pm 2, \pm 3, \ldots, \pm 2^k$, define $X_s \subseteq \{0, 1\}^n$ by (4.5). Then by Theorem 4.3, with probability at least $1 - \mathrm{e}^{-n/3}$ there is a distribution μ_s on X_s for each s such that

$$(5.2) \qquad \mathop{\mathbf{E}}_{\mu_s} [p(x)] = \mathop{\mathbf{E}}_{\mu_r} [p(x)]$$

for any $s, r \in \{\pm 1, \pm 2, \pm 3, \ldots, \pm 2^k\}$ and any polynomial p of degree no greater than k. In the remainder of the proof, we will work with a fixed choice of weights $w_1, w_2, \ldots, w_n$ for which the described distributions μ_s exist.

Suppose that $R^+(f, d) < \epsilon$ where $0 < \epsilon < 1$ and $0 \leqslant d \leqslant k$. Then there are degree-$d$ polynomials p, q on $\mathbb{R}^n \times \mathbb{R}$ such that on the domain of f,

$$(5.3) \qquad 0 < (1 - \epsilon) q(x, t) \leqslant p(x, t) f(x, t) \leqslant (1 + \epsilon) q(x, t).$$

On the support of μ_s (for $s = \pm 1, \pm 2, \pm 3, \ldots, \pm 2^k$), the linear form

$$\ell(x, s) = 2^{-k-1}\left(\sum_{i=1}^{n} w_i x_i - s \right)$$

obeys $\ell(x, s) \in \{0, 1, 2, \ldots, n\}$ and $f(x, \ell(x, s)) = \operatorname{sgn} s$. Letting $t = \ell(x, s)$ in (5.3) and passing to expectations,

$$0 < \mathop{\mathbf{E}}_{x \sim \mu_s} [q(x, \ell(x, s))] (1 - \epsilon) \leqslant \mathop{\mathbf{E}}_{x \sim \mu_s} [p(x, \ell(x, s))] \operatorname{sgn} s$$

$$\leqslant \mathop{\mathbf{E}}_{x \sim \mu_s} [q(x, \ell(x, s))] (1 + \epsilon).$$

It follows from (5.2) that $\mathbf{E}_{\mu_s}[p(x, \ell(x, s))] = P(s)$ and $\mathbf{E}_{\mu_s}[q(x, \ell(x, s))] = Q(s)$ for some $P, Q \in P_d$ and all s. As a result, $R^+(\{\pm 1, \pm 2, \pm 3, \ldots, \pm 2^k\}, d) \leqslant \epsilon$, the approximant in question being P/Q. $\qquad\Box$

It remains to rewrite the previous theorem in terms of functions on the hypercube $\{0, 1\}^{2n}$ rather than the set $\{0, 1\}^n \times \{0, 1, 2, \ldots, n\}$.

THEOREM 5.2. *Put* $k = \lfloor \alpha n \rfloor$, *where* $\alpha > 0$ *is the absolute constant from Theorem 4.3. Choose* $w_1, w_2, \ldots, w_n$ *uniformly at random from* $\{0, 1, \ldots, 2^{k+1} - 1\}$. *Define* $f: \{0, 1\}^{2n} \to \{-1, +1\}$ *by*

$$f(x) = \operatorname{sgn}\left(\frac{1}{2} + \sum_{i=1}^{n} w_i x_i - 2^{k+1} \sum_{i=n+1}^{2n} x_i\right).$$

Then with probability at least $1 - e^{-n/3}$ *over the choice of* $w_1, w_2, \ldots, w_n$, *one has*

$$R^+(f, d) \geqslant R^+(\{\pm 1, \pm 2, \pm 3, \ldots, \pm 2^k\}, d), \qquad d = 0, 1, \ldots, k.$$

Proof. Immediate from Proposition 2.7 and Theorem 5.1. □

6. MAIN RESULT AND GENERALIZATIONS

We now combine the newly obtained result on rational approximation with known results from Section 2 to prove the main theorem of this work.

