POLYNOMIAL THRESHOLD FUNCTIONS, HYPERPLANE ARRANGEMENTS, AND RANDOM TENSORS

PIERRE BALDI AND ROMAN VERSHYNIN

Abstract. A simple way to generate a Boolean function is to take the sign of a real polynomial in n variables. Such Boolean functions are called polynomial threshold functions. How many low-degree polynomial threshold functions are there? The partial case of this problem for degree $d = 1$ was solved by Zuev in 1989, who showed that the number $T(n, 1)$ of linear threshold functions satisfies $\log_2 T(n, 1) \approx n^2$, up to smaller order terms. However the number of polynomial threshold functions for any higher degrees, including $d = 2$, has remained open. We settle this problem for all fixed degrees $d \geq 1$, showing that $\log_2 T(n, d) \approx n\binom{n}{\leq d}$. The solution relies on connections between the theory of Boolean threshold functions, hyperplane arrangements, and random tensors. Perhaps surprisingly, it uses also a recent result of E. Abbe, A. Shpilka, and A. Wigderson on Reed-Muller codes.

Contents

1. Introduction

1.1. **The problem.** Neural networks and deep learning models and applications [85] rely on a simplified neuronal model that goes back at least to the work of W. McCulloch and W. Pitts in the 1940s [56]. In this model, a neuron is viewed as a processing unit which, given n inputs $x = (x_1, \ldots, x_n)$, produces an output of the form $y = f(s) = f(\sum w_i x_i + t)$ where the coefficients w_i represent the synaptic weights, t is a threshold, the weighted average s is the activation, and f is the transfer function. When the inputs are restricted to the Boolean cube $\{-1, 1\}^n$ and f is the sign function ($f = \mathrm{sgn}$), the neuron operates as a Boolean linear threshold function.

Date: July 23, 2019.
Work in part supported by DARPA grant D17AP00002 to P. B. and U.S. Air Force grant FA9550-18-1-0031 to R. V.

In search for both more powerful computational models, as well as greater biological realism that may take into account non-linear interactions between synapses along neuronal dendritic trees, it is natural to replace the linear activation with a polynomial activation $p(x)$ of degree d so that $y = f(p(x))$ [9]. Again, considering the Boolean case, a Boolean function $f : \{-1,1\}^n \to \{-1,1\}$ is called a *polynomial threshold function* if it has the form:

$$f(x) = \operatorname{sgn}(p(x))$$

for some real-valued polynomial $p : \mathbb{R}^n \to \mathbb{R}$ that has no roots in the Boolean cube $\{-1,1\}^n$. Thus, f takes the value 1 at any point where the polynomial p is positive, and -1 at any point where the polynomial p is negative. Up to a factor of two due to the sgn operation, polynomial threshold functions can be identified with partitions of the Boolean cube $\{-1,1\}^n$ into two classes by polynomial surfaces corresponding to $p(x) = 0$. Figure 1 illustrates the partitions of the two-dimensional cube obtained by linear and quadratic threshold functions, associated with polynomials of degree one and two respectively. *The main goal in this paper is to estimate the number of different polynomial threshold functions of degree d.*

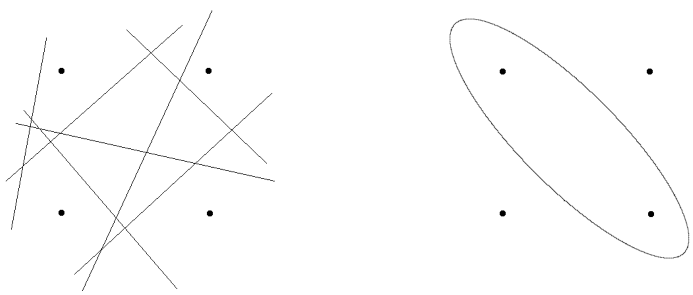

Figure 1. On the left, all 7 possible linear partitions of the Boolean square $\{-1,1\}^2$ are shown. These define 14 Boolean linear threshold functions of two variables. On the right, using quadratic surfaces (ellipses), one can realize on additional partition, for a total of 8 possible quadratic partitions. These partitions are associated with $2 \cdot 8 = 16$ Boolean quadratic threshold functions of two variables corresponding, in this case, to all possible Boolean functions of two variables.

1.2. **General background.** As previously indicated, the study of linear and polynomial threshold functions is implicit in some of the first models of neural activity by W. McCulloch and W. Pitts in the 1940s [56]. Linear threshold functions were studied by T. Cover [28], S. Muroga [60], M. Minsky and S. Papert in their book on perceptrons [59], and others in the 1960s. Since then, linear and polynomial threshold functions have been extensively used and studied in complexity theory, machine learning, and network theory; see, for instance, [8, 9, 10, 21, 83, 16, 51, 87, 7, 18, 3, 46, 47, 30, 68, 69, 17, 43]. An introduction to polynomial threshold functions can be found in [67, Chapter 5], [6, Chapter 4], and [83]. Linear and polynomial threshold functions remain a fundamental model for biological or neuromorphic neurons and, together with their continuous approximations, are at the center of all the current developments and applications of deep learning [85, 12].

1.3. **The notion of capacity.** The standard description of the neuronal model given above focuses on the processing aspect of the neuron, its input-output function. However, there is a second, at least equally important aspect, which is the storage aspect. As described, a neuron is also a storage device capable of storing information in the linear or polynomial synaptic weights of its activation function. Thus it is natural to ask how many bits can be stored in a neuron. This question is intimately related to the recently-introduced notion of capacity [13, 14] .

Given a class of functions A, such as all the functions that can be implemented by a neuron–or a network of neurons–as the synaptic weights are varied, we define the capacity of $C(A)$ as the binary logarithm $C(A) = \log_2 |A|$. In the discrete case, the only one to be considered here, $|A|$ is simply the number of functions in A (in the continuous case, one must define a notion of volume). Thus the capacity is the number of bits required to specify an element of A. Remarkably, when A is associated with a neural architecture comprising one or many neurons, the capacity can be viewed as the number of effective bits that can be extracted from a training set and stored in the neural architecture[14]. Estimating the capacity of a single linear threshold function has a long history reviewed below and, recently, we were able to estimate the capacity of networks of linear threshold function. Thus, the primary goal here is to begin extending these results beyond the linear case by estimating the capacity of a single polynomial threshold function.

1.4. **The capacity of linear threshold functions.** There are 2^{2^n} Boolean functions of n variables. Let $T(n,d)$ denote the number of polynomial threshold functions of fixed degree d. The asymptotical behavior of $T(n,d)$ has been known only for the linear case $d = 1$. The work of T. Cover [28] and others used a simple hyperplane counting argument to show that $T(n,1)$ is upperbounded by 2^{n^2}. Recursive constructions by S. Muroga [60] and others provided lower bounds of the form $2^{\alpha n^2}$ with values of α that were significantly below 1.

Yu. Zuev [109, 110] was finally able to show in 1989 that the upper bound 2^{n^2} is asymptotically tight: the number of linear threshold functions satisfies:

$$\left(1 - \frac{10}{\log n}\right) \cdot n^2 \le \log_2 T(n,1) \le n^2. \tag{1.1}$$

J. Kahn, J. Komlós, E. Szemerédi [41, Section 4] further improved this result to:

$$\log_2 T(n,1) = n^2 - n \log_2 n \pm O(n).$$

Although Zuev's breakthrough led to some progress in understanding linear threshold functions (see e.g. [71, 111, 31, 39, 40, 44]), the same problem for higher degrees has remained open. M. Saks explicitly asked about the asymptotical behavior of $T(n,d)$ in 1993 [83, Problem 2.35]. Even the asymptotic behavior of the number of quadratic threshold functions $T(n,2)$ has so far remained unknown.

1.5. **The capacity of polynomial threshold functions: the main result.** Each of the 2^{2^n} Boolean functions of n variables can be expressed as a polynomial of degree at most n: to see this, write the function f in conjunctive (or disjunctive) normal form, or take the Fourier transform of f. In particular, every Boolean function f is a polynomial threshold function, but the polynomial that represents f often has high degree. A conjecture of J. Aspnes *et al.* [7] and C. Wang and A. Williams [104] states that, for most Boolean functions $f(x)$, the lowest degree of $p(x)$ such that $f(x) = \operatorname{sgn}(p(x))$ is either $\lfloor n/2 \rfloor$ or $\lceil n/2 \rceil$. M. Anthony [5] and independently N. Alon (see [83]) proved one half of this conjecture, showing that for most Boolean functions the lower degree of $p(x)$ is at least $\lceil n/2 \rceil$. The other half of the conjecture was settled, in an approximate sense, by R. O'Donnell and R. A. Servedio [68] who gave an upper bound $n/2 + O(\sqrt{n \log n})$ on the degree of $p(x)$.

While low degree polynomial threshold functions may be relatively rare within the space of Boolean functions, they are of particular interest both theoretically and practically, due to their functional simplicity and their utilization in biological modeling and neural network applications. Thus the most fundamental open question regarding low-degree polynomial threshold functions is:

> *How many low-degree polynomial threshold functions are there? Equivalently, how many different ways are there to partition the Boolean cube by polynomial surfaces of low degree?* Equivalently how many bits can effectively be stored in the coefficients of a polynomial threshold function?

In the following theorem, we settle the problem for all fixed degrees $d \in \mathbb{N}$.

Theorem 1.1. *For any positive integers n and d such that $1 \le d \le n^{0.9}$, the number of Boolean polynomial threshold functions $T(n,d)$ satisfies*[1]

$$\Big(1 - \frac{C}{\log n}\Big)^d \cdot n \binom{n}{\le d} \le \log_2 T(n,d) \le n \binom{n}{\le d}.$$

In this theorem and the rest of the paper, C denotes a positive absolute constant; its value does not depend on n or d. The exact value of C may be different in different parts of this paper. For linear threshold functions, i.e. for $d = 1$, Theorem 1.1 yields Zuev's result (1.1) up to the absolute constant C.

The upper bound in Theorem 1.1 holds for all $1 \le d \le n$; this bound is known and can be derived from counting regions in hyperplane arrangements; we reprove it in Section 3 for completeness. The lower bound in Theorem 1.1 is new. It will be clear from the argument that the exponent 0.9 in the constraint on d can be replaced by any constant strictly less than 1 at the cost of changing the absolute constant C.

For small degrees d, namely for $d = o(\log n)$, the factor $(1 - C/\log n)^d$ becomes $1 - o(1)$ and Theorem 1.1 yields in this case the asymptotically tight bound on the capacity:

$$\log_2 T(n,d) = (1 - o(1)) \cdot n \binom{n}{\le d}.$$

To better understand this bound, note that a general polynomial of degree d has $\binom{n}{\le d}$ monomial terms. Thus, Theorem 1.1 yields the following result on communication complexity, and learning:

> *To communicate a polynomial threshold function, one needs to spend approximately n bits per monomial term. During learning, approximately n bits can be stored per monomial term.*

In some situations, it may be desirable to have a simpler estimate of $T(n,d)$ that is free of binomial sums. For this purpose, we can simplify the conclusion of Theorem 1.1 and state it as follows:

Corollary 1.2. *For any integers n and d such that $n > 1$ and $1 \le d \le n^{0.9}$, the number of Boolean polynomial threshold functions $T(n,d)$ satisfies:*

$$\Big(1 - \frac{C}{\log n}\Big)^d \cdot \frac{n^{d+1}}{d!} < \log_2 T(n,d) < \frac{n^{d+1}}{d!}.$$

The upper bound in Corollary 1.2 actually holds for all $n > 1$, $1 \le d \le n$. We derive Corollary 1.2 from Theorem 1.1 in Section 7 by a careful analysis of the underlying binomial sums.

