BOOLEAN POLYNOMIAL THRESHOLD FUNCTIONS 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 $\log_2 T(n, d) \approx n^2$, 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 $\log_2 T(n, d) \approx n^{d+1}/d!$. The solution relies on establishing connections between the theory of Boolean functions and high-dimensional probability theory and leads to a more general program of extending random matrix theory to random tensors.

Contents

1. Introduction

A Boolean function $f : \{-1, 1\}^n \to \{-1, 1\}$ is called a *polynomial threshold function* if it has the form

$$f(x) = \mathrm{sign}(p(x))$$

Date: March 28, 2018.

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

for some real polynomial p which has no roots in the Boolean cube. 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 [52]. Linear threshold functions were studied by T. Cover [24], S. Muroga [56], M. Minsky and S. Papert in their book on perceptrons [55], 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, [7, 8, 9, 17, 78, 12, 47, 82, 6, 14, 2, 42, 43, 26, 64, 65, 13, 38]. An introduction to polynomial threshold functions can be found in [63, Chapter 5], [5, Chapter 4], and [78]. 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 [80].

There are 2^{2^n} Boolean functions of n variables. Each one of them can be expressed as a polynomial of degree at most n: to see this, write the function f in conjunctive normal form or take the Fourier transform of f. In particular, every Boolean function f is a polynomial threshold function, but the the polynomial that represents f often has high degree. A conjecture of J. Aspnes *et al.* [6] and C. Wang and A. Williams [97] states that for most Boolean functions $f(x)$, the lowest degree of $p(x)$ such that $f(x) = \operatorname{sign}(p(x))$ is either $\lfloor n/2 \rfloor$ or $\lceil n/2 \rceil$. M. Anthony [4] and independently N. Alon (see [78]) 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 approximately by R. O'Donnell and R. A. Servedio [64] who gave an upper bound $n/2 + O(\sqrt{n \log n})$ on the degree of $p(x)$.

Of special interest are low-degree polynomial threshold functions, and some major conjectures about them remain open. One of the most basic questions is:

How many low-degree polynomial threshold functions are there?

More precisely, what is the asymptotic behavior of $T(n,d)$, the number of polynomial threshold functions of fixed degree d? A tight answer is known only for the linear case $d = 1$. The work of T. Cover [24] 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 [56] and others provided lower bounds of the form $2^{\alpha n^2}$ with values of α that were significantly below 1. Yu. Zuev [101, 102] was finally able to show in 1989 that the upper bound was tight in the sense that the number of linear threshold functions satisfies

$$\log_2 T(n,1) = n^2 + o(n^2). \tag{1.1}$$

Although this result led to some further progress in understanding linear threshold functions (see e.g. [67, 103, 27, 35, 36, 39]), the same problem for higher degrees remained open. M. Saks explicitly asked about the asymptotical behavior of $T(n,d)$ in 1993 [78, Problem 2.35]. Even the number of quadratic threshold functions $T(n,2)$ has been unknown.

In this paper, we settle the problem for all degrees $d \in \mathbb{N}$.

Theorem 1.1. *For any fixed degree $d \geq 1$, the number of polynomial threshold functions $T(n,d)$ satisfies*

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

as $n \to \infty$.

Theorem 1.1 states that one needs approximately $n^{d+1}/d!$ bits to specify an n-variable polynomial threshold function of degree d. Since a general polynomial consists of $\binom{n+d}{d} \approx n^d/d!$ monomial terms, Theorem 1.1 can be equivalently stated as follows:

> *To specify a polynomial threshold function of any fixed degree,*
> *one needs to spend approximately n bits per monomial term.*

Furthermore, Theorem 1.1 determines the complexity of a polynomial *classification problem*. Suppose we want to separate the points of the Boolean cube $\{-1,1\}^n$ into two classes by some polynomial surface of degree d (the zero set of a polynomial). Theorem 1.1 says that there are approximately $2^{n^{d+1}/d!}$ different ways one can achieve this.

Although we stated Theorem 1.1 for a fixed degree d, we can allow d to grow mildly with n, and our result still holds if $d = o(\sqrt{\log n / \log\log n})$. This follows from a sharper, quantitative version of Theorem 1.1 that is valid for fixed n and d; see Theorem 3.6 for a (known) upper bound and Theorem 9.3 for a (new) lower bound.

1.1. **Prior work and overview.**

Prior to our work, the best known general bounds on the number of polynomial threshold functions were given by

$$\binom{n}{d+1} \leq \log_2 T(n,d) \leq \frac{n^{d+1}}{d!}. \tag{1.2}$$

Versions of the upper bound go back to 1960s [24] and the present form was given by P. Baldi [8] in 1988. The lower bound was established by M. Saks [78] in 1993. We include a proof of the upper bound in Section 3.5; see [5, Sections 4.5, 4.6] for detailed derivation of both bounds. As we see from Theorem 1.1, the upper bound in (1.2) turns out to be optimal. In contrast, the lower bound in (1.2) is approximately $n^{d+1}/(d+1)!$, which leaves a multiplicative gap $O(d)$ between the the upper and lower bounds in (1.2). Our work closes this gap.

The asymptotically sharp result (1.1) about linear threshold functions has a remarkably short proof [102]. It can be quickly deduced 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* [99], and the probabilistic result is Odlyzko's theorem on spans of random ±1 vectors [62]. Odlyzko's theorem, in turn, is closely related to a theorem on the *singularity of random matrices*, which states that random matrices with ±1 entries have full rank with high probability. The original results on the singularity problem are due to J. Komlós [45, 46]. More recently, the singularity problem has been actively studied in random matrix theory, and a significant number of extensions and improvements on the result of J. Komlós are now available, see e.g. [40, 85, 22, 86, 70, 90, 71, 72, 1, 73, 88, 16, 74, 58, 60, 94, 76, 32, 11, 91, 92, 50, 20, 93].