THEOREM 6.1 (Main result). *Fix sufficiently small absolute constants* $\alpha > 0$ *and* $\beta = \beta(\alpha) > 0$. *Choose integers* $w_1, w_2, \ldots, w_n \in \{0, 1, \ldots, 2^{\lfloor \alpha n \rfloor + 1} - 1\}$ *uniformly at random. Then with probability at least* $1 - e^{-n/3}$, *the function* $f: \{0, 1\}^{2n} \to \{-1, +1\}$ *given by*

$$f(x) = \operatorname{sgn}\left(\frac{1}{2} + \sum_{i=1}^{n} w_i x_i - 2^{\lfloor \alpha n \rfloor + 1} \sum_{i=n+1}^{2n} x_i\right)$$

obeys

(6.1) $\deg_\pm(f \wedge f) \geqslant \lfloor \beta n \rfloor.$

Proof. Theorem 5.2 shows that with probability at least $1 - e^{-n/3}$ over the choice of $w_1, w_2, \ldots, w_n$, one has

(6.2) $R^+(f, d) \geqslant R^+(S, d), \qquad d = 0, 1, \ldots, \lfloor \alpha n \rfloor,$

where $S = \{\pm 1, \pm 2, \pm 3, \ldots, \pm 2^{\lfloor \alpha n \rfloor}\}$ and $\alpha > 0$ is the absolute constant from Theorem 4.3. In the remainder of the proof, we will condition on this event.

Suppose now that $\deg_\pm(f \wedge f) < \lfloor \beta n \rfloor$, where β is a constant to be chosen later subject to $0 < \beta < \alpha/4$. Then Theorem 2.5 implies that $R^+(f, \lfloor 4\beta n \rfloor) < 1/2$, which in view of (6.2) leads to $R^+(S, \lfloor 4\beta n \rfloor) < 1/2$. The last inequality violates Theorem 2.2 for small enough $\beta > 0$. Thus, (6.1) holds for β small enough. □

Recall that the technical crux of this paper is an optimal lower bound for the rational approximation of a halfspace. We will have occasion to appeal to this result again, and for this reason we formulate it as a theorem in its own right.

THEOREM 6.2. *A family of halfspaces* $h_n: \{0, 1\}^n \to \{-1, +1\}$, $n = 1, 2, 3, \ldots$, *exists such that*

(6.3) $R^+(h_n, d) = 1 - \exp\left\{-\Theta\left(\frac{n}{d}\right)\right\}, \qquad d = 1, 2, \ldots, \Theta(n).$

Proof. The lower bound in (6.3) is immediate from Theorem 5.2 and the univariate lower bounds in Theorem 2.2.

Next, every halfspace $h_n: \{0, 1\}^n \to \{-1, +1\}$ constructed in Theorem 5.2 trivially obeys $R^+(h_n, 1) < 1 - \exp\{-\Theta(n)\}$. For $0 < \xi < 1$, Newman's classical work [28] shows

that $R^+([-1, -\xi] \cup [\xi, 1], d) \leqslant 1 - \xi^{\Theta(1/d)}$, whence by composition of the approximants one obtains the upper bound in (6.3). □

Mixed intersection. Theorem 6.1 shows that the intersection of two halfspaces has the asymptotically highest threshold degree. At the same time, Beigel et al. [6] showed that the intersection of a constant number of majority functions on $\{0, 1\}^n$, which are particularly simple halfspaces, has threshold degree $O(\log n)$. We now derive a lower bound of $\Omega(\sqrt{n \log n})$ on the threshold degree of the intersection of a halfspace and a majority function, which improves quadratically on the previous bound in [34] and essentially matches the upper bound, $O(\sqrt{n} \log n)$, given below in Remark 6.4.