---

[1] Here and in the rest of the paper, $\binom{n}{\le d}$ denotes the binomial sum up to term d, i.e. $\binom{n}{\le d} := \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d}$.

For small degrees d, namely for $d = o(\log n)$, the factor $(1 - C/\log n)^d$ becomes $1 - o(1)$ and Corollary 1.2 yields in this case the asymptotically tight bound on the capacity:

$$\log_2 T(n, d) = (1 - o(1)) \cdot \frac{n^{d+1}}{d!}.$$

1.6. **Prior work.** Prior to our work, an upperbound on $T(n, d)$ that scales like $n^{d+1}/d!$ was known [9, 6]. The upperbounds given in Theorem 1.1 and Corollary 1.2 are more precise and general. The best known lower bound was given by M. Saks [83] in 1993:

$$\log_2 T(n, d) \geq \binom{n}{d+1}. \tag{1.2}$$

For all small degrees d, there is a multiplicative gap of size approximately $d + 1$ between the optimal upper bound $n\binom{n}{\leq d}$ in Theorem 1.1 and the lower bound (1.2). For example, $T(n, 2) \leq n\binom{n}{\leq 2} \approx n^3/2$ but (1.2) only gives $T(n, 2) \geq n^3/6$. Our work closes this gap.

The asymptotically sharp result (1.1) about linear threshold functions has a remarkably short proof [110]. It quickly follows from a combination of two results, one in enumerative combinatorics and the other one in probability. The combinatorial result is a consequence of Zaslavsky's formula for *hyperplane arrangements* [107], and the probabilistic result is Odlyzko's theorem on spans of *random vectors* with independent ± 1 coordinates [66]. Odlyzko's theorem is a stronger, *resilience* version of known results on the *singularity of random matrices*, results that state that a random matrix with ± 1 entries has full rank with high probability. The original results on the singularity problem are due to J. Komlós [49, 50]. More recently, the singularity problem has been actively studied in random matrix theory. A significant number of extensions and improvements on the result of J. Komlós are now available, see e.g. [41, 90, 26, 91, 75, 95, 76, 77, 2, 78, 93, 20, 79, 62, 64, 100, 81, 36, 15, 96, 97, 54, 24, 98], culminating in the very recent proof by K. Tikhomirov [99] providing optimal estimates for the probability that a random ± 1 matrix be singular.

1.7. **Our approach.** Zuev's approach can be extended from linear threshold functions to polynomial threshold functions by lifting the problem into the tensor product space $(\mathbb{R}^n)^{\otimes d}$. The combinatorial part of the argument generalizes without any problem, but the probabilistic part is less obvious, because the theory of random tensors is not sufficiently developed yet. In [11] we developed some of theory in order to prove a version of Theorem 1.1, but recently found a theorem by E. Abbe, A. Shpilka, and A. Wigderson [1] on the singularity of random tensors, developed in the context of Reed-Muller codes, that allows one to simplify the proof. The forthcoming paper [102] gives an alternative, more general and quantitative, approach to the singularity problem for random tensors. In the current paper, we prove a version of Odlyzko's resilience result [66] for random tensors. Our proof of resilience is inspired by Odlyzko's proof and it uses the result of [1]. The argument is non-trivial due to the lack of independence – even if the entries of a random vector x are stochastically independent, the entries of the random tensor $x^{\otimes d}$ are not.

1.8. **Outline of paper.** The rest of the paper is devoted to the proof of Theorem 1.1. The next two sections deal with the combinatorial part of the argument. In Section 2, we explain a canonical correspondence between linear threshold functions and hyperplane arrangements, and we discuss known bounds on the number of regions determined by hyperplane arrangements. In Section 3, we linearize polynomial threshold functions by lifting them into the tensor product space, which then reduces polynomial threshold functions to linear threshold functions and the corresponding hyperplane arrangements. This allows us to quickly prove the (known) upper bound in Theorem 1.1 at the end of Section 3.

Next we turn to the probabilistic part of the argument, in order to derive the lower bound. In Section 4, we explain the result of E. Abbe, A. Shpilka, and A. Wigderson [1] on the linear independence of random tensors (Theorem 4.1), and state a new resilience version of this result (Theorem 4.2). In Section 4.4, using this resilience version, we derive the lower bound of Theorem 1.1. In Sections 5 and 6, we prove the resilience result for random tensors. The argument uses the Littlewood-Offord Lemma for sums of independent random variables, so we explain the Littlewood-Offord lemma in Section 5. In Section 6, we prove the resilience result (Theorem 4.2), thus completing the entire argument. In Section 7, we deduce Corollary 1.2. In Section 8, we describe several possible extensions and related open questions.

Acknowledgements. The authors are grateful to Michael Forbes who drew their attention to the work of E. Abbe, A. Shpilka, and A. Wigderson [1], which lead to a great simplification of the original proof of Theorem 1.1. The authors also thank the anonymous reviewers for their useful comments and suggestions.

2. Hyperplane arrangements

There is a natural correspondence between threshold functions and regions of hyperplane arrangements, a classical topic in enumerative combinatorics that has been studied for decades [107]; see [88], [55, Section 6]. To see the connection, let us fix a finite subset $S \subset \mathbb{R}^n \setminus \{0\}$ and consider all *homogeneous linear threshold functions* on S, i.e. functions $f : S \to \mathbb{R}$ of the form

$$f_a(x) = \operatorname{sgn}(\langle a, x\rangle)$$

where $a \in \mathbb{R}^n$ is a fixed vector. Consider the collection ("arrangement") of hyperplanes

$$\{x^\perp : \ x \in S\},$$

where $x^\perp = \{z \in \mathbb{R}^n : \ \langle z, x\rangle = 0\}$ is the hyperplane through the origin with normal vector x. Two vectors a and b define the same homogeneous linear threshold function $f_a = f_b$ if and only if a and b lie on the same side of each of these hyperplanes. In other words, $f_a = f_b$ if and only if a and b lie in the same open component of the partition of $\mathbb{R}^n$ created by the hyperplanes $x^\perp$, with $x \in S$. Such open components are called the *regions* of the hyperplane arrangement (Figure 2). Thus we have the following lemma:

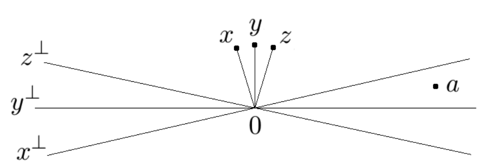

Figure 2. The set of points $S = \{x, y, z\}$ in the plane defines the arrangement of hyperplanes (lines in this cases) $x^\perp, y^\perp, z^\perp$. These hyperplanes partition the plane into six regions. Each region defines a different linear threshold function on S by the rule $f_a(x) = \operatorname{sgn}(\langle a, x\rangle)$, where a is any point taken from that region.

Lemma 2.1 (Threshold functions and hyperplane arrangements)**.** *The number of homogeneous linear threshold functions on a given finite set $S \subset \mathbb{R}^n \setminus \{0\}$ equals the number of regions of the hyperplane arrangement $\{x^\perp : \ x \in S\}$.*

This leads us to the following general question: how many regions are there in a given hyperplane arrangement? An exact formula was found by Zaslavsky [107]; see [88]. It expresses the number of regions via the Möbius function of the poset of the intersection spaces associated with the hyperplanes. Computing the Möbius function, however, may be a challenging task. Nevertheless, there are convenient bounds on the Möbius function, which yields the following result:

Lemma 2.2 (Counting regions of hyperplane arrangements)**.** *Consider an arrangement of p affine hyperplanes in $\mathbb{R}^m$, where $p \geq m$. Let $r(p,m)$ denote the number of regions of this arrangement.*

1. *We have:*
$$r(p,m) \leq \binom{p}{\leq m}. \tag{2.1}$$
2. *$r(p,m)$ is bounded below by the number of all intersection subspaces[2] defined by the hyperplanes.*
3. *If the hyperplanes are in general position, then the upper and lower bounds are the same, and each bound becomes an equality.*
4. *If all hyperplanes are central, i.e. pass through the same point, then the upper bound improves to*
$$r(p,m) \leq 2\binom{p-1}{\leq m-1}. \tag{2.2}$$
5. *If the normal vectors to the hyperplanes are in general position, then the inequality in (2.2) becomes an equality.*

To illustrate the first three parts of Lemma 2.2, consider first the line arrangement on the left in Figure 3. This line arrangement is in general position, it has seven regions, which is the same as $\binom{3}{\leq 2}$ and also the same as the number of intersection subspaces, corresponding to: three points, three lines, and one plane. As for the last two parts of Lemma 2.2, consider the line arrangement on the right of Figure 3. It has six regions but only five intersection subspaces. Moreover, since the normal vectors to the lines are in general position, part 4 of Lemma 2.2 states that the number of regions must be bounded by $2\binom{2}{\leq 1} = 6$, which is sharp.

Figure 3. Two hyperplane arrangements in $\mathbb{R}^2$

Lemma 2.2 can be derived from Zaslavsky's formula [107] and the basic properties of the Möbius function of geometric lattices [74, Section 7]; see e.g. [88, Proposition 2.4] for the derivation of (2.1) and [71] for (2.2). Originally, the upper bound (2.1) goes back to R. C. Buck [22] and the lower bound (property 2 in the lemma) was noted by Yu. Zuev [110].

[2] An intersection subspace in this lemma refers to the intersection of any subfamily of the original hyperplanes. The dimensions of an intersection subspace may range from zero (a single point) to m (intersecting an empty set of hyperplanes gives the entire space $\mathbb{R}^m$).

Both the upper and lower bounds can also be proven directly – without Zaslavsky's formula – using simple inductive arguments, see [55, Section 6] for the upper bound and [110] for the lower bound. The bound (2.2) can also be proven by induction, as in the original proof of this result due to L. Schläfli [86, pp. 209–212], which is reproduced e.g. in [106] and, more explicitly, in [6, Theorem 4.1]. See also [105].

3. Tensor lift

Our next goal is to extend the correspondence between linear threshold functions and the regions associated with hyperplane arrangements to *polynomial* threshold functions of arbitrary degree. There is a very simple way to achieve this by lifting the problem into a tensor product space.

Definition 3.1 (d-th power of vectors and sets)**.** *The d-th power of a vector $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ is the vector $x^{\le d} \in \mathbb{R}^{\binom{n}{\le d}}$ whose coordinates are purely homogeneous monomials in $x_1, \ldots, x_n$. More precisely, the components of $x^{\le d}$ are indexed by all subsets $I \subset [n] = \{1, 2, \ldots, n\}$ that have at most d elements, with:*

$$(x^{\le d})_I = \prod_{i \in I} x_i.$$

The product over an empty set is defined to be 1. The d-th power of a set $S \subset \mathbb{R}^n$ is the set $S^{\le d} \subset \mathbb{R}^{\binom{n}{\le d}}$ defined as

$$S^{\le d} = \left\{ x^{\le d} : \; x \in S \right\}.$$

For example, for $n = 3$ and $d = 2$ we have

$$(x_1, x_2, x_3)^{\le 2} = (1, x_1, x_2, x_3, x_1 x_2, x_1 x_3, x_2 x_3).$$

The notion of the d-th power of a vector is closely related to the notion of *tensor power* of a vector. Indeed, the d-th tensor power of $x \in \mathbb{R}^n$ can be identified with the vector $x^{\otimes d} \in \mathbb{R}^{n^d}$ indexed by *multisets* $I \subset [n]$ that have exactly d elements, possibly with repetitions, and where the coefficients are defined similarly:

$$(x^{\otimes d})_I = \prod_{i \in I} x_i.$$

Although not as ubiquitous as tensor power, the concept of tensor lift $x^{\le d}$ appears naturally, although often implicitly, in coding theory. The list of all vectors $x^{\le d}$ for $x \in \{0, 1\}^n$ be seen as the $\binom{n}{\le d} \times 2^n$ evaluation matrix, or a *truth table*, of all Boolean monomials of degree d [1]; the rows of this matrix can be identified with Walsh functions that arise in Fourier analysis of Boolean functions [67].