One may attempt to extend the approach used for $T(n,1)$ to higher degrees $d > 1$ by lifting it into the tensor product space $(\mathbb{R}^n)^{\otimes d}$. This attempt however hits a bottleneck: while the theory of random matrices is well developed, and in particular the singularity problem of random matrices is well understood, its extension to *random tensors* is still in its infancy. Hence a broader goal of this paper is to explore new directions in the theory of random tensors and create new methods for their analysis. In particular, we develop several tools for random tensors: decoupling of small ball probabilities (Section 4), contractions (Section 5), restrictions (Section 6), singularity (Section 7), and uniqueness of spans (Section 8). But before we develop and apply these tools, let us pause to give a heuristic argument that leads to Theorem 1.1.

2. Overview of the argument

Let us describe the main ideas of the proof of Theorem 1.1. We shall focus here on the lower bound on the number of threshold functions. The upper bound is known and relatively simple; we will prove it in Section 3.5.

2.1. **Tensor lift.** First we linearize the problem using a standard tensor lifting trick. Instead of $\mathbb{R}^n$ we will mostly work in $\mathrm{Sym}^d(\mathbb{R}^n)$, the vector space of symmetric tensors of order d, which has dimension

$$N(n,d) = \binom{n+d-1}{d} = \frac{n^d}{d!} + o(n^d). \tag{2.1}$$

Instead of the Boolean hypercube $\{-1,1\}^n$, we consider its tensor lift $\mathcal{X} \subset \mathrm{Sym}^d(\mathbb{R}^n)$, which we define as

$$\mathcal{X} := \Big\{ x^{\otimes d} : \; x \in \{-1,1\}^n \Big\}. \tag{2.2}$$

Every real homogeneous polynomial $p(x)$ in n variables and of degree d can be represented as

$$p(x) = \sum_{i_1,\ldots,i_d=1}^{n} A_{i_1,\ldots,i_d}\, x_{i_1} \cdots x_{i_d} = \langle A, x^{\otimes d} \rangle$$

for some $A \in \mathrm{Sym}^d(\mathbb{R}^n)$. This means that every homogeneous polynomial threshold function in n variables of degree d has the form

$$f(x) = \mathrm{sign}\langle A, x^{\otimes d} \rangle.$$

This gives a one-to-one correspondence between homogeneous *polynomial* threshold functions on the Boolean hypercube $\{-1,1\}^n$ and homogeneous *linear* threshold functions on $\mathcal{X}$.

Remark 2.1. Since $x_i^2 = 1$ for $x_i = \pm 1$, some homogeneous polynomials on the Boolean cube may be reduced further; for example we have $2x_1^2 - 3x_1x_2 - x_1^2 = 1 - 3x_1x_2$. For this reason, the term *homogeneous* is sometimes used in the literature for a smaller class of polynomials $p(x)$ which consist of at most $\binom{n}{d}$ pure monomial terms of the form $A_{i_1,\ldots,i_d}\, x_{i_1} \ldots x_{i_d}$, where all the indices i_k are different [8]. For example, the polynomial $2x_1^2 - 3x_1x_2 - x_1^2$ is not of this "pure" kind but $3x_1x_2$ is. These "purely homogeneous" polynomials correspond to symmetric tensors with no extended diagonal terms, i.e. no terms containing at least two identical indices. The techniques we develop in this paper should be generalizable for purely homogeneous polynomials.

2.2. **Hyperplane arrangements.** Counting linear threshold functions is related to a problem in enumerative combinatorics that has been studied for decades, namely the problem of counting regions in hyperplane arrangements [99]; see [83], [51, Section 6] for an introduction.

For each tensor $x^{\otimes d}$ in $\mathcal{X}$, consider the hyperplane orthogonal to it, i.e.

$$(x^{\otimes d})^\perp = \left\{ A \in \mathrm{Sym}^d(\mathbb{R}^n) : \langle A, x^{\otimes d} \rangle = 0 \right\}.$$

These hyperplanes partition the space $\mathrm{Sym}^d(\mathbb{R}^n)$ into open connected components called *regions* of the hyperplane arrangement. Clearly, two polynomial threshold functions $f_A(x) = \mathrm{sign}\langle A, x^{\otimes d}\rangle$ and $f_B(x) = \mathrm{sign}\langle B, x^{\otimes d}\rangle$ are identical if and only if the tensors A and B lie on the same side of all the hyperplanes $(x^{\otimes d})^\perp$, i.e. A and B belong to the same region of the hyperplane arrangement. Therefore:

> *There are as many different polynomial threshold functions as there are regions of the hyperplane arrangement $(x^{\otimes d})^\perp$, $x \in \{-1,1\}^n$.*

Counting regions (and, more generally, faces of any dimension) of hyperplane arrangements is a well known problem in enumerative combinatorics. There exist general counting methods that give exact expressions for the number of regions [99, 83], but such methods are often hard to apply. Instead, we will be satisfied with the following simple lower bound, which was first noted and used in a similar way by Yu. Zuev [102]:

> *The number of regions in any hyperplane arrangement is bounded below by the number of all intersection subspaces.*

An intersection subspace here refers to the intersection of any number of the hyperplanes, ranging from dimension zero (a point) to N (intersecting an empty set of hyperplanes gives the entire space $\mathbb{R}^N$). For example, the line arrangement in Figure 1 on the left has seven regions and seven intersection subspaces – three points, three lines and one plane. The line arrangement in Figure 1 on the right has six regions and five intersection subspaces.

Figure 1. Two hyperplane arrangements in $\mathbb{R}^2$

2.3. **General position.** Suppose for a moment that the points in $\mathcal{X}$ are in general position (in reality they are not). Then every subset $x_1^{\otimes d}, \ldots, x_m^{\otimes d} \in \mathcal{X}$ of $m = N(n,d) - 1$ points must be linearly independent, and the hyperplane spanned by this set may not contain any other points from $\mathcal{X}$, up to a sign. It follows that there are as many intersection subspaces as $N(n,d) - 1$ element subsets of $\mathcal{X}$, which is $\binom{|\mathcal{X}|}{N(n,d)-1}$. By the previous reductions, this gives

$$T(n,d) \geq \binom{|\mathcal{X}|}{N(n,d)-1}.$$

To simplify this bound, we may note that $|\mathcal{X}| \geq 2^{n-1}$ and recall the value of $N(n,d)$ from (2.1). A simple asymptotic analysis lets us conclude that

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

as claimed in Theorem 1.1.