THEOREM 6.3. *A family of halfspaces* $h_n \colon \{0, 1\}^n \to \{-1, +1\}$, $n = 1, 2, 3, \ldots$, *exists such that*

$$(6.4) \qquad \deg_{\pm}(h_n \wedge \mathrm{MAJ}_n) = \Theta(\sqrt{n \log n}).$$

Proof. The lower bound in (6.4) is immediate from Theorems 2.3, 2.5, and 6.2. The upper bound in (6.4) is immediate from Theorems 2.3, 2.4, and 6.2. □

REMARK 6.4. The construction of Theorem 6.3 is essentially best possible in that every sequence of halfspaces $h_n \colon \{0, 1\}^n \to \{-1, +1\}$, $n = 1, 2, 3, \ldots$, obeys

$$(6.5) \qquad \deg_{\pm}(h_n \wedge \mathrm{MAJ}_n) = O(\sqrt{n} \log n).$$

To derive this upper bound, recall that $R^+(h_n, 1) < 1 - \exp\{-\Theta(n \log n)\}$ for every halfspace $h_n \colon \{0, 1\}^n \to \{-1, +1\}$, by a classical result due to Muroga [26]. Since $R^+([-1, -\xi] \cup [\xi, 1], d) < 1 - \xi^{\Theta(1/d)}$ for $0 < \xi < 1$ by Newman [28], we obtain by composition of approximants that $R^+(h_n, d) < 1 - \exp\{-\Theta(\{n \log n\}/d)\}$. This settles (6.5) in view of Theorems 2.3 and 2.4.

Threshold density. In addition to threshold degree, several other complexity measures are of interest when sign-representing Boolean functions by real polynomials. One such complexity measure is *density*, defined as the least k for which a given function can be sign-represented by a linear combination of k parity functions. Formally, for a given function $f \colon \{0, 1\}^n \to \{-1, +1\}$, the *threshold density* $\mathrm{dns}(f)$ is the minimum size $|\mathscr{S}|$ of a family $\mathscr{S} \subseteq \mathscr{P}(\{1, 2, \ldots, n\})$ such that

$$f(x) \equiv \mathrm{sgn}\left(\sum_{S \in \mathscr{S}} \lambda_S \chi_S(x)\right)$$

for some reals λ_S, $S \in \mathscr{S}$. It is clear from the definition that $\mathrm{dns}(f) \leqslant 2^n$ for all functions $f \colon \{0, 1\}^n \to \{-1, +1\}$, and we will show that the intersection of two halfspaces on $\{0, 1\}^n$ has threshold density $2^{\Theta(n)}$.

To this end, we recall an elegant technique for converting Boolean functions with high threshold degree into Boolean functions with high threshold density, due to Krause and Pudlák [21, Prop. 2.1]. Their construction sends a function $f \colon \{0, 1\}^n \to \{-1, +1\}$ to the function $f^{\mathrm{KP}} \colon (\{0, 1\}^n)^3 \to \{-1, +1\}$ given by

$$f^{\mathrm{KP}}(x, y, z) = f(\ldots, (\overline{z_i} \wedge x_i) \vee (z_i \wedge y_i), \ldots).$$

THEOREM 6.5 (Krause and Pudlák). *Every function* $f\colon \{0, 1\}^n \to \{-1, +1\}$ *obeys*

$$\mathrm{dns}(f^{\mathrm{KP}}) \geqslant 2^{\deg_{\pm}(f)}.$$

We are now in a position to obtain the desired density results.

THEOREM 6.6. *A family of halfspaces* $h_n\colon \{0, 1\}^n \to \{-1, +1\}$, $n = 1, 2, 3, \ldots,$ *exists such that*

$$(6.6) \qquad\qquad\qquad \mathrm{dns}(h_n \wedge h_n) \geqslant \exp\{\Theta(n)\},$$

$$(6.7) \qquad\qquad\qquad \mathrm{dns}(h_n \wedge \mathrm{MAJ}_n) \geqslant \exp\{\Theta(\sqrt{n \log n})\}.$$