The notion of d-th power allows us to establish a canonical correspondence between affine polynomials $p(x)$ in n variables of degree at most d with homogeneous monomials and homogeneous linear functions of $x^{\le d}$:

$$p(x) = \sum_{I : |I| \le d} a_I \prod_{i \in I} x_i = \langle a, x^{\le d} \rangle, \quad \text{where} \quad x \in \mathbb{R}^n, \; a \in \mathbb{R}^{\binom{n}{\le d}}.$$

This correspondence give the following Lemma.

Lemma 3.2 (Linearizing polynomial threshold functions)**.** *Let $S \subset \{-1, 1\}^n$. There are as many different polynomial threshold functions of degree d on S as there are homogeneous linear threshold functions on $S^{\le d}$.*

This simple fact and the link to hyperplane arrangements described in Section 2 leads to the following upper known bound on $T(n,d)$, which is implicit in [9]; see [6, Theorem 4.7].

Theorem 3.3 (Upper bound)**.** *For any $1 \le d \le n$, we have:*
$$T(n,d) \le 2\binom{2^n-1}{\le m-1} \quad \text{where} \quad m = \binom{n}{\le d}.$$

Proof. By Lemma 3.2, the number of Boolean polynomial threshold functions $T(n,d)$ equals the number of homogeneous threshold functions on $(\{-1,1\}^n)^{\le d}$. By Lemma 2.1, this is the same as the number of regions of the arrangement of hyperplanes $(x^{\le d})^\perp$, $x \in \{-1,1\}^n$. These are 2^n central hyperplanes in $\mathbb{R}^m$, where $m = \binom{n}{\le d}$. Thus part 4 of Lemma 2.2 yields:
$$T(n,d) \le r(2^n, m-1) \le 2\binom{2^n-1}{\le m-1},$$
as claimed. □

By simplifying this bound, we can prove the upper bound of Theorem 1.1:

Proof of the upper bound in Theorem 1.1. Due to Theorem 3.3, it is enough to check that
$$B(n,m) \coloneqq 2\binom{2^n-1}{\le m-1} \le 2^{nm}$$
for all integers $n \ge 1$ and $m \ge 2$.

If $m \ge 4$, then an elementary bound on the binomial coefficients (Lemma A.1 in the Appendix) gives
$$B(n,m) \le 2\Big(\frac{e2^n}{m-1}\Big)^{m-1} \le 2 \cdot 2^{n(m-1)} \le 2^{nm}.$$
Here we used that $m-1 \ge 3 \ge e$.

The two remaining cases to check are $m=2$ and $m=3$. If $m=2$, then, setting $N \coloneqq 2^n$, we get
$$B(n,m) = 2\binom{N-1}{\le 1} = 2N \le N^2 = 2^{nm},$$
since $N = 2^n \ge 2$. Finally, if $m=3$, then we similarly have
$$B(n,m) = 2\binom{N-1}{\le 2} = N^2 - N + 2 \le N^2 \le N^3 = 2^{nm}.$$
Theorem 1.1 is proved. □

4. Random tensors

4.1. **General position.** It remains to prove the lower bound in Theorem 1.1. Let us try to reverse the argument that we gave for the upper bound in the proof of Theorem 3.3. There we noted that $T(n,d)$ is the same as the number of regions of the hyperplane arrangement $x^\perp$, $x \in \mathcal{X}$, where
$$\mathcal{X} \coloneqq (\{-1,1\}^n)^{\le d} \subset \mathbb{R}^{\binom{n}{\le d}}.$$
Suppose for a moment that the points in $\mathcal{X}$ are in general position (in reality they are not). Then every subset of $\mathcal{X}$ consisting of $m < \binom{n}{\le d}$ points would span a different subspace in $\mathbb{R}^{\binom{n}{\le d}}$. The orthogonal complements of these different subspaces are different, too. These complements are intersections of some hyperplanes from our arrangement $x^\perp$, $x \in \mathcal{X}$; we called them *intersection subspaces* in Lemma 2.2. According to part 2 of this lemma, the

number of regions of our hyperplane arrangement is bounded below by the number of all intersection subspaces, which is at least as many as there are m-element subsets of $\mathcal{X}$. This gives:

$$T(n,d) \geq \binom{|\mathcal{X}|}{m} = \binom{2^n}{\binom{n}{\leq d} - 1}$$

if we choose $m = \binom{n}{\leq d} - 1$. This would be an almost matching lower lower bound for the upper bound in Theorem 3.3.

4.2. **Linear independence.** The problem with this heuristic argument is that the points in the d-th power of the Boolean hypercube $(\{-1,1\}^n)^{\leq d}$, and even in the Boolean hypercube $\{-1,1\}^n$ itself, are very far from being in general position. For example, the affine hyperplane spanned by a $(n-1)$-dimensional face of the Boolean cube contains 2^{n-1} points. Nevertheless, we might be able to say that most subsets of points are in the general position. This is where probabilistic reasoning becomes useful, allowing us to interpret "most" as *random*.

The first result in this direction was recently proved by E. Abbe, A. Shpilka, and A. Wigderson [1, Theorem 4.5] in the context of their study of Reed-Solomon codes. Their result states that a random subset $(\{-1,1\}^n)^{\leq d}$ is linearly independent with high probability:

Theorem 4.1 (Linear independence [1])**.** *Let n, d, m be positive integers such that*

$$m < \binom{n - \log\binom{n}{\leq d} - t}{\leq d}.$$

Consider independent random vectors $x_1, \ldots, x_m$ uniformly distributed in $\{-1,1\}^n$. Then, with probability larger than $1 - 2^{-t}$, the random vectors $x_1^{\leq d}, \ldots, x_m^{\leq d} \in \mathbb{R}^{\binom{n}{\leq d}}$ are linearly independent.

The special case of Theorem 4.1 obtained with $d = 1$ states that a $n \times m$ random matrix with columns x_k has full rank with high probability, if $m < n - C\log n$. This statement is not difficult to prove, and much more precise results are known about the singularity of random matrices [41, 90, 26, 91, 75, 95, 76, 77, 2, 78, 93, 20, 79, 62, 64, 100, 81, 36, 15, 96, 97, 54, 24, 98].

The original version of Theorem 4.1 from [1] guaranteed linear independence over the finite field $\mathbb{F}_2$, which is stronger than the linear independence over $\mathbb{R}$.[3] Moreover, in the original version of this theorem, the coordinates of the vectors $x_1, \ldots, x_m$ were uniformly distributed in $\{0,1\}$ rather than in $\{-1,1\}$, but the proof can be easily adapted to random $\{-1,1\}$ valued random variables.[4]

4.3. **Resilience.** However, linear independence of random vectors $x_1^{\leq d}, \ldots, x_m^{\leq d}$ is still too weak for our purposes. What we really need is to be able to show that the span of random vectors $x_1^{\leq d}, \ldots, x_m^{\leq d}$ does not contain any vector of the form $u^{\leq d}$, $u \in \{-1,1\}^n$, that is different from all the $\pm x_i^d$. This is a *resilience* property of linear independence, as it states

[3]If some Boolean vectors are linearly dependent over $\mathbb{R}$, then one can find a non-trivial linear combination that equals zero, and whose all coefficients are rational, and thus even integer. (This can be seen e.g. by performing the Gauss elimination.) Moreover, without loss of generality, not all of the coefficients are even: otherwise we can divide both sides by 2. Taking mod 2 of both sides of this equation, we obtain a linear dependence over $\mathbb{F}^2$.

[4]Most of the argument of [1] extends to $\{-1,1\}$ without any change. The only place that needs attention is Lemma 4.10 [1], which states that there exists a lot of linearly independent polynomials on any large subset of $\mathbb{F}_2^m$. Obviously, there is the same number of polynomials on any affine transformation of $\mathbb{F}_2^m$, in particular on $\{-1,1\}^m$.

that not only the set of vectors are independent, but independence holds even if we add any vector $u^{\le d}$, $u \in \{-1,1\}^n$, to the set. The following new result establishes the resilience of linear independence for random tensors:

Theorem 4.2 (Resilience of linear independence)**.** *Let n, d, m be positive integers such that $dn^{0.02} \le t \le 0.001n$ and*

$$m < \binom{n - \frac{Cn}{\log n} - t}{\le d}.$$

Consider independent random vectors $x_1, \dots, x_m$ uniformly distributed in $\{-1,1\}^n$. Then, with probability larger than $1 - 2^{-t/4}$, the span of the random vectors $x_1^{\le d}, \dots, x_m^{\le d}$ does not contain any other vector of the form $u^{\le d}$, $u \in \{-1,1\}^n$, that is different from $\pm x_i^{\le d}$.

This result is not an easy consequence of Theorem 4.1, and it requires additional probabilistic tools. The partial case of Theorem 4.2 where $d = 1$ was proved by Odlyzko [66] with a very sharp bound on the probability, which we do not need here. It was later noticed that for $d = 1$, the probabilities for linear independence and resilience are asymptotically equivalent [103]. Odlyzko's result was used in Zuev's argument [110] to prove (1.1). For tensors of any order $d > 1$, Theorem 4.2 is new. We derive it from Theorem 4.1 using an argument that is inspired by Odlyzko's method.

4.4. **Proof of the lower bound in Theorem 1.1.** Let us assume for a moment that the Resilience Theorem 4.2 is valid, and show how it yields the lower bound in our main result, Theorem 1.1.

Lemma 4.3 (Lots of unique subspaces)**.** *Let n, d, m be positive integers such that $1 \le d \le n^{0.9}$ and*

$$m \le \binom{n - \frac{Cn}{\log n}}{\le d}.$$

Then there exist at least $\frac{1}{2}\binom{2^n}{m}$ different subspaces of the form $\operatorname{span}(S^{\le d})$ *where S are subsets of $\{-1,1\}^n$ of cardinality m.*

Proof. Given a subset $S \subset \{-1,1\}^n$, let us call it *good* if $\operatorname{span}(S^{\le d})$ does not contain any vector of the form $u^{\le d}$, $u \in \{-1,1\}^n$, that is different from all $\pm x^{\le d}$, $x \in S$. Call S *bad* otherwise. Obviously, if S and T are two different good subsets, then $\operatorname{span}(S^{\le d})$ and $\operatorname{span}(T^{\le d})$ are two different subspaces. Thus, to complete the proof, it suffices to show that at least half of all m-element subsets S of the Boolean hypercube $\{-1,1\}^n$ are good.

Let us apply Theorem 4.2 with the value $t \coloneqq Cn/\log n$, which obviously satisfies the assumptions for sufficiently large n. The theorem implies that if S is a random set obtained by sampling m points from $\{-1,1\}^n$ *with replacement*, then:

$$\mathbb{P}\{S \text{ is bad}\} \le \frac{1}{4}. \tag{4.1}$$

The probability that there are no repetitions among these m points is:

$$\mathbb{P}\{\text{no repeat}\} = \prod_{i=1}^{m-1}\left(1 - \frac{i}{2^n}\right) \ge \left(1 - \frac{m-1}{2^n}\right)^{m-1} \ge 1 - \frac{(m-1)^2}{2^n} \ge \frac{1}{2}. \tag{4.2}$$

To check the last bound, recall that $d \le n^{0.9}$, so for sufficiently large n we get:

$$\log m \le \log\binom{n}{\le d} \le d\log(en) \le \frac{n-1}{2},$$

which implies $(m-1)^2 \le 2^{n-1}$ and yields the bound in (4.2).