2.4. **Random tensors: linear independence.** The problem with our argument is that the points in the tensor lift $\mathcal{X}$ of the Boolean hypercube (2.2), 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 core of this paper is the following result, which is interesting on its own. It states that stochastically independent simple tensors tends to be linearly independent. The number of such tensors can be almost as large as the dimension $N(n,d)$ of the ambient space $\mathrm{Sym}^d(\mathbb{R}^n)$, which is an obvious upper bound.

Theorem 2.2 (Random tensors are linearly independent)**.** *Let $x_1, \ldots, x_m$ be independent random vectors uniformly distributed in $\{-1,1\}^n$. Then, for any fixed degree d, the set of*

$$m = N(n,d)(1 - o(1))$$

random tensors $x_1^{\otimes d}, \ldots, x_m^{\otimes d}$ is linearly independent with high probability.

Theorem 7.3 gives a more formal, quantitative statement of this result.

The partial case of Theorem 2.2 where $d = 1$ is known and simple (see Proposition 3.9) and can be rephrased in terms of random matrices. Indeed, consider the $n \times m$ matrix whose columns are x_k. This is an almost square random matrix with independent ± 1 entries. Theorem 2.2 for $d = 1$ states that such matrix is non-singular (i.e. has full rank) with high probability. Singularity of random matrices, and in particular those with ± 1 entries, has been extensively studied in random matrix theory, and results that are more general and stronger than the case $d = 1$ of Theorem 2.2 are known [40, 85, 22, 86, 70, 90, 71, 72, 1, 73, 88, 16, 74, 58, 60, 94, 76, 32, 11, 91, 92, 50, 20].

For tensors of any higher order $d > 1$, Theorem 2.2 is new and considerably harder to prove. This result is non-trivial even for $d = 2$, where it states that *stochastically independent random matrices tend to be linearly independent.* More precisely, in this case we have $m = n^2/2 - o(n^2)$ and Theorem 2.2 yields that m independent symmetric rank-one ± 1 matrices $x_1 x_1^\mathsf{T}, \ldots, x_m x_m^\mathsf{T}$ are linearly independent with high probability.

It is perhaps surprising to note that for $d > 1$, the random vectors $x_1, \ldots, x_m$ in Theorem 2.2 must be linearly *dependent* as in this case we have $m > n$. Thus, the theorem shows that tensor products tend to create linearly independent tensors from linearly dependent vectors.

The principal difficulty in proving Theorem 2.2 is the *lack of independence.* While a random vector x sampled uniformly from the Boolean hypercube $\{-1, 1\}^n$ has independent entries, the random tensor $x^{\otimes d}$ does not. Indeed, the tensor $x^{\otimes d}$ has n^d entries that are generated by just n random independent bits. This difficulty prompts us to develop several new tools for the analysis of random tensors in Sections 4–7, which we will describe now.

2.4.1. *A heuristic proof of Theorem 2.2.* Let us outline our approach to in the simplest nontrivial case, namely $d = 2$. Here we are looking at the space of $n \times n$ symmetric matrices, which has dimension $N(n, 2) = n(n+1)/2$. Let us try to establish linear independence of $m = (1 - \varepsilon)n^2/2$ random matrices $x_k \otimes x_k = x_k x_k^\mathsf{T}$. They are linearly independent if and only if each one of them does not lie in the span L of the other matrices. If we condition on all except one x_k, we see that the task reduces to showing that

$$P := \mathbb{P}\left\{ x x^\mathsf{T} \in L \right\} \le \text{something small}$$

where x is a random vector uniformly distributed in $\{-1, 1\}^n$ and L is a fixed subspace of dimension at most m in the space of symmetric matrices. Consider $E := L^\perp$, a subspace of dimension at least $\varepsilon n^2/2$. We have

$$P = \mathbb{P}\left\{ \langle A, x x^\mathsf{T} \rangle = 0 \quad \forall A \in E \right\}. \tag{2.3}$$

We develop a new *decoupling* technique in order to replace $x x^\mathsf{T}$ by $x y^\mathsf{T}$ where y is an independent copy of x. While many decoupling results are known in the probability literature (see [25] for a comprehensive treatment), none of them applies to a version of (2.3) for higher-order tensors. In Section 4 we establish a new decoupling inequality for *small ball probabilities of random tensors*, which is of independent interest. Paired with a *restriction* technique we develop in Section 6, the decoupling inequality (roughly) gives

$$P \lesssim \mathbb{P}\Big\{ \langle A, xy^\mathsf{T} \rangle = 0 \quad \forall A \in E \Big\} = \mathbb{P}\Big\{ \langle Ay, x \rangle = 0 \quad \forall A \in E \Big\}.$$

Consider the random subspace $Ey := \{Ay : \ A \in E\}$; then

$$P \lesssim \mathbb{P}\Big\{ \langle z, x \rangle = 0 \quad \forall A \in Ey \Big\} = \mathbb{P}\left\{ x \in (Ey)^\perp \right\}.$$

This reduces our problem to a question that is much simpler and betterunderstood. Namely, we seek to bound the probability that a random vector x (rather than a random matrix xx^T) falls into a given subspace $Ey \subset \mathbb{R}^n$. We can view Ey as the *contraction* of a fixed matrix subspace E by a random vector y. In Section 5 we prove a key result about the dimensions of tensor contractions. It says roughly that

$$\dim(Ey) \gtrsim \frac{\dim(E)}{n}$$

with high probability, which is an optimal bound. Since $\dim(E) \ge \varepsilon n^2/2$, we get $\dim(Ey) \gtrsim \varepsilon n$ and thus

$$\dim(Ey)^\perp \lesssim (1-\varepsilon)n.$$

It is well known (see Section 3.6.1) that a random vector x does not fall into a given subspace of $\mathbb{R}^n$ of dimension at most $(1-\varepsilon)n$ with high probability, namely with probability $1 - \exp(-c\varepsilon n)$. This gives

$$P \lesssim \exp(-c\varepsilon n)$$

which leads to the conclusion of Theorem 2.2.

2.5. **Random tensors: unique spans.**

An important consequence of Theorem 2.2 is that the linear spans of the random tensors $x_k^{\otimes d}$ are unique with high probability – these spans do not contain any other vector of the same kind up to a sign.