Proof. The parity of several parity functions is another parity function. As a result,

$$(6.8) \qquad\qquad\qquad \max_{h_n}\{\mathrm{dns}(h_n \wedge h_n)\} \geqslant \max_{F}\{\mathrm{dns}(F \wedge F)\},$$

where the maximum on the left is over all halfspaces $h_n\colon \{0, 1\}^n \to \{-1, +1\}$ and the maximum on the right is over arbitrary functions $F\colon \{0, 1\}^m \to \{-1, +1\}$ (for arbitrary m) such that $\mathrm{dns}(F) \leqslant n$. For each $n = 1, 2, 3, \ldots,$ Theorem 6.1 ensures the existence of a halfspace $f_n\colon \{0, 1\}^n \to \{-1, +1\}$ with $\deg_{\pm}(f_n \wedge f_n) \geqslant \Omega(n)$. By Theorem 6.5, the function $(f_n \wedge f_n)^{\mathrm{KP}} = f_n{}^{\mathrm{KP}} \wedge f_n{}^{\mathrm{KP}}$ has threshold density $\exp\{\Omega(n)\}$. Since $\mathrm{dns}(f_n{}^{\mathrm{KP}}) \leqslant 4n + 1$, the right member of (6.8) is at least $\exp\{\Omega(n)\}$. This completes the proof of (6.6).

The proof of (6.7) is analogous, with Theorem 6.3 used instead of Theorem 6.1. $\square$

The lower bounds in Theorem 6.6 are essentially optimal. Specifically, (6.6) is tight for trivial reasons, whereas the lower bound (6.7) nearly matches the upper bound of $\exp\{\Theta(\sqrt{n} \log^2 n)\}$ that follows from (6.5).

ACKNOWLEDGMENTS

REFERENCES

[1] M. Alekhnovich, M. Braverman, V. Feldman, A. R. Klivans, and T. Pitassi. The complexity of properly learning simple concept classes. *J. Comput. Syst. Sci.*, 74(1):16–34, 2008.

[2] A. Ambainis, A. M. Childs, B. Reichardt, R. Špalek, and S. Zhang. Any AND-OR formula of size N can be evaluated in time $N^{1/2+o(1)}$ on a quantum computer. In *Proceedings of the Forty-Eighth Annual IEEE Symposium on Foundations of Computer Science* (FOCS), pages 363–372, 2007.

[3] R. I. Arriaga and S. Vempala. An algorithmic theory of learning: Robust concepts and random projection. *Mach. Learn.*, 63(2):161–182, 2006.

[4] J. Aspnes, R. Beigel, M. L. Furst, and S. Rudich. The expressive power of voting polynomials. *Combinatorica*, 14(2):135–148, 1994.

[5] R. Beigel. Perceptrons, PP, and the polynomial hierarchy. *Computational Complexity*, 4:339–349, 1994.

[6] R. Beigel, N. Reingold, and D. A. Spielman. PP is closed under intersection. *J. Comput. Syst. Sci.*, 50(2):191–202, 1995.

[7] A. Blum and R. Kannan. Learning an intersection of a constant number of halfspaces over a uniform distribution. *J. Comput. Syst. Sci.*, 54(2):371–380, 1997.

[8] A. L. Blum and R. L. Rivest. Training a 3-node neural network is NP-complete. *Neural Networks*, 5:117–127, 1992.

[9] H. Buhrman, N. K. Vereshchagin, and R. de Wolf. On computation and communication with small bias. In *Proceedings of the Twenty-Second Annual IEEE Conference on Computational Complexity* (CCC), pages 24–32, 2007.

[10] E. Farhi, J. Goldstone, and S. Gutmann. A quantum algorithm for the Hamiltonian NAND tree. *Theory of Computing*, 4(1):169–190, 2008.

[11] S. A. Gershgorin. Über die Abgrenzung der Eigenwerte einer Matrix. *Izv. Akad. Nauk. U.S.S.R. Otd. Fiz.-Mat. Nauk*, 7:749–754, 1931.

[12] J. C. Jackson. An efficient membership-query algorithm for learning DNF with respect to the uniform distribution. *J. Comput. Syst. Sci.*, 55(3):414–440, 1997.

[13] S. Jukna. *Extremal Combinatorics with Applications in Computer Science*. Springer-Verlag, Berlin, 2001.

[14] S. Khot and R. Saket. On hardness of learning intersection of two halfspaces. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing* (STOC), pages 345–354, 2008.

[15] A. R. Klivans. *A Complexity-Theoretic Approach to Learning*. PhD thesis, MIT, 2002.