Thus, the probability that a random m-element subset S, sampled *without replacement* from $\{-1,1\}^n$, is bad, is be bounded by:

$$\frac{\mathbb{P}\{S \text{ is bad}\}}{\mathbb{P}\{\text{no repeat}\}} \le \frac{1/4}{1/2} = \frac{1}{2}.$$

Hence, at most a half of the m-elements subsets of $\{-1,1\}^n$ are bad. The proof of the lemma is complete. □

Proof of the lower bound in Theorem 1.1. According to the Linearization Lemma 3.2, $T(n,d)$ is the number of linear threshold functions on the set

$$\mathcal{X} := (\{-1,1\}^n)^{\le d} \subset \mathbb{R}^{\binom{n}{\le d}}.$$

This number, according to Lemmas 2.1 and 2.2, is bounded below by the number of all intersection subspaces, which are the linear subspaces generated by intersecting various hyperplanes $z^\perp$, $z \in \mathcal{X}$. The orthogonal complement of each intersection subspace is the linear span of a subset of $\mathcal{X}$. Thus, the number of intersection subspaces equals the number of subspaces obtained as spans of subsets of $\mathcal{X}$. The number of spans can be bounded below using Lemma 4.3. This line of reasoning yields:

$$T(n,d) \ge \frac{1}{2}\binom{2^n}{m} \quad \text{where} \quad m = \binom{n - \frac{Cn}{\log n}}{\le d}.$$

It remains to simplify this bound. Taking logarithms of both sides and using a simple bound on binomial coefficients (Lemma A.1 in the Appendix), we get:

$$\log_2 T(n,d) \ge m(n - \log m) - 1 \ge m(n - 2\log m).$$

Another elementary bound on binomial sums (Lemma A.2 in the Appendix) gives

$$m \ge \Big(1 - \frac{2C}{\log n}\Big)^d \binom{n}{\le d},$$

and using Lemma A.1 again we see that

$$\log m < \log \binom{n}{\le d} \le d\log(en) \le \frac{Cn}{\log n}.$$

Thus

$$\log_2 T(n,d) > \Big(1 - \frac{2C}{\log n}\Big)^{d+1} n \binom{n}{\le d} \ge \Big(1 - \frac{4C}{\log n}\Big)^d n \binom{n}{\le d}.$$

This competes the proof of the main theorem. □

5. The Littlewood-Offord lemma

In this section and the next one, we prove the Resilience Theorem 4.2. Our argument is inspired by the proof of the partial case of this result for $d = 1$ due to Odlyzko [66]. The proof is based on the classical Littlewood-Offord lemma about *anti-concentration* of sums of independent random variables.

Let $\xi_1, \ldots, \xi_n$ be independent random variables and $a_1, \ldots, a_n \in \mathbb{R}$ be fixed coefficients. A classical question, which goes back to J. E. Littlewood and A. C. Offord [53] is to determine the probability that the sum of independent random variables $\sum a_k \xi_k$ hits a given number $u \in \mathbb{R}$. The first general result on this problem, now commonly known as the Littlewood-Offord Lemma, was proved by J. E. Littlewood and A. C. Offord [53] and sharpened by P. Erdös [32].

Lemma 5.1 (Littlewood-Offord Lemma [32])**.** *Let $\xi_1, \dots, \xi_n$ be independent, zero-mean, random variables taking values in $\{-1, 1\}$, and let $a_1, \dots, a_n$ be nonzero real numbers. Then, for every fixed $u \in \mathbb{R}$, we have*[5]

$$\mathbb{P}\left\{\sum_{k=1}^{n} a_k \xi_k = u\right\} \le 2^{-n}\binom{n}{\lfloor n/2 \rfloor} =: P(n).$$

A slightly more general version of Lemma 5.1, which bounds the probability that the sum falls in a given neighborhood of u, quickly follows from Sperner's theorem in combinatorics [32], see [19, Chapter 4].

Note that the probability bound in the Littlewood-Offord lemma is sharp: it reduces to an equality if all coefficients a_k are the same and $u = 0$. For many other vectors of coefficient $a = (a_1, \dots, a_n)$, one can obtain better bounds depending on the arithmetic structure of a. Such bounds have been extensively studied in connection to number theory, combinatorics and, more recently, random matrix theory; see, for instance, [32, 84, 37, 34, 95, 76, 78, 94, 65, 100, 63, 25, 58, 82], and the surveys [92, 79].

Using Stirling's approximation to estimate the binomial coefficient, we can derive the following, less precise but simpler, bound on the probability in the Littlewood-Offord Lemma.

Lemma 5.2 (Probability bounds in Littlewood-Offord Lemma)**.** *We have:*

$$P(n) \le \frac{C}{\sqrt{n}} \text{ for all } n \ge 1; \qquad P(n) \le \frac{3}{8} \text{ for all } n \ge 3.$$

Proof. The first bound follows from Stirling's formula. Furthermore, one can easily check that the numbers $P(n)$ form a non-increasing sequence and $P(3) = 3/8$. This gives the second bound. □

The bound in the Littlewood-Offord Lemma 5.1 can be slightly strengthened if $u \ne 0$. Although the following may seem like a small improvement, it can be critical for small values of n.

Lemma 5.3. *If $u \ne 0$ in the Littlewood-Offord Lemma 5.1, then*

$$\mathbb{P}\left\{\sum_{k=1}^{n} a_k \xi_k = u\right\} \le P(n+1).$$

Proof. Let ξ_{n+1} be a mean zero random variable taking values in $\{-1, 1\}$, and which is independent of $\xi_1, \dots, \xi_n$. Then

$$\begin{aligned}
\mathbb{P}\left\{\sum_{k=1}^{n} a_k \xi_k = u\right\} &= \mathbb{P}\left\{\sum_{k=1}^{n} a_k \xi_k = u\xi_{n+1}\right\} \quad \text{(by symmetry)} \\
&= \mathbb{P}\left\{\sum_{k=1}^{n+1} a_k \xi_k = 0\right\} \quad (\text{where } a_{n+1} := -u) \\
&\le P(n+1).
\end{aligned}$$

The proof is complete. □

[5] Here $\lfloor m \rfloor$ denotes the floor of m, i.e. the largest integer that is less or equal to m.

6. Resilience

To prove Theorem 4.2, we have to show that it is unlikely that there exists a vector $u \in \{-1,1\}^n$, and coefficients $a_1, \ldots, a_m \in \mathbb{R}$ at least two of which are non-zero, such that:

$$\sum_{k=1}^{m} a_k x_k^{\leq d} = u^{\leq d}. \tag{6.1}$$

Our argument will be a little different depending on how many coefficients a_k are nonzero. We will first analyze the case of "long combinations" where at least $n^{0.01}$ coefficients are nonzero, and then the remaining case of "short combinations".

6.1. Long combinations.

Let P_{long} denote the probability that there exists a vector $u \in \{-1,1\}^n$ and coefficients $a_1, \ldots, a_m \in \mathbb{R}$ at least $n^{0.01}$ of which are nonzero, and such that (6.1) holds. Our goal is to bound P_{long}.

Step 1. Extracting two batches of equations. We can view (6.1) as a system of $\binom{n}{\leq d}$ linear equations in variables $a_1, \ldots, a_m$, and we can write it as

$$\sum_{k=1}^{m} a_k \big(x_k^{\leq d}\big)_I = \big(u^{\leq d}\big)_I, \quad \text{or equivalently as} \quad \sum_{k=1}^{m} a_k \prod_{i \in I} x_{ki} = \prod_{i \in I} u_i,$$

for each subset $I \subset [n]$ with at most d elements. We will consider two subsets, or "batches", of these equations. The first batch will be used to determine the coefficients (a_k), and the second batch will be used to bound the probability.

Fix an integer $1 \leq n_0 \leq n$ whose value will be determined later. The first batch will be defined by the subsets:

$$I \subset \binom{[n_0]}{\leq d}$$

and the second batch, by the subsets:

$$I \subset \binom{n_0+1, \ldots, n}{1}.$$

Thus, $\sum_{k=1}^{m} a_k \prod_{i \in I} x_{ki}$ is a polynomial in the first n_0 variables $x_1, \ldots, x_{n_0}$ in the first batch, and we have a linear form in the remaining variables $x_{n_0+1}, \ldots, x_n$ in the second batch.

This gives us two systems of stochastically independent equations. We can rewrite them as follows. The first batch is given by:

$$\sum_{k=1}^{m} a_k \bar{x}_k^{\leq d} = \bar{u}^{\leq d} \tag{6.2}$$

where the vector $\bar{u} \in \{-1,1\}^{n_0}$ is obtained from $u \in \{-1,1\}^n$ by keeping only the first n_0 coefficients, and similarly for the vectors $\bar{x}_k \in \{-1,1\}^{n_0}$. The second batch is given by:

$$\sum_{k=1}^{m} a_k x_{ki} = u_i, \; n_0 < i \leq n.$$

Step 2: The first batch determines the coefficients, the second batch bounds the probability. Suppose that n_0 is chosen so that:

$$m < \binom{n_0 - \log \binom{n_0}{\leq d} - t}{\leq d}. \tag{6.3}$$

Then, by Theorem 4.1, the random vectors $\bar{x}_1^{\le d}, \dots, \bar{x}_m^{\le d}$ are linearly independent with probability larger than $1 - 2^{-t}$. In this case, the first batch of equations has full rank. Let us condition on a realization of $\bar{x}_1, \dots, \bar{x}_m$ for which the linear independence does hold.

Suppose the event P_{long} occurs. The vector $u \in \{-1,1\}^n$ in (6.1) can be chosen in 2^n ways; let us fix it. Since $\bar{x}_k^{\le d}$ and $\bar{u}^{\le d}$ are fixed at this point, linear independence implies that the coefficients (a_k) are uniquely determined by the first batch of equations. Thus the coefficients (a_k) are now fixed, too. Put them in the second batch of equations, which is stochastically independent from the first. This reasoning gives

$$P_{\text{long}} \le 2^{-t} + 2^n \cdot \max_{a,u} \mathbb{P}\left\{ \sum_{k=1}^m a_k x_{ki} = u_i, \; n_0 < i \le n \right\} \tag{6.4}$$

where the maximum is over all vectors $a = (a_k)$ with at least $n^{0.01}$ nonzero coefficients, and over all vectors $u = (u_i)$ with ± 1 coefficients.

Step 3: Applying the Littlewood-Offord Lemma. The Littlewood-Offord Lemma (Lemma 5.1 and Lemma 5.2) can help us bound the probability of each equation in (6.4). Indeed, since at least $n^{0.01}$ coefficients a_k are nonzero, we get

$$\mathbb{P}\left\{ \sum_{k=1}^m a_k x_{ki} = u_i \right\} \le P(n^{0.01}) \le \frac{C'}{n^{1/8}}$$

for each i. Since all $n - n_0$ such equations in (6.4) are stochastically independent, this implies

$$P_{\text{long}} \le 2^{-t} + 2^n \cdot \left(\frac{C'}{n^{1/8}} \right)^{n-n_0}.$$

Now is a good time to select a value for n_0. Let us set:

$$n_0 = n - \frac{C''n}{\log n} \tag{6.5}$$

where the absolute constant C'' is large enough. This choice allows us to have: $2^n \cdot (C''/n^{1/8})^{n-n_0} \le 2^n \cdot 2^{-2n} = 2^{-n}$. Furthermore, since $t \le n$ by assumption, $2^{-n} \le 2^{-t}$, and we obtain the bound:

$$P_{\text{long}} \le 2^{1-t}.$$

Step 4: Checking the bound on m. It remains to check that our choice of n_0 also makes our bound (6.3) valid. An elementary bound on binomial sums (Lemma A.1 in the Appendix) yields:

$$\log \binom{n}{\le d} \le d \log(en_0) \le d \log(en) \le \frac{n}{\log n}$$

by the assumption on d. Here we also used that n can be assumed to be sufficiently large (larger than any given absolute constant); otherwise the assumption on m in Theorem 4.2 is vacuous if the absolute constant C is chosen large enough. Similarly, m and t can be assumed to be sufficiently large. We will repeatedly use this in the rest of the argument.