Theorem 2.3 (Uniqueness of spans). *Let $x_1, x_2, \ldots, x_m$ be independent random vectors uniformly distributed in $\{-1,1\}^n$. Then, for any fixed degree d, with high probability, the span of*

$$m = N(n,d)(1-o(1))$$

random tensors $x_1^{\otimes d}, \ldots, x_m^{\otimes d}$ does not contain any simple tensor $u^{\otimes d}$ that is different from $\pm x_k^{\otimes d}$.

A more formal, quantitative version of this result is Theorem 8.1 below.

The uniqueness property is not a direct consequence of Theorem 2.2, and it requires additional probabilistic tools. For the partial case $d = 1$, this derivation was made by Odlyzko [62] with a very sharp bound on the probability (which we do not need here). It was later noticed that the singularity and uniqueness problems are actually asymptotically equivalent [96]. Odlyzko's result was used in Zuev's argument [102] to prove (1.1).

For tensors of any order $d > 1$, Theorem 2.3 is new. We derive it from Theorem 2.2 using a (non-trivial) development of Odlyzko's method.

2.6. Plan for the rest of the paper.

In Section 3, we provide some background material on tensors (Section 3.3), hyperplane arrangements (Section 3.4), and high-dimensional probability theory (Section 3.6). In Section 3.5 we give a proof of the (known) upper bound in Theorem 1.1. In Section 4 we establish a new decoupling inequality for small ball probabilities of random tensors. In Section 5 we study the dimension of random contractions of tensor subspaces. In Section 6 we show how to restrict any tensor subspace onto a set of independent coordinates without sacrificing a lot of the dimension; this result becomes especially useful when applied together with decoupling. In Section 7, all of these tools are put together to prove Theorem 2.2 on the linear independence of random tensors (the formal result being Theorem 7.3). In Section 8, we deduce Theorem 2.3 on the uniqueness of spans of random tensors (the formal result being Theorem 8.1). In Section 9, we prove our main result, the lower bound in Theorem 1.1 (the formal result being Theorem 9.3). Finally, in Section 10 we describe several extensions and open problems.

3. Preliminaries

3.1. Basic notation and conventions.

By $C, c, C_1, c_1, \dots$ we denote positive absolute constants, whose precise values can be different from line to line. We stress that such constants may not depend on the number of variables n or the degree d (or on any other variables in question).

For an integer m, we use the following notation for the integer interval: $[m] = \{1, \dots, m\}$. We will routinely use the following elementary and well known bounds on the binomial sums:

$$\left(\frac{n}{m}\right)^m \le \binom{n}{m} \le \sum_{k=0}^{m} \binom{n}{k} \le \left(\frac{en}{m}\right)^m \tag{3.1}$$

for all integers $1 \le m \le n$, see e.g. [95, Exercise 0.0.5].

3.2. Polynomial threshold functions.

We call a polynomial threshold function $f(x) = \operatorname{sign}(p(x))$ *homogeneous* if $p(x)$ is a real homogeneous polynomial. If furthermore $p(x)$ is a linear function, we call $f(x)$ a homogeneous linear threshold function. As the following lemma shows, homogeneity restriction does not significantly affect the count of threshold functions.

Lemma 3.1 (The effect of homogeneity)**.** *For any n and d, the number of homogeneous polynomial threshold functions $\bar{T}(n,d)$ and the number of all polynomial threshold functions $T(n,d)$ satisfy*

$$\bar{T}(n,d) \le T(n,d) \le \bar{T}(n+1,d).$$

Proof. The first bound is trivial. To prove the second, let $p(x)$ be a polynomial in n variables $x = (x_1, \dots, x_n)$ of degree d. We can realize $p(x)$ as a restriction of some homogeneous polynomial $\bar{p}(x, x_{n+1})$ in $n+1$ variables of degree d onto $x_{n+1} = 1$. For example, $p(x_1, x_2) = x_1^2 + 5x_2 + 7$ is a restriction of $\bar{p}(x_1, x_2, x_3) = x_1^2 + 5x_2x_3 + 7x_3^2$ onto $x_3 = 1$. Obviously the map $p \mapsto \bar{p}$ is injective. This proves the second bound in the lemma. □

By default, the polynomial threshold functions are defined on the Boolean hypercube $\{-1, 1\}^n$. However, occasionally we will need to consider functions $f(x) = \operatorname{sign}(p(x))$ on different domains $\mathcal{X} \subset \mathbb{R}^n$. In such cases, we call these f polynomial threshold functions on $\mathcal{X}$.

3.3. **Tensors and operations on them.** For additional background and references on tensors, please see [53, 23, 44]. In this paper, by a *tensor* we mean a multi-dimensional array or real numbers

$$A = (A_{i_1,\dots,i_d}) \in \mathbb{R}^{n_1 \times \cdots \times n_d}.$$

Thus, tensors for which $d = 1$ are vectors, and tensors for which $d = 2$ are matrices. A *simple tensor* is the tensor product of vectors $x_1 \in \mathbb{R}^{n_1}, \dots, x_d \in \mathbb{R}^{n_d}$, which is defined as

$$x_1 \otimes \cdots \otimes x_d = (x_{1i_1} \cdots x_{di_d}) \in \mathbb{R}^{n_1 \times \cdots \times n_d}.$$

For example, $x \otimes y = (x_i y_j) = xy^\mathsf{T}$ is a rank-one matrix. If we tensor-multiply a vector x by itself d times, we often write the result as

$$x^{\otimes d} = x \otimes \cdots \otimes x.$$

3.3.1. *Space of symmetric tensors.* A symmetric tensor is a tensor that is invariant under permutation of the indices, i.e.

$$A_{i_1,\dots,i_d} = A_{i_{\sigma(1)},\dots,i_{\sigma(d)}}$$

for every permutation σ of the set $\{1, 2, \dots, d\}$. For $d = 2$, symmetric tensors are just symmetric matrices. The space of symmetric tensors is denoted $\operatorname{Sym}^d(\mathbb{R}^n)$. The dimension of this space equals

$$N(n, d) := \dim\left(\operatorname{Sym}^d(\mathbb{R}^n)\right) = \binom{n+d-1}{d}. \tag{3.2}$$

If d is fixed and $n \to \infty$, we have

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

The following lemma gives a quantitative form of this asymptotic statement.