[16] A. R. Klivans, P. M. Long, and A. K. Tang. Baum's algorithm learns intersections of halfspaces with respect to log-concave distributions. In *Proceedings of the Thirteenth International Workshop on Randomization and Computation* (RANDOM), pages 588–600, 2009.

[17] A. R. Klivans, R. O'Donnell, and R. A. Servedio. Learning intersections and thresholds of halfspaces. *J. Comput. Syst. Sci.*, 68(4):808–840, 2004.

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

[19] A. R. Klivans and R. A. Servedio. Learning intersections of halfspaces with a margin. *J. Comput. Syst. Sci.*, 74(1):35–48, 2008.

[20] A. R. Klivans and A. A. Sherstov. Cryptographic hardness for learning intersections of halfspaces. *J. Comput. Syst. Sci.*, 75(1):2–12, 2009. Preliminary version in *Proceedings of the Forty-Seventh Annual IEEE Symposium on Foundations of Computer Science* (FOCS), 2006.

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

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

[23] S. Kwek and L. Pitt. PAC learning intersections of halfspaces with membership queries. *Algorithmica*, 22(1/2):53–75, 1998.

[24] T. Lee. A note on the sign degree of formulas, 2009. Available at http://arxiv.org/abs/0909.4607.

[25] M. L. Minsky and S. A. Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, Cambridge, Mass., 1969.

[26] S. Muroga. *Threshold Logic and Its Applications*. John Wiley & Sons, New York, 1971.

[27] J. Myhill and W. H. Kautz. On the size of weights required for linear-input switching functions. *IRE Trans. on Electronic Computers*, 10(2):288–290, 1961.

[28] D. J. Newman. Rational approximation to $|x|$. *Michigan Math. J.*, 11(1):11–14, 1964.

[29] R. O'Donnell and R. A. Servedio. New degree bounds for polynomial threshold functions. In *Proceedings of the Thirty-Fifth Annual ACM Symposium on Theory of Computing* (STOC), pages 325–334, 2003.

[30] R. Paturi and M. E. Saks. Approximating threshold circuits by rational functions. *Inf. Comput.*, 112(2):257–272, 1994.

[31] A. A. Razborov and A. A. Sherstov. The sign-rank of AC^0. *SIAM J. Comput.*, 39(5):1833–1855, 2010. Preliminary version in *Proceedings of the Forty-Ninth Annual IEEE Symposium on Foundations of Computer Science* (FOCS), 2008.

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

[33] M. E. Saks. Slicing the hypercube. *Surveys in Combinatorics*, pages 211–255, 1993.

[34] A. A. Sherstov. The intersection of two halfspaces has high threshold degree. In *Proceedings of the Fiftieth Annual IEEE Symposium on Foundations of Computer Science* (FOCS), pages 343–362, 2009.

[35] A. A. Sherstov. Separating AC^0 from depth-2 majority circuits. *SIAM J. Comput.*, 38(6):2113–2129, 2009. Preliminary version in *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing* (STOC), 2007.

[36] A. A. Sherstov. The pattern matrix method. *SIAM J. Comput.*, 40(6):1969–2000, 2011. Preliminary version in *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing* (STOC), 2008.

[37] A. A. Sherstov. The unbounded-error communication complexity of symmetric functions. *Combinatorica*, 31(5):583–614, 2011. Preliminary version in *Proceedings of the Forty-Ninth Annual IEEE Symposium on Foundations of Computer Science* (FOCS), 2008.

[38] K.-Y. Siu and J. Bruck. On the power of threshold circuits with small weights. *SIAM J. Discrete Math.*, 4(3):423–435, 1991.

[39] K.-Y. Siu, V. P. Roychowdhury, and T. Kailath. Rational approximation techniques for analysis of neural networks. *IEEE Transactions on Information Theory*, 40(2):455–466, 1994.

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

[41] S. Vempala. A random sampling based algorithm for learning the intersection of halfspaces. In *Proceedings of the Thirty-Eighth Annual IEEE Symposium on Foundations of Computer Science* (FOCS), pages 508–513, 1997.

[42] E. I. Zolotarev. Application of elliptic functions to questions of functions deviating least and most from zero. *Izvestiya Imp. Akad. Nauk*, 30(5), 1877.