This and our choice (6.5) of n_0 give:

$$\binom{n_0 - \log \binom{n_0}{\le d} - t}{\le d} \ge \binom{n - \frac{(C''+1)n}{\log n} - t}{\le d} > m$$

where the last inequality is the theorem's assumption on m with $C = C'' + 1$. We have checked that (6.3) is valid, and thus completed the analysis of the long combinations.

6.2. **Short combinations.** Let P_{short} denote the probability that there exists a vector $u \in \{-1,1\}^n$, and coefficients $a_1, \ldots, a_m \in \mathbb{R}$ at least two and at most $n^{0.01}$ of which are nonzero, such that (6.1) holds. Our goal is to bound P_{short}.

Step 1: Fixing the pattern of non-zero coefficients. Let us first address a simpler problem where the pattern of non-zero coefficients a_k is fixed. Namely, let us require that the non-zero coefficients a_k be exactly the first m_0 ones, where

$$m_0 \in [2, n^{0.01}]$$

is a fixed integer. Denote the probability in this simplified problem by P_{m_0}. Thus, P_{m_0} is the probability that there exists a vector $u \in \{-1,1\}^n$ and nonzero coefficients $a_1, \ldots, a_{m_0} \in \mathbb{R}$, satisfying:

$$\sum_{k=1}^{m_0} a_k x_k^{\leq d} = u^{\leq d}. \tag{6.6}$$

Step 2: Two batches of equations. We define and analyze two batches of equations in the same way as in our analysis of long combinations in Section 6.1, except that the value of n_0 in this case will be chosen differently later in the proof. Indeed, if:

$$m_0 < \binom{n_0 - \log\binom{n_0}{\leq d} - t}{\leq d}, \tag{6.7}$$

repeating the argument from Section 6.1 we get:

$$P_{m_0} \leq 2^{-t} + 2^n \cdot \max_{a,u} \mathbb{P}\left\{\sum_{k=1}^{m} a_k x_{ki} = u_i,\ n_0 < i \leq n\right\} \tag{6.8}$$

where the maximum is over all vectors $a = (a_k)$ with all nonzero coefficients, and over all vectors $u = (u_i)$ with ± 1 coefficients.

Step 3: Applying the Littlewood-Offord Lemma. The Littlewood-Offord Lemma (Lemma 5.3 and Lemma 5.2) can help us bound each probability term in (6.4). Indeed, since all a_k and u_i are nonzero, we get for each i:

$$\mathbb{P}\left\{\sum_{k=1}^{m_0} a_k x_{ki} = u_i\right\} \leq P(m_0 + 1) \leq \frac{3}{8}.$$

Since all $n - n_0$ such equations in (6.8) are stochastically independent, this implies:

$$P_{m_0} \leq 2^{-t} + 2^n \left(\frac{3}{8}\right)^{n-n_0}.$$

Now is a good time to select a value for n_0. Let us set:

$$n_0 = 0.1n. \tag{6.9}$$

This choice guarantees that $2^n (3/8)^{n-n_0} \leq (0.82)^n$. Furthermore, since $t \leq n/4$ by assumption, we have $(0.82)^n \leq 2^{-t}$. Thus we conclude that:

$$P_{m_0} \leq 2^{1-t}.$$

Step 4: Unfixing the pattern of non-zero coefficients. In the beginning of the proof, we made a simplifying assumption that the support of the coefficient vector $a = (a_k)$ be the set $[m_0]$. The same argument holds if we replace $[m_0]$ by any other subset of $[m]$ of cardinality

m_0. Thus, taking the union bound over all $\binom{m}{m_0}$ ways of choosing the support of a, and over all $m_0 \in [2, n^{0.01}]$ allowed sizes of the support, we obtain:

$$P_{\text{short}} \le \sum_{m_0=2}^{n^{0.01}} \binom{m}{m_0} P_{m_0} \le \binom{m}{n^{0.01}} 2^{1-t}.$$

Recall that n, t and $m < \log \binom{n}{\le d}$ are sufficiently large, so using elementary bounds on the binomial coefficients (Lemma A.1 in the Appendix) we get:

$$\log \binom{m}{n^{0.01}} \le 2n^{0.01} d \log n \le t/2$$

by the assumption on t. This allows us to conclude that:

$$P_{\text{short}} \le 2^{t/2} 2^{1-t} = 2^{1-t/2} \le 2^{-t/3}.$$

Step 4: Checking the bound on m_0. It remains to be checked that our choice of n_0 is consistent with the bound (6.7). Note that using elementary bounds on binomial coefficients (Lemma A.1):

$$\log \binom{n_0}{\le d} \le d \log(en_0) \le d \log n \le \frac{0.01n}{3} = \frac{n_0}{3}$$

by our assumption on d. Similarly:

$$t \le 0.001n \le \frac{0.01n}{3} = \frac{n_0}{3}.$$

This yields:

$$\binom{n_0 - \log \binom{n_0}{\le d} - t}{\le d} > \binom{n_0/3}{\le d} \ge \frac{n_0}{3} \ge 0.01n \ge m_0$$

where the last inequality follows for large n from the constraint $m_0 \le n^{0.01}$. Thus we have checked that (6.3) is valid, completing the analysis of the case of short combinations.

Step 5: Conclusion of the proof of Theorem 4.2. Combining our results for long and short combinations, we obtain that the overall failure probability is at most

$$P_{\text{long}} + P_{\text{short}} \le 2^{1-t} + 2^{-t/3} \le 2^{-t/4}$$

for sufficiently large t. This concludes the proof of Theorem 4.2. □

7. Proof of Corollary 1.2

We will derive both lower and upper bounds from Theorem 1.1. Let us start from the lower bound, which is much simpler.

7.1. **Lower bound.** Note that

$$\binom{n}{\le d} > \binom{n}{d} = \frac{n(n-1)\cdots(n-d+1)}{d!} > \frac{(n-d)^d}{d!} = \Big(1 - \frac{d}{n}\Big)^d \cdot \frac{n^d}{d!}.$$

Then Theorem 1.1 yields

$$\log_2 T(n,d) > \Big[\Big(1 - \frac{C}{\log n}\Big)\Big(1 - \frac{d}{n}\Big)\Big]^d \cdot \frac{n^{d+1}}{d!} \ge \Big(1 - \frac{2C}{\log n}\Big)^d \cdot \frac{n^{d+1}}{d!},$$

The last inequality holds whenever $d/n \le C/\log n$; if C is sufficiently large, this always holds in the range $1 \le d \le n^{0.9}$. Rename $2C$ to C to complete the proof of the lower bound in Corollary 1.2.

7.2. **Upper bound: five cases.** As we are about to show, the upper bound in Corollary 1.2 holds in the entire range $n > 1$, $1 \le d \le n$. The argument will again follow from Theorem 1.1 together with fine estimates of binomial sums. The desired inequality is exact and holds even for small values of d and n. The argument is elementary but somewhat long since several different cases must be considered.

Our reasoning will be different depending on whether the numbers n and d are small or large, and also on the size of the ratio n/d. These distinctions break the argument into five cases, which are summarized in Figure 4 below. As we see from the figure, the critical ratio for n/d in the argument is given by:

$$\alpha = 3.0528. \tag{7.1}$$

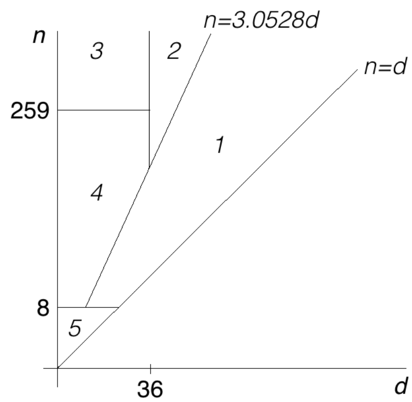

Figure 4. The proof has five cases, which are determined by the five regions the points (n, d) belong to. In region 1, the bound is trivial since it exceeds the number of all Boolean functions. In regions 2 and 3, the result is derived using fine approximations of binomial sums. In the finite regions 4 and 5, the bound is tested numerically by computer for all pairs (n, d) in those regions.

7.3. **Case 1:** $8 \le n \le \alpha d$**.** We are going to check that in this range:

$$\frac{n^{d+1}}{d!} > 2^n. \tag{7.2}$$

This would trivially yield the upper bound in Corollary 1.2, since $T(n, d)$ is bounded by the total number of Boolean functions 2^{2^n}. And this would yield:

$$\log_2 T(n, d) \le 2^n < \frac{n^{d+1}}{d!}.$$

In order to show (7.2), Stirling's approximation can be used to check that the bound:

$$d! \le (d/e)^d \sqrt{d}\, e \tag{7.3}$$

holds for all positive integers d, and so we have:

$$\frac{n^{d+1}}{d!} \ge \left(\frac{en}{d}\right)^d \frac{n}{e\sqrt{d}} > \left(\frac{en}{d}\right)^d.$$

In the last step, we used that $\frac{n}{e\sqrt{d}} > 1$ due to the constraints $n \geq 8$ and $d \leq n$. Thus, in order to complete the proof in this case, it suffices to show that

$$\left(\frac{en}{d}\right)^d \geq 2^n.$$

Taking logarithms on both sides, we see that this is the same as showing that

$$\log_2(ex) \geq x \quad \text{where } x = n/d.$$

A quick check shows that $\log_2(ex) \geq x$ for all values $x \in [1, \alpha]$ where $\alpha = 3.0528$ as in (7.1). It remains to note that the specific value $x = n/d$ falls into the interval $[1, \alpha]$: the lower bound holds since $d \leq n$, and the upper bound is exactly our assumption that $n \leq \alpha d$.

The proof in Case 1 is complete.