Lemma 3.2. *Let $1 \le d \le c\sqrt{n}$. Then the dimension of the space of symmetric tensors $\operatorname{Sym}^d(\mathbb{R}^n)$ satisfies*

$$\frac{n^d}{d!} \le N(n, d) \le \frac{n^d}{d!}\Big(1 + \frac{2d^2}{n}\Big).$$

Proof. The lower bound follows immediately from the identity (3.2), which we can write as

$$N(n,d) = \frac{n(n+1)\cdots(n+d-1)}{d!}. \tag{3.3}$$

As for the upper bound, (3.3) implies that $N(n,d) \le (n+d)^d/d!$. Furthermore, one can easily check that $(n+d)^d \le n^d(1+2d^2/n)$ if $d \le \sqrt{n}$. Combining these two inequalities completes the proof. □

Occasionally, the following much weaker but simpler bounds can be useful, too:

$$n \le N(n,d) \le n^d. \tag{3.4}$$

To check the upper bound, it is enough to recall that $N(n,d)$ is dimension of the space of symmetric tensors while n^d is the dimension of the space of all tensors. The lower bound follows from the fact that $N(n,d)$ is the dimension of the space of symmetric tensors, so $N(n,d) \ge N(n,1) = n$.

3.3.2. *Inner product of tensors.* The inner product of two tensors $A, B \in \mathbb{R}^{n_1\times\cdots\times n_d}$ is defined by

$$\langle A, B\rangle := \sum_{i_1,\ldots,i_d} A_{i_1,\ldots,i_d} B_{i_1,\ldots,i_d}. \tag{3.5}$$

For $d = 1$, this definition reduces to the canonical inner product of vectors in $\mathbb{R}^n$. For $d = 2$, we obtain the *Frobenius inner product* of matrices

$$\langle A, B\rangle = \sum_{ij} A_{ij}B_{ij} = \operatorname{tr}(AB^\mathsf{T}).$$

In particular,

$$\langle A, x\otimes y\rangle = \sum_{ij} A_{ij}x_iy_j = x^\mathsf{T}Ay$$

is a *bilinear form.* If $x = y$, this is a *quadratic form.* For general d, the inner product with a simple tensor defines a *multilinear form.* For example, if $d = 3$ we have

$$\langle A, x\otimes y\otimes z\rangle = \sum_{i,j,k} A_{ijk}\, x_iy_jz_k.$$

If $x = y = z$, this is a homogeneous cubic polynomial.

3.3.3. *Tensor-vector multiplication.* There seems to be no standard notion of tensor-vector multiplication in the literature. We define the product of a tensor $A \in \mathbb{R}^{n_1\times\cdots\times n_d}$ by the vector $x \in \mathbb{R}^{n_d}$ as the tensor $Ax \in \mathbb{R}^{n_1\times\cdots\times n_{d-1}}$ whose entries are computed as follows:

$$(Ax)_{i_1,\ldots,i_{d-1}} := \sum_{i_d=1}^{n_d} A_{i_1,\ldots,i_d}\, x_{i_d}.$$

Multiplication defined in this way contracts the *last* coordinate of the tensor. In the particular case where $d = 2$, tensor-vector multiplication corresponds to the standard matrix-vector multiplication. Its result is the vector with coordinates

$$(Ax)_i = \sum_j A_{ij} x_j.$$

If $d = 3$, the result of the tensor-vector multiplication is the matrix with entries

$$(Ax)_{ij} = \sum_k A_{ijk} x_k.$$

3.3.4. *A cyclic identity.* The definitions of inner product and tensor-vector multiplication lead to the following elementary and useful identity:

$$\langle A, x_1 \otimes \cdots \otimes x_d \rangle = \langle A x_d, x_1 \otimes \cdots \otimes x_{d-1} \rangle. \tag{3.6}$$

In the particular case where $d = 2$, we have the following identity for matrices:

$$\langle A, x \otimes y \rangle = \langle A, xy^\mathsf{T} \rangle = \langle Ay, x \rangle.$$

By induction, one can deduce from (3.6) the following cyclic identity:

$$\langle A, x_1 \otimes \cdots \otimes x_d \rangle = \langle A x_d x_{d-1} \cdots x_2, x_1 \rangle. \tag{3.7}$$

Here $A x_d x_{d-1} \cdots x_2$ is the vector in $\mathbb{R}^{n_1}$ that is obtained by recursively multiplying A by x_d then by x_{d-1} and so on. The coordinates of this vector are

$$(A x_d x_{d-1} \cdots x_2)_{i_1} = \sum_{i_2,\ldots,i_d} A_{i_1,\ldots,i_d} \, x_{i_2} \cdots x_{i_d}.$$

For example, if $d = 3$, we have

$$(Azy)_i = \sum_{jk} A_{ijk} \, y_j z_k.$$

3.4. **Hyperplane arrangements.** Counting linear threshold functions is equivalent to counting regions of hyperplane arrangements, a problem that has been studied in enumerative combinatorics for decades [99]; see [83], [51, Section 6] for an introduction. The following elementary and well known lemma describes this connection.

Lemma 3.3 (Threshold functions and hyperplane arrangements)**.** *Let $\mathcal{X} \subset \mathbb{R}^N$ be a finite subset. Then the number of homogeneous linear threshold functions $f : \mathcal{X} \to \{-1, 1\}$ equals the number of open components in the partition of the space $\mathbb{R}^N$ by the hyperplanes $x^\perp$, $x \in \mathcal{X}$.*

Proof. A homogeneous threshold function $f_a(x) = \operatorname{sign}\langle a, x \rangle$ on $\mathcal{X}$ is determined by the vector $a \in \mathbb{R}^N$. Two vectors a and b define the same function $f_a = f_b$ if and only if a and b are on the same side of each hyperplane $x^\perp$ for each $x \in \mathcal{X}$. In other words, $f_a = f_b$ if and only if a and b lie in the same open component of the partition of $\mathbb{R}^N$ by the hyperplanes $x^\perp$. This completes the proof. □

Lemma 3.3 leads to the following general question: given an arrangement of affine hyperplanes, how many open components ("regions") does it divide the space into? An exact formula for the number of regions was found by Zaslavsky [99]; see [83]. It expresses the number of components via the Möbius function of the poset of the intersection spaces of the hyperplanes. Zaslavsky theorem implies the following convenient and asymptotically tight bounds.