7.4. **Case 2:** $n > \alpha d$**,** $d \geq 36$**.** In this regime, we use Theorem 3.3. To simplify its conclusion, let us find a convenient bound on:

$$m = \binom{n}{\leq d}.$$

Lemma A.3 gives:

$$m \leq \frac{n^d}{d!} \cdot \frac{n+1-d}{n+1-2d}.$$

Furthermore, we have:

$$\frac{n+1-d}{n+1-2d} = 1 + \frac{d/n}{1+1/n-2d/n} \leq 1 + \frac{d/n}{1-2/\alpha}$$

where the last inequality follows since $1/n \geq 0$ and $d/n \leq 1/\alpha$. Thus:

$$m \leq \left\lfloor \frac{n^d}{d!}\left(1 + \frac{d/n}{1-2/\alpha}\right)\right\rfloor =: m_0. \tag{7.4}$$

Using Theorem 3.3, we obtain:

$$T(n,d) < 2\binom{2^n}{\leq m_0} \leq 2\left(\frac{e2^n}{m_0}\right)^{m_0},$$

where we used Lemma A.1 in the last step. Thus:

$$\log_2 T(n,d) < 1 + m_0\left(\log_2 e + n - \log_2 m_0\right). \tag{7.5}$$

To proceed, we need a convenient lower bound on $\log_2 m_0$. The definition of m_0 and Stirling's approximation (7.3) give:

$$m_0 \geq \frac{n^d}{d!} \geq \left(\frac{en}{d}\right)^d \frac{1}{\sqrt{d}\,e},$$

so:

$$\begin{aligned}\log_2 m_0 &\geq d\log_2\left(\frac{en}{d}\right) - \frac{1}{2}\log_2 d - \log_2 e \\ &\geq d\log_2(e\alpha) - \frac{1}{2}\log_2 d - \log_2 e \quad \text{(since } n > \alpha d \text{ by assumption)}.\end{aligned}$$

Substituting this lower bound and the definition (7.4) of m_0 into (7.5), yields:

$$\log_2 T(n,d) < 1 + \frac{n^d}{d!}\left(1 + \frac{d/n}{1-2/\alpha}\right)(n - f(d))$$

where

$$f(d) := d\log_2(e\alpha) - \frac{1}{2}\log_2 d - 2\log_2 e.$$

Redistributing the powers of n, this is the same as:

$$\log_2 T(n,d) \le 1 + \frac{n^{d+1}}{d!}\left(1 + \frac{d/n}{1-2/\alpha}\right)\left(1 - \frac{f(d)}{n}\right).$$

We claim that:

$$f(d) \ge \frac{d}{1-2/\alpha} \quad \text{for all } d \ge 36. \tag{7.6}$$

By definition of $f(d)$, we can write this desired inequality as:

$$\left(\log_2(e\alpha) - \frac{1}{1-2/\alpha}\right)d - \frac{1}{2}\log_2 d - 2\log_2 e > 0. \tag{7.7}$$

Since $\alpha = 3.0528$ by (7.1) and since $2\log_2 e \approx 2.88539$, this inequality is (slightly) weaker than:

$$0.1528d - 0.5\log_2 d - 2.8854 \ge 0.$$

A quick verification shows that this inequality holds for all real values $d \ge 36$, so the claim (7.6) is true.

Using (7.6) we obtain:

$$\begin{aligned}
\log_2 T(n,d) &< 1 + \frac{n^{d+1}}{d!}\left(1 + \frac{d/n}{1-2/\alpha}\right)\left(1 - \frac{d/n}{1-2/\alpha}\right) \\
&= 1 + \frac{n^{d+1}}{d!}\left(1 - \left(\frac{d/n}{1-2/\alpha}\right)^2\right) \\
&= \frac{n^{d+1}}{d!} + 1 - \frac{n^{d+1}}{d!}\left(\frac{d/n}{1-2/\alpha}\right)^2.
\end{aligned}$$

Now, by (7.1) $0 < 2/\alpha < 1$, so $(1-2/\alpha)^2 < 1$. Thus:

$$\frac{n^{d+1}}{d!}\left(\frac{d/n}{1-2/\alpha}\right)^2 > \frac{n^{d+1}}{d!}\left(\frac{d}{n}\right)^2 = \frac{n^{d-1}}{(d-1)!}\cdot d \ge 1$$

since $n \ge d \ge 1$. This shows that:

$$\log_2 T(n,d) < \frac{n^{d+1}}{d!},$$

which completes the proof in Case 2.

7.5. **Case 3: $n \ge 259$, $d \le 35$.** In this regime:

$$\frac{n}{d} \ge \frac{259}{35} = 7.4.$$

Thus, we may repeat the argument of Case 2 for the larger, and thus better, value $\alpha = 7.4$. With this value of α, the inequality (7.7) is (slightly) weaker than:

$$2.9598d - 0.5\log_2 d - 2.8854 \ge 0.$$

It is easy to check numerically by computer that this inequality holds for all real values $d \ge 1$. The rest of the argument is identical to that of Case 2.

7.6. **Case 4: $8 \le n \le 258$, $d \le 35$, $n > \alpha d$.** Theorem 3.3 and Lemma A.1 on the binomial sum yield:

$$T(n,d) < 2\binom{2^n-1}{\le m} \le 2\left(\frac{e(2^n-1)}{m}\right)^m.$$

Taking logarithms on both sides gives:

$$\log_2 T(n,d) < 1 + m\log_2\left(\frac{e(2^n-1)}{m}\right) =: t(n,d).$$

It is easy to check numerically by computer that:

$$t(n,d) \le \frac{n^{d+1}}{d!}$$

for all pairs (n,d) in the current range.

7.7. **Case 5: $1 < n \le 7$, $1 \le d \le n$.** This is similar to Case 4, except we will not use any bounds on the binomial sum. As before:

$$T(n,d) < 2\binom{2^n-1}{\le m}$$

and it is easy to check numerically by computer that:

$$\log_2\left[2\binom{2^n-1}{\le m}\right] \le \frac{n^{d+1}}{d!}$$

for all pairs (n,d) in the current range.

The proof of Corollary 1.2 is complete. □

8. Further questions

The results and especially the methods of this paper lead to a number of interesting directions for further study.

8.1. **Polynomial capacity of sets.** Given a finite set S of points in $\mathbb{R}^n$, we can define the capacity $C_d(S)$ to be the base two logarithm of the number of different ways S can be split by polynomials of degree d. We can derive some bounds on the polynomial capacity of sets.

Theorem 8.1 (Polynomial set capacity)**.** *Consider a finite subset $S \subset \mathbb{R}^n$, where $n > 1$. Then, for any degree $1 < d \le n$, we have:*

$$C_d(S) \le 1 + \log_2\binom{|S|-1}{\le m-1} \le m\log_2|S|,$$

where:

$$m = \binom{n+k-1}{\le d} \le \left(\frac{2en}{d}\right)^d.$$

Proof. First, it is easy to see that the number of coefficients of a polynomial of degree d in n variables $x_1, \dots, x_n$ is given by $m = \binom{n+k-1}{\le d}$, including the constant term (bias). A vector $x \in \mathbb{R}^n$ can be canonically and injectively mapped into a vector $f(x) \in \mathbb{R}^{m-1}$ whose components are the various monomials. Using this mapping, we can represent any polynomial $p(x)$ of degree d over $\mathbb{R}^n$ as a linear affine function over $f(x)$. And vice versa, any linear affine function over $f(x)$ is a polynomial of degree d over x. For example, if $d = 2$, the vector $x = (x_1, x_2) \in \mathbb{R}^2$ is canonically mapped to the vector $f(x) = (x_1, x_2, x_1x_2, x_1^2, x_2^2) \in \mathbb{R}^5$. Any polynomial $p(x) = a_0 + a_1x_1 + a_2x_2 + a_{12}x_1x_2 + a_{11}x_1^2 + a_{22}x_2^2$ over $\mathbb{R}^2$ is clearly an affine function of $f(x)$, and vice versa. Therefore:

$$C_d(S) = C_1(f(S)) = C(f(S)).$$

We complete the proof by applying the known bounds on the capacity of sets with respect to linear threshold functions (see [14]) to the set $f(S) \subset \mathbb{R}^{m-1}$, noting that $f(S)$ has the same cardinality as S since f is injective. □

Note that if we apply Theorem 8.1 to the hypercube $S = H^n$, we get:

$$C_d(H^n) \le n\Big(\frac{2en}{d}\Big)^d,$$

which is somewhat weaker asymptotically than the result in [11] giving:

$$C_d(H^n) = C_d(n,1) = \frac{n^{d+1}}{d!}(1+o(1)).$$

Note also that the general lower bound: $1 + \log_2|S| \le C_d(S)$, and its improved version when S is a subset of the Boolean cube ([14]: $\log_2^2|S|/16 \le C_d(S)$ are trivially satisfied. Thus an open research area here is to obtain better estimates of the polynomial capacity of finite sets of points, including subsets of the hypercube.

8.2. **Polynomial threshold functions with high degrees.** In this paper we determined the asymptotic behavior of $T(n,d)$, the number of n-variable polynomial threshold functions with bounded or slowly growing degrees d. It would be interesting to find out what happens if the degree d grows rapidly, for example linearly with n. It is plausible that the upper bound on $T(n,d)$ that we stated in Theorem 3.3 may be tight, and the following conjecture mentioned by M. Anthony [5] could hold:

Conjecture 8.2. *The number $T(n,d)$ of n-variable polynomial threshold functions of degree d satisfy*

$$T(n,d) = (2-o(1))\binom{2^n-1}{\le m-1} \quad \text{where} \quad m = \binom{n}{\le d} \tag{8.1}$$

for all degrees $1 \le d \le n$ as $n \to \infty$.

This conjecture might be too strong and it possibly holds only after we take logarithms on both sides of (8.1).

For bounded or mildly growing degrees d, Conjecture 8.2 easily implies the main result of this paper, Theorem 1.1. For $d = n/2$, Conjecture 8.2 and a careful asymptotic analysis of the bound (8.2) implies the Wang-Williams conjecture mentioned in the introduction, which states that most Boolean functions can be expressed as polynomial threshold functions of degree $n/2$ (see [5]). Finally, for $d = n$, Conjecture 8.2 is trivial. In this case it gives $T(n,n) = 2^{2^n}$, which is equivalent to the fact that all Boolean functions are polynomial threshold functions of degree at most n.

8.3. **Polynomial threshold functions with restricted coefficients.** In some applications (e.g. discrete synapses in neural networks), it is useful to consider polynomial threshold functions $f(x) = \operatorname{sgn}(p(x))$ where $p(x)$ is required to have bounded, discrete, or positive coefficients.

8.3.1. *Integer coefficients.* By an easy perturbation argument, we can always force $p(x)$ to have integer coefficients. How large are these coefficients? If $d = 1$, i.e. in the case of linear threshold functions, all coefficients of $p(x)$ are bounded by $n^{n/2+o(n)}$. This bound is tight due to results of J. Håstad [38] and N. Alon and V. Vu [4]. For any higher degree $d \geq 2$, a similar result is described in [73].

Beyond these results, it may be natural to look for a bound on the coefficients of $p(x)$ that holds for *most* (or many) polynomial threshold functions $f(x) = \mathrm{sgn}(p(x))$, e.g. for $2^{(1+o(1))n^{d+1}/d!}$ of them. What is then the optimal bound? Is it significantly smaller than $n^{n/2+o(n)}$, the bound that holds for *all* functions $f(x)$? This question seems to be open for all degrees including $d = 1$.

A related problem is when we require the integer coefficients of $p(x)$ to be bounded by a given number M. How many polynomial threshold functions $f(x) = \mathrm{sgn}(p(x))$ can be generated with this restriction? What if we consider polynomials $p(x)$ with all ± 1 coefficients? Since each such polynomial consists of $\binom{n}{\leq d}$ monomial terms and each term is assigned an ± 1 coefficient, there are at most $2^{\binom{n}{\leq d}}$ polynomials with ± 1 coefficients. Thus the number of corresponding polynomial threshold functions is bounded by $2^{\binom{n}{\leq d}}$. Is this bound asymptotically tight? It is easy to check that the answer is positive for $d = 1$, but for higher degrees the problem is non-trivial.

8.3.2. *Positive coefficients.* In some other situations (e.g. excitatory neurons in neural networks), it is natural to consider polynomial threshold functions $f(x) = \mathrm{sgn}(p(x))$ where the polynomials $p(x)$ have positive coefficients. How many polynomial threshold functions can be generated with this restriction?

For $d = 1$, one can answer this question easily by leveraging the symmetry of the Boolean cube $\{-1, 1\}^n$ with respect to signs. Due to this symmetry, the number of homogeneous linear threshold functions $f(x) = \mathrm{sgn}(a_1 x_1 + \cdots a_n x_n)$ whose coefficients a_k follow a given sign pattern is the same for each pattern. It follows that

$$\bar{T}^+(n, 1) = \frac{\bar{T}(n, 1)}{2^n}$$

where $\bar{T}(n, 1)$ denotes the number of homogeneous linear threshold functions, and $\bar{T}^+(n, 1)$ denotes the number of such functions with positive coefficients. Since $\log_2 \bar{T}(n, 1) = n^2 - o(n^2)$, we get

$$\log_2 \bar{T}^+(n, 1) = n^2 - o(n^2) - n = n^2 - o(n^2).$$

However, for higher degrees $d \geq 2$, the symmetry argument fails and the problem remains open.

8.4. **Connections to information theory.** The main result of this paper, Theorem 1.1, implies that we need essentially $n^{d+1}/d!$ bits to communicate a polynomial threshold function of degree d. This can be viewed as $n^d/d!$ binary vectors of dimension n and can intuitively be understood as communicating $n^d/d!$ *support* vectors, that is the $n^d/d!$ vectors of the Boolean hypercube that are closest to and on one side of the corresponding separating surface $p(x) = 0$. Thus in this case, the set of vectors depends on the function and thus it needs to be be communicated. However we may consider fixing a set of vectors in advance – one set for any function being communicated. In this scheme, we need to send only the value of the function $f(x)$ on this set. How well will such a scheme work? How large does the set of vectors needs to be for exact or approximate communication of any (or most) polynomial threshold functions of a given degree d?

8.5. **Boolean networks.** In neural networks and other applications, one is interested in the behavior of entire networks (or circuits) of polynomial threshold functions, rather than single polynomial threshold functions. For a given circuit, one would like to estimate the number of different Boolean functions that can be realized using different weights. This question has seemed hopeless for a long time but we believe the results presented here can be used to make some progress, at least in the case of particular circuits that are widely used in applications. Results in this direction are described in [13, 14] where we show how to estimate the number of functions that can be computed by fully connected networks, as well as shallow and deep layered feedforward networks of polynomial threshold gates.

8.6. **The geometry of boolean threshold functions.** As we noted in Section 2, homogeneous linear threshold functions correspond to the regions of the hyperplane arrangement $x^\perp$, $x \in \{-1,1\}^n$. These regions are polyhedral cones in $\mathbb{R}^n$, and to study their geometry it is convenient to intersect them with the unit Euclidean sphere. Thus we are looking at a decomposition of the sphere by 2^n central hyperplanes. From (1.1) we know that there are approximately 2^{n^2} regions in this decomposition. What else do we know about them? For example, what is the distribution of their area? We can of course ask the same questions for $d > 1$ as well.

These problems are related to the classical study of random Poission *tessellations* in stochastic geometry [23]; see also [72] for random tessellations on the sphere. However, the main new challenge here is to handle the discrete distribution induced by the Boolean cube.

Appendix A. Bounds on binomial sums

Throughout the main body of this paper, we repeatedly used the following elementary and well known bounds on binomial sums.

Lemma A.1 (see e.g. Exercise 0.0.5 in [101])**.** *For any integers $1 \le d \le n$, we have:*

$$\Big(\frac{n}{d}\Big)^d \le \binom{n}{d} \le \binom{n}{\le d} \le \Big(\frac{en}{d}\Big)^d.$$

Lemma A.2. *For any integers d and n such that $1 \le d \le n/2$ and for any integer $1 \le k \le n-d+1$, we have:*

$$\Big(1-\frac{2k}{n}\Big)^d \binom{n}{\le d} \le \binom{n-k}{\le d} \le \binom{n}{\le d}.$$

Proof. Let us first prove a similar but simpler fact for binomial coefficients:

$$\Big(1-\frac{2k}{n}\Big)^d \binom{n}{d} \le \binom{n-k}{d} \le \binom{n}{d} \tag{A.1}$$

The upper bound is non-trivial. To prove the lower bound, combine the terms in the definition of the binomial coefficients and get:

$$\frac{\binom{n-k}{d}}{\binom{n}{d}} = \prod_{i=0}^{d-1} \Big(1-\frac{k}{n-i}\Big) \ge \Big(1-\frac{k}{n-d+1}\Big)^d.$$

Since $n-d+1 \ge n/2$, the lower bound in (A.1) follows.

Next, we can deduce the conclusion of the lemma from (A.1) as follows:

$$\sum_{i=1}^d \binom{n-k}{i} \ge \sum_{i=1}^d \Big(1-\frac{2k}{n}\Big)^i \binom{n}{i} \ge \Big(1-\frac{2k}{n}\Big)^d \sum_{i=1}^d \binom{n}{i},$$

which is the lower bound we claimed. □

Lemma A.3. *For any integers $1 \le d \le n/2$, we have:*

$$\binom{n}{\le d} \le \binom{n}{d} \frac{n+1-d}{n+1-2d}.$$

Proof. Let us bound the ratio

$$\begin{aligned}
\frac{\binom{n}{\le d}}{\binom{n}{d}} &= \frac{\binom{n}{d} + \binom{n}{d-1} + \binom{n}{d-2} + \cdots}{\binom{n}{d}} \\
&= 1 + \frac{d}{n-d+1} + \left(\frac{d}{n-d+1}\right)\left(\frac{d-1}{n-d+2}\right) + \cdots \\
&\le 1 + \frac{d}{n-d+1} + \left(\frac{d}{n-d+1}\right)^2 + \cdots \\
&= \frac{n+1-d}{n+1-2d} \quad \text{(by summing the geometric series).}
\end{aligned}$$

This proves the lemma. □

References

[1] E. Abbe, A. Shpilka, and A. Wigderson, *ReedMuller codes for random erasures and errors,* IEEE Transactions on Information Theory 61 (2015), 5229–5252.

[2] R. Adamczak, O. Guédon, A. Litvak, A. Pajor, N. Tomczak-Jaegermann, *Smallest singular value of random matrices with independent columns,* C. R. Math. Acad. Sci. Paris 346 (2008), 853–856.

[3] J. Alman, T. Chan, R. Williams, *Polynomial representations of threshold functions and algorithmic applications,* 57th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2016).

[4] N. Alon, V. Vu, *Anti-Hadamard matrices, coin weighing, threshold gates, and indecomposable hypergraphs,* J. Combin. Theory Ser. A 79 (1997), 133–160.

[5] M. Anthony, *Classification by polynomial surfaces,* Discrete Applied Mathematics 61 (1995), 91–103.

[6] M. Anthony, *Discrete mathematics of neural networks. Selected topics.* SIAM Monographs on Discrete Mathematics and Applications. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2001.

[7] J. Aspnes, R. Beigel, M. Furst, S. Rudich, *The expressive power of voting polynomials,* Combinatorica, 14 (1994), 1–14.

[8] P. Baldi, *Symmetries and learning in neural network models,* Phys. Rev. Lett. 59 (1987), no. 17, 1976–1978.

[9] P. Baldi, *Neural networks, orientations of the hypercube, and algebraic threshold functions,* IEEE Trans. Inform. Theory 34 (1988), no. 3, 523–530.

[10] P. Baldi, *Group actions and learning for a family of automata,* J. Comput. System Sci. 36 (1988), no. 1, 1–15.

[11] Pierre Baldi and Roman Vershynin. *Boolean polynomial threshold functions and random tensors.* arXiv preprint arXiv:1803.10868, v1, 2018.

[12] P. Baldi. *Deep learning in biomedical data science.* Annual Review of Biomedical Data Science, 1 (1988),181–205.

[13] P. Baldi and R. Vershynin. *On neuronal capacity.* Advances in Neural Information Processing Systems, (2018), 7740–7749.

[14] P. Baldi and R. Vershynin. *The capacity of feedforward neural networks.* Neural Networks, 116, (2019), 288-311. See also arXiv preprint: arXiv:1901.00434, 2018.

[15] A. Basak, M. Rudelson, *Invertibility of sparse non-Hermitian matrices,* Adv. Math. 310 (2017), 426–483.

[16] R. Beigel, *The polynomial method in circuit complexity,* Proc. of 8th Annual Structure in Complexity Theory Conference (1993), 82–95.

[17] A. Bhattacharyya, S. Ghoshal, R. Saket, *Hardness of learning noisy halfspaces using polynomial thresholds,* preprint (2017).

[18] R. Beigel, N. Reingold, D. Spielman, *PP is closed under intersection,* Journal of Computer & System Sciences 50 (1995), 191–202.
[19] B. Bollobás, *Combinatorics. Set systems, hypergraphs, families of vectors and combinatorial probability.* Cambridge University Press, Cambridge, 1986.
[20] J. Bourgain, V. Vu, P. Wood, *On the singularity probability of discrete random matrices,* J. Funct. Anal. 258 (2010), 559–603.
[21] J. Bruck, *Harmonic analysis of polynomial threshold functions,* SIAM J. Discrete Math. 3 (1990), 168–177.
[22] R. C. Buck, *Partition of space,* Amer. Math. Monthly 50 (1943), 541–544.
[23] P. Calka, *Tessellations.* New perspectives in stochastic geometry, 145–169, Oxford Univ. Press, Oxford, 2010.
[24] N. Cook, *On the singularity of adjacency matrices for random regular digraphs,* Probab. Theory Related Fields 167 (2017), 143–200.
[25] K. Costello, *Bilinear and quadratic variants on the Littlewood-Offord problem,* Israel J. Math. 194 (2013), 359–394.
[26] K. Costello, T. Tao, V. Vu, *Random symmetric matrices are almost surely nonsingular,* Duke Math. J. 135 (2006), 395–413.
[27] P. Comon, G. Golub, L.-H. Lim, B. Mourrain, *Symmetric tensors and symmetric tensor rank,* SIAM J. Matrix Anal. Appl. 30 (2008), 1254–1279.
[28] T. Cover, *Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition,* IEEE Transactions on Electronic Computers 3 (1965), 326–334.
[29] V. de la Peña, E. Giné, *Decoupling: from dependence to independence.* Springer Verlag, 1999.
[30] I. Diakonikolas, R. O'Donnell, R. Servedio, Y. Wu, *Hardness results for agnostically learning low-degree polynomial threshold functions.* Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms, 1590–1606, SIAM, Philadelphia, PA, 2011.
[31] I. Diakonikolas, R. A. Servedio, L.-Y. Tan, A. Wan, *A regularity lemma and low-weight approximators for low-degree polynomial threshold functions,* Theory Comput. 10 (2014), 27–53.
[32] P. Erdös, *On a lemma of Littlewood and Offord,* Bull. Amer. Math. Soc. 51 (1945), 898–902.
[33] P. Erdös, *Extremal problems in number theory,* Proc. Sympos. Pure Math., vol.VIII, AMS, Providence, RI, 1965, pp.181–189.
[34] P. Frankl,Z. Füredi, *Solution of the Littlewood-Offord problem in high dimensions,* Ann. of Math. (2) 128 (1988), 259–270.
[35] F. Götze, *Asymptotic expansions for bivariate von Mises functionals,* Z. Wahrsch. Verw. Gebiete 50 (1979), 333–355.
[36] F. Götze, A. Naumov, A. Tikhomirov, *On minimal singular values of random matrices with correlated entries,* Random Matrices Theory Appl. 4 (2015), no. 2, 1550006.
[37] G. Halász, *Estimates for the concentration function of combinatorial number theory and probability,* Period. Math. Hungar. 8 (1977) 197–211.
[38] J. Håstad, *On the size of weights for threshold gates,* SIAM J. Discrete Math. 7 (1994), 484–492.
[39] A. A. Irmatov, *On the number of threshold functions,* Diskret. Mat. 5 (1993), 40–43; translation in Discrete Math. Appl. 3 (1993), 429–432.
[40] A. A. Irmatov, *Arrangements of hyperplanes and the number of threshold functions,* Acta Appl. Math. 68 (2001), 211–226.
[41] J. Kahn, J. Komlós, E. Szemerédi, *On the probability that a random* ± 1*-matrix is singular,* J. Amer. Math. Soc. 8 (1995), 223–240.
[42] N. Kalton, *Rademacher series and decoupling,* New York J. Math. 11 (2005), 563–595.
[43] D. Kane, *A structure theorem for poorly anticoncentrated polynomials of Gaussians and applications to the study of polynomial threshold functions,* Ann. Probab. 45 (2017), 1612–1679.
[44] Z. Kovijanić Vukićević, *An enumerative problem in threshold logic,* Publ. Inst. Math. (Beograd) (N.S.) 82(96) (2007), 129–134.
[45] R. Kannan, *Decoupling and partial independence,* Building bridges, 321–331, Bolyai Soc. Math. Stud., 19, Springer, Berlin, 2008.
[46] A. Klivans, R. O'Donnell, R. Servedio, *Learning intersections and thresholds of halfspaces.* In Proceedings of the 43rd Annual Symposium on Foundations of Computer Science (2002), 177–186.
[47] A. Klivans, R. Servedio, *Learning DNF in time* $2^{O(n^{1/3})}$, J. Computer and System Sciences 68 (2004), 303–318.
[48] T. Kolda, B. Bader, *Tensor decompositions and applications,* SIAM Rev. 51 (2009), no. 3, 455-?500.