Lemma 3.4 (Counting components of hyperplane arrangements)**.** *Consider an arrangement of $K \geq N$ affine hyperplanes in $\mathbb{R}^N$. Let $r(K,N)$ denote the number of regions of this arrangement.*

(1) We have

$$r(K,N) \leq \binom{K}{0} + \binom{K}{1} + \cdots + \binom{K}{N}.$$

(2) $r(K,N)$ is bounded below by the number of all intersection subspaces[1] *defined by the hyperplanes.*

If the hyperplanes are in general position, then the upper and lower bounds in Lemma 3.4 are clearly the same, and each bound becomes an equality.

Lemma 3.4 immediately follows from Zaslavsky's formula [99] and basic properties of the Möbius function of geometric lattices [69, Section 7]; see e.g. [83, Proposition 2.4] for the derivation of the upper bound. The upper bound goes back to R. C. Buck [18] and the lower bound was noted by Yu. Zuev [102]. Both bounds can be proved directly – without Zaslavsky's formula – using simple inductive arguments, see [51, Section 6] for the upper bound and [102] for the lower bound.

3.4.1. *Central arrangments.* A central arrangement in $\mathbb{R}^n$ is one in which the intersection of all hyperplanes is nonempty. This happens, in particular, if all hyperplanes pass through the origin, which happens in all applications we consider in this paper. For central arrangements, the upper bound in Lemma 3.4 can be slightly improved, namely we have

$$r(K,N) \leq 2\left[\binom{K-1}{0} + \binom{K-1}{1} + \cdots + \binom{K-1}{N-1}\right]. \tag{3.8}$$

Moreover, if the normal vectors to the hyperplanes are in general position, then the inequality in (3.8) becomes an equality.

Consider, for example, the central hyperplane arrangement in Figure 1 on the right. The normal vectors of the hyperplanes are in general position. There are six regions, and (3.8) gives the optimal bound $r(3,2) \leq 2\left[\binom{2}{0} + \binom{2}{1}\right] = 6$.

[1] An intersection subspace in this lemma refers to the intersection of any subfamily of the hyperplanes. The dimensions of intersection subspaces may range from zero (the origin is the intersection of all hyperplanes) to N (intersecting an empty set of hyperplanes gives the entire space $\mathbb{R}^N$).

The bound (3.8), just like Lemma 3.4, can be quickly derived from Zaslavsky's formula, see e.g. [67]. Alternatively, one can prove it by simple induction. Such was the original proof of (3.8) due to L. Schläfli [81, pp. 209–212], which is reproduced e.g. in [98] and, more explicitly, in [5, Theorem 4.1]

3.5. **The upper bound.** Lemmas 3.3 and 3.4 quickly lead to the tight (and known) upper bound on the number of threshold functions [8].

Proposition 3.5. *Let $1 \le d \le c \log n$. Then the number of homogeneous polynomial threshold functions $\bar{T}(n,d)$ satisfies*

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

Proof. Every homogeneous polynomial in n variables of degree d can be represented as $p(x) = \langle A, x^{\otimes d} \rangle$ for some $A \in \mathrm{Sym}^d(\mathbb{R}^n)$. Thus, every homogeneous polynomial threshold function in n variables of degree d has the form

$$f(x) = \mathrm{sign}\langle A, x^{\otimes d} \rangle.$$

This gives a one-to-one correspondence between homogeneous polynomial threshold functions on the Boolean hypercube $\{-1,1\}$ and homogeneous linear threshold functions on the tensor lift $\mathcal{X} \subset \mathrm{Sym}^d(\mathbb{R}^n)$ of the Boolean hypercube, which is defined as

$$\mathcal{X} := \Big\{ x^{\otimes d} : \; x \in \{-1,1\}^n \Big\}.$$

We can apply Lemmas 3.3 and 3.4 to $\mathcal{X}$, where $K = |\mathcal{X}| \le 2^n$ and $N = N(n,d)$. This yields

$$\bar{T}(n,d) \le \sum_{k=0}^{N} \binom{2^n}{k} \le \Big(\frac{e2^n}{N} \Big)^N,$$

where we used the elementary bound (3.1) for the binomial sum. Taking the logarithm of both sides, we get

$$\log_2 \bar{T}(n,d) \le N(2 + n - \log_2 N).$$

Applying Lemma 3.2 to bound $N = N(n,d)$ and redistributing the powers of n, we obtain

$$\log_2 \bar{T}(n,d) \le \frac{n^{d+1}}{d!} \Big(1 + \frac{2d^2}{n}\Big)\Big(1 + \frac{2 - \log_2 N}{n}\Big). \tag{3.9}$$

The lower bound in Lemma 3.2 and the elementary inequality $d! \ge (d/e)^d$ give $N(n,d) \ge (en/d)^d$, and thus $\log_2 N \ge d \log(en/d)$. Therefore, if $d < c \log n$ with sufficiently small absolute constant $c > 0$, we have $\log_2 N - 2 \ge 2d^2$. Substituting this into (3.9), we conclude that

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

as claimed. □

Proposition 3.5 can be extended to all polynomial threshold functions, not necessarily homogeneous. The following result is due to P. Baldi [8].

Theorem 3.6 (Upper bound)**.** *For any $1 \le d \le n$, the number of polynomial threshold functions $T(n, d)$ satisfies*

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

Proof. Recall from Proposition 3.1 that $T(n, d) \le \bar{T}(n + 1, d)$. Thus we can essentially repeat the proof of Proposition 3.5 replacing n by $n + 1$ where needed. We skip the details. □

3.6. **Probability background.** The lower bound in Theorem 1.1 is more challenging. In addition to using hyperplane arrangements (Lemma 3.4), our argument will critically rely on probabilistic arguments. For convenience, we describe here several basic tools of high-dimensional probability theory.

3.6.1. *Distances to subspaces.* Our first result concerns the probability that a random vector falls inside a fixed subspace in $\mathbb{R}^n$. The following result gives a useful concentration inequality for the distance between a vector and a subspace.

Lemma 3.7 (Distance to a subspace)**.** *Let F be a subspace of $\mathbb{R}^n$ with codimension $q = \operatorname{codim}(F)$. Let x be a random vector uniformly distributed in $\{-1, 1\}^n$. Then, for every $t \ge 0$, we have*

$$\mathbb{P}\left\{|\operatorname{dist}(x, F) - \sqrt{q}| \ge t\right\} \le 2\exp(-ct^2).$$

Lemma 3.7 is a well known and widely used result in high-dimensional probability. It can be derived from the Hanson-Wright concentration inequality for quadratic forms. Indeed, we can represent the distance as

$$\operatorname{dist}(x, F) = \|Px\|_2$$

where P is the orthogonal projection in $\mathbb{R}^n$ onto the q-dimensional subspace $F^\perp$. Thus $\|Px\|_2^2$ is a quadratic form in x, and the Hanson-Wright inequality can be used to study its concentration properties, see e.g. [75], [95, Section 6.3, Exercise 6.3.4]. Like most results in this paper, Lemma 3.7 applies not only to binary ±1 distributions but can be extended to general sub-gaussian distributions.

Alternatively, one can note that $f(x) := \|Px\|_2$ is a convex and Lipschitz function on $\mathbb{R}^n$ and deduce Lemma 3.7 from Talagrand's concentration inequality for convex Lipschitz functions [84, Theorem 6.6]; see [48, Corollary 4.10] and [95, Theorem 5.2.16].

One consequence of Lemma 3.7 will be especially useful for us, namely that a random vector x falls into a fixed subspace F with exponentially small probability.

Lemma 3.8 (A random vector is unlikely to fall in a subspace)**.** *Let F be a subspace of $\mathbb{R}^n$; denote $q = \operatorname{codim}(F)$. Let x be a random vector uniformly distributed in $\{-1,1\}^n$. Then*

$$\mathbb{P}\left\{x \in F\right\} \leq 2\exp(-cq).$$

Proof. The event $x \in F$ means that $\operatorname{dist}(x,F) = 0$, which obviously implies that $|\operatorname{dist}(x,F) - \sqrt{q}| \geq \sqrt{q/2}$. To finish the proof, apply Lemma 3.7 for $t := \sqrt{q}/2$. □

It should be noted that there exist more advanced bounds for distances between random vectors and subspaces, which take into account the "arithmetic structure" of the subspaces (see e.g. [73, Section 4]). These more advanced bounds will not be needed here. .

3.6.2. *Singularity of random matrices.* The distance bounds we mentioned quickly imply that random vectors sampled uniformly from $\{-1,1\}^n$ are likely to be linearly independent. The following result is well known.

Proposition 3.9 (Linear independence of random vectors)**.** *Let $q > C\log n$. Let x_k, $k = 1,\ldots,n-q$, be independent random vectors uniformly distributed in $\{-1,1\}^n$. Then the vectors x_k are linearly independent with probability at least $1 - 2\exp(-cq)$.*

Proof. The random vectors x_k are linearly independent if and only if each vector x_k does not lie in the linear span of the other vectors, which we denote by $L_k = \operatorname{span}(x_j)_{j\neq k}$. The codimension on L_k is greater than q. (It equals $q+1$ if the vectors x_k forming L_k are in general position, and is larger if they are not). Applying Lemma 3.8, we see that

$$\mathbb{P}\left\{x_k \in L_k\right\} \leq 2\exp(-cq)$$

for each k. Taking the union bound over all $k = 1,\ldots,n-q$, we see that

$$\mathbb{P}\left\{\text{linear dependence}\right\} = \mathbb{P}\left\{\exists k:\ x_k \in L_k\right\} \leq 2(n-q)\exp(-cq).$$

The assumption on q completes the proof, provided that we choose the absolute constant C large enough. □

Proposition 3.9 can be rephrased in terms of random matrices. It states that the $n \times (n-q)$ random matrix with independent random ±1 entries is non-singular with high probability as long as $q > C\log n$. Singularity of random matrices is a topic that has been extensively studied in probability in recent years. Several sharper and more general versions of Proposition 3.9 are known [40, 85, 22, 86, 70, 90, 71, 72, 1, 73, 88, 16, 74, 58, 60, 94, 76, 32, 11, 91, 92, 50, 20, 93].

One of the main technical goals of the current paper is to extend Proposition 3.9 to *random tensors* as claimed in Theorem 2.2.

3.6.3. *The Littlewood-Offord Lemma.* Let $\xi_1, \dots, \xi_n$ be independent random variables and $a_1, \dots, a_n \in \mathbb{R}$ be fixed coefficients. A classical question, which goes back to J. E. Littlewood and A. C. Offord [49] 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 [49] and sharpened by P. Erdös [28].

Lemma 3.10 (Littlewood-Offord Lemma [28])**.** *Let $\xi_1, \dots, \xi_n$ be independent mean zero 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*[2]

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

A slightly more general version of Lemma 3.10, which bounds the probability that the sum falls in a given neighborhood of u, quickly follows from Sperner's theorem in combinatorics [28], see [15, 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, [28, 79, 33, 30, 90, 71, 73, 89, 61, 94, 59, 21, 54, 77], surveys [87, 74], and many others.

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 3.11 (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 3.10 can be slightly strengthened if $u \neq 0$. Although the following may seem like a small improvement, it can be critical for small values of n.

Lemma 3.12. *If $u \neq 0$ in the Littlewood-Offord Lemma 3.10, then*

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

[2] Here $\lfloor m \rfloor$ denotes the integer part of an integer $m > 0$.

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

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

The proof is complete. □

4. Decoupling small ball probabilities

In probability theory, decoupling refers to a collection of techniques that allow one to break (decouple) dependencies among random variables. The theory of decoupling originated from problems in the geometry of Banach spaces, stochastic processes, and U-statistics. The book [25] offers a comprehensive study of decoupling problems and methods, [95, Section 6.1] for a short introduction to decoupling, and [37, 41, 66] for some recent developments.

A typical situation where decoupling may help is where one deals with a homogeneous polynomial of independent random variables. As we noted in Section 2.1, a homogeneous polynomial of degree d in n variables can be expressed as

$$p(x) = \langle A, x \otimes \cdots \otimes x \rangle$$

where $A \in \mathrm{Sym}^d(\mathbb{R}^n)$ is a symmetric tensor and $x \in \mathbb{R}^n$. Suppose the coordinates of x are independent random variables, and we are interested in the distribution of $p(x)$. Decoupling methods seek to replace the polynomial $p(x)$ by a multilinear form

$$\langle A, x_1 \otimes \cdots \otimes x_d \rangle$$

where $x_1, \ldots, x_n$ are independent copies of the random vector x. After this is done, one typically conditions on all random variables x_k except one. This reduces the problem to studying a *linear* form, or a sum of independent random variables, which is a simpler task.

Classical decoupling techniques are tailored to handle the concentration of $p(x)$ around the mean. Thus, if one wants to bound the probability

$$\mathbb{P}\left\{|p(x) - \mathbb{E}\, p(x)| > t\right\},$$

decoupling may reduce this task to a similar problem for the multilinear form $\langle A, x_1 \otimes \cdots \otimes x_d \rangle$ instead of $p(x)$.

In this paper, however, we will not be concerned with concentration but rather with *small ball probabilities*, which measure the "spread" of the distribution of $p(x)$. We would like to bound the probability that $p(x)$ is near any fixed number $u \in \mathbb{R}^n$, namely

$$\mathbb{P}\left\{|p(x) - u| \le \varepsilon\right\}. \tag{4.1}$$

The supremum of small ball probabilities (4.1) over all $u \in \mathbb{R}$ is called the *Lévy concentration function* of $p(x)$. There seem to be no general decoupling methods for small ball probabilities (4.1) in the literature, with the exception of some techniques for quadratic polynomials ($d = 2$) [31, 21, 22, 94].

In this section we provide a general decoupling inequality for the Lévy concentration function of polynomials of any degree $d \ge 1$. An especially transparent version of this inequality holds for polynomials in Gaussian random variables. We prove this simpler result first, and then extend in Section 4.5 for general distributions and for multiple tensors.

Theorem 4.1 (Decoupling small ball probabilities for Gaussian tensors)**.** *Let $d \ge 2$. Consider a fixed tensor $A \in \mathrm{Sym}^d(\mathbb{R}^n)$, and let x be a standard normal random vector*[3] *in $\mathbb{R}^n$. Let $x_1, \dots, x_d$ be independent copies of x. Then for any $u \in \mathbb{R}$ and $\varepsilon \ge 0$ we have*

$$\mathbb{P}\left\{|\langle A, x \otimes \cdots \otimes x\rangle - u| \le \varepsilon\right\} \le \mathbb{P}\left\{|\langle A, x_1 \otimes \cdots \otimes x_d\rangle| \le 3\varepsilon\right\}^{2^{-d}}.$$

4.1. A polynomial identity.

Our approach to decoupling is based on the following identity for real polynomials.

Lemma 4.2 (Polynomial identity)**.** *Let $d \ge 2$. For any real numbers $x_1, \dots, x_d$ and $x_1', \dots, x_d'$, we have*

$$\sum_{I \subset [d]} (-1)^{|I|} \Big(\sum_{i \in I^c} x_i + \sum_{i \in I} x_i' \Big)^d = d! \prod_{i=1}^d (x_i - x_i'). \tag{4.2}$$

The summation is over all 2^d subsets $I \subset [d]$.

To give a simple example, consider what happens for quadratic polynomials. For degree $d = 2$ the identity (4.2) becomes

$$(x_1 + x_2)^2 - (x_1 + x_2')^2 - (x_1' + x_2)^2 + (x_1' + x_2')^2 = 2(x_1 - x_1')(x_2 - x_2') \tag{4.3}$$

which is easy to check directly.

Proof. Expanding the right-hand side of (4.2), we see that it is equal to

$$d! \sum_{I \subset [d]} \prod_{i \in I^c} x_i \prod_{i \in I} (-x_i') = d! \sum_{I \subset [d]} (-1)^{|I|} \prod_{i \in I^c} x_i \prod_{i \in I} x_i'. \tag{4.4}$$

We claim that expanding the left side of (4.2) yields the same expression. For example, for $d = 2$ one can check directly that both sides of (4.3) are equal to

$$2(x_1 x_2 - x_1 x_2' - x_1' x_2 + x_1' x_2').$$

[3] This means that the the coordinates of x are independent $N(0,1)$ random variables.

For general d, let us fix I and see what happens when we expand the polynomial

$$\Big(\sum_{i \in I^c} x_i + \sum_{i \in I} x_i' \Big)^d$$

into a sum of monomials. The only multilinear monomial in this expansion is

$$d! \prod_{i \in I^c} x_i \prod_{i \in I} x_i'.$$

We claim that all other monomials cancel when we take the sum over I with signs $(-1)^{|I|}$ in (4.2). Indeed, any monomial that is not multilinear contains neither x_k nor x_k' for at least one k. Therefore, each such monomial in the left side (4.2) is canceled by another monomial, which is exactly the same except it has the opposite sign. These two monomials correspond to the sets I that differ from each other in exactly one member k. For example, in (4.3) the monomial x_2^2 coming from expanding $(x_1+x_2)^2$ gets canceled by a similar monomial coming from $(x_1'+x_2)^2$; these correspond to $I = \emptyset$ and $I = \{1\}$.

Thus, the left side of (4.2) equals (4.4). □

4.2. **Tensorization.** We would like to extend the polynomial identity in Lemma 4.2 to vectors $x_i, x_i' \in \mathbb{R}^n$. This is not possible for the tensor product since it is not commutative. We can circumvent this obstacle by considering the *symmetric tensor product* of d vectors in $\mathbb{R}^n$, which is defined by averaging their tensor products over all permutations:

$$x_1 \odot \cdots \odot x_d := \frac{1}{d!} \sum_{\sigma \in S_d} x_{\sigma(1)} \otimes \cdots \otimes x_{\sigma(d)}.$$

We will use the notation $x^{\odot d} = x \odot \cdots \odot x$ for multiplying a vector by itself d times.

Since the symmetric tensor product is commutative, the polynomial identity of Lemma 4.2 automatically generalizes to it, and we have

$$\sum_{I \subset [d]} (-1)^{|I|} \Big(\sum_{i \in I^c} x_