[49] J. Komlós, *On the determinant of* $(0, 1)$ *matrices,* Studia Sci. Math. Hungar 2 (1967), 7–21.

[50] J. Komlós, *On the determinant of random matrices,* Studia Sci. Math. Hungar. 3 (1968), 387–399.

[51] M. Krause, P. Pudlak, *Computing boolean functions by polynomials and threshold circuits,* Computational Complexity 7 (1998), 346–370.

[52] M. Ledoux, *The concentration of measure phenomenon.* Mathematical Surveys and Monographs, 89. American Mathematical Society, Providence, RI, 2001.

[53] J. E. Littlewood, A. C. Offord, *On the number of real roots of a random algebraic equation. III,* Rec. Math. [Mat. Sbornik] N.S. 12 (1943), 277–286; in *Collected Papers of J. E. Littlewood*, Vol. 2, pp. 1333–1342, Oxford University Press, London, 1982.

[54] A. Litvak, A. Lytova, K. Tikhomirov, N. Tomczak-Jaegermann, P. Youssef, *Adjacency matrices of random digraphs: singularity and anti-concentration,* J. Math. Anal. Appl. 445 (2017), 1447–1491.

[55] J. Matoušek, *Lectures on discrete geometry.* Graduate Texts in Mathematics, 212. Springer-Verlag, New York, 2002.

[56] W. McCulloch, W. Pitts, *A logical calculus of the ideas immanent in nervous activity,* Bull. Math. Biophys. 5 (1943), 115–133.

[57] P. McCullagh, *Tensor methods in statistics.* Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1987.

[58] R. Meka, O. Nguyen, V. Vu, *Anti-concentration for polynomials of independent random variables,* Theory Comput. 12 (2016), Paper No. 11, 16 pp.

[59] M. Minsky, S. Papert, *Perceptrons: an introduction to computational geometry* (expanded edition); MIT Press, Cambridge, MA, 1988.

[60] S. Muroga, *Lower bounds of the number of threshold functions and a maximum weight,* IEEE Transactions on Electronic Computers 2 (1965), 136–148.

[61] E. Nering, *Linear Algebra and Matrix Theory.* Second edition. New York: Wiley, 1970.

[62] H. Nguyen, *On the least singular value of random symmetric matrices,* Electron. J. Probab. 17 (2012), no. 53, 19 pp.

[63] H. Nguyen, *Inverse Littlewood-Offord problems and the singularity of random symmetric matrices,* Duke Math. J. 161 (2012), 545–586.

[64] H. Nguyen, *On the singularity of random combinatorial matrices,* SIAM J. Discrete Math. 27 (2013), 447–458.

[65] H. Nguyen, V. Vu, *Optimal inverse Littlewood-Offord theorems,* Adv. Math. 226 (2011), 5298–5319.

[66] A. M. Odlyzko, *On subspaces spanned by random selections of* ± 1 *vectors,* Journal of Combinatorial Theory, Series A 47 (1988), 124–133.

[67] R. O'Donnell, *Analysis of Boolean functions.* Cambridge University Press, New York, 2014.

[68] R. O'Donnell, R. A. Servedio, *Extremal properties of polynomial threshold functions,* J. Comput. System Sci. 74 (2008), 298–312.

[69] R. O'Donnell, R. A. Servedio, *New degree bounds for polynomial threshold functions,* Combinatorica 30 (2010), 327–358.

[70] R. O'Donnell, Y. Zhao, *Polynomial bounds for decoupling, with applications,* 31st Conference on Computational Complexity, Art. No. 24, 18 pp., LIPIcs. Leibniz Int. Proc. Inform., 50, Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2016.

[71] P. C. Ojha, *Enumeration of linear threshold functions from the lattice of hyperplane intersections,* IEEE Trans. Neural Networks 11 (2000), 839–850.

[72] Y. Plan, R. Vershynin, *Dimension reduction by random hyperplane tessellations,* Discrete and Computational Geometry 51 (2014), 438–461.

[73] V.V. Podolskii, *Perceptrons of large weight,* Problems of Information Transmission 45, 1, (2009), 46–53.

[74] G.-C. Rota, *On the foundations of combinatorial theory. I. Theory of Mbius functions,* Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 2 (1964), 340–368.

[75] M. Rudelson, *Invertibility of random matrices: norm of the inverse,* Ann. of Math. (2) 168 (2008), 575–600.

[76] M. Rudelson, R. Vershynin, *The Littlewood-Offord Problem and invertibility of random matrices,* Advances in Mathematics 218 (2008), 600–633.

[77] M. Rudelson, R. Vershynin, *The least singular value of a random square matrix is* $O(n^{-1/2})$*,* C. R. Math. Acad. Sci. Paris 346 (2008), 893–896.

[78] M. Rudelson, R. Vershynin, *Smallest singular value of a random rectangular matrix,* Communications on Pure and Applied Mathematics 62 (2009), 1707–1739.

[79] M. Rudelson, R. Vershynin, *Non-asymptotic theory of random matrices: extreme singular values.* Proceedings of the International Congress of Mathematicians. Volume III, 1576–1602, Hindustan Book Agency, New Delhi, 2010.

[80] M. Rudelson, R. Vershynin, *Hanson-Wright inequality and sub-gaussian concentration,* Electronic Communications in Probability 18 (2013), 1–9.

[81] M. Rudelson, R. Vershynin, *Invertibility of random matrices: unitary and orthogonal perturbations,* Journal of the AMS 27 (2014), 293–338.

[82] M. Rudelson, R. Vershynin, *No-gaps delocalization for general random matrices,* Geometric and Functional Analysis 26 (2016), 1716–1776.

[83] M. Saks, *Slicing the hypercube.* Surveys in combinatorics, 1993 (Keele), 211–255, London Math. Soc. Lecture Note Ser., 187, Cambridge Univ. Press, Cambridge, 1993.

[84] A. Sárközy, E. Szemeredi, Über ein Problem von Erdös und Moser, Acta Arith. 11 (1965), 205–208.

[85] J. Schmidhuber, *Deep learning in neural networks: An overview,* Neural Networks 61 (2015), 85–117.

[86] L. Schläfli, *Gesammelte mathematische Abhandlungen.* Band I. (German) Verlag Birkhäuser, Basel, 1950.

[87] A. Sherstov, *Separating AC0 from depth-2 majority circuits,* SIAM J. Computing 38 (2009), 2113–2129.

[88] R. Stanley, *An introduction to hyperplane arrangements.* Geometric combinatorics, 389–496, IAS/Park City Math. Ser., 13, Amer. Math. Soc., Providence, RI, 2007.

[89] M. Talagrand, *A new look at independence,* Ann. Probab. 24 (1996), 1–34.

[90] T. Tao, V. Vu, *On random* ± 1 *matrices: singularity and determinant,* Random Structures and Algorithms 28 (2006), 1–23.

[91] T. Tao, V. Vu, *On the singularity probability of random Bernoulli matrices,* J. Amer. Math. Soc. 20 (2007), 603–628.

[92] T. Tao, V. Vu, *From the Littlewood-Offord problem to the circular law: universality of the spectral distribution of random matrices,* Bull. Amer. Math. Soc. (N.S.) 46 (2009), 377–396.

[93] T. Tao, V. Vu, *Random matrices: the distribution of the smallest singular values,* Geom. Funct. Anal. 20 (2010), 260–297.

[94] T. Tao, V. Vu, *A sharp inverse Littlewood-Offord theorem,* Random Structures Algorithms 37 (2010), 525–539.

[95] T. Tao, V. Vu, *Inverse Littlewood-Offord theorems and the condition number of random discrete matrices,* Annals of Math. 169 (2009), 595–632.

[96] K. Tikhomirov, *The limit of the smallest singular value of random matrices with i.i.d. entries,* Adv. Math. 284 (2015), 1–20.

[97] K. Tikhomirov, *The smallest singular value of random rectangular matrices with no moment assumptions on entries,* Israel J. Math. 212 (2016), 289–314.

[98] K. Tikhomirov, *Sample covariance matrices of heavy-tailed distributions,* Int. Math. Res. Notes, to appear.

[99] K. Tikhomirov, *Singularity of random Bernoulli matrices,* preprint.

[100] R. Vershynin, *Invertibility of symmetric random matrices,* Random Structures and Algorithms 44 (2014), 135–182.

[101] R. Vershynin, *High-dimensional probability. An introduction with applications in data science.* Cambridge University Press, 2017.

[102] R. Vershynin, *Invertibility of random tensors,* in preparation.

[103] T. Voigt, G. Ziegler, *Singular 0/1-matrices, and the hyperplanes spanned by random 0/1-vectors,* Combin. Probab. Comput. 15 (2006), no. 3, 463–471.

[104] C. Wang, A. Williams, *The threshold order of a boolean function,* Discrete Applied Mathematics, 31 (1991), 51–69.

[105] R. O. Winder. *Partitions of n-space by hyperplanes,* SIAM Journal on Applied Mathematics, 14 (1966), no. 4, 811–818.

[106] J. Wendel, *A problem in geometric probability,* Math. Scand. 11 (1962), 109–111.

[107] T. Zaslavsky, *Facing up to arrangements: face-count formulas for partitions of space by hyperplanes.* Mem. Amer. Math. Soc. 1 (1975), issue 1, no. 154, vii+102 pp.

[108] T. Zhang, G. Golub, *Rank-one approximation to high order tensors,* SIAM J. Matrix Anal. Appl. 23 (2001), no. 2, 534–550.

[109] Yu. A. Zuev, *Asymptotics of the logarithm of the number of Boolean threshold functions.* (Russian) Dokl. Akad. Nauk SSSR 306 (1989), 528–530; translation in Soviet Math. Dokl. 39 (1989), no. 3, 512–513.

[110] Yu. A. Zuev, *Combinatorial-probability and geometric methods in threshold logic.* (Russian) Diskret. Mat. 3 (1991), no. 2, 47–57; translation in Discrete Math. Appl. 2 (1992), no. 4, 427–438.

[111] J. Zunic, *On encoding and enumerating threshold functions,* IEEE Transactions on Neural Networks 15 (2004), 261–267.

Department of Computer Science, University of California, Irvine
E-mail address: pfbaldi@uci.edu

Department of Mathematics, University of California, Irvine
E-mail address: rvershyn@uci.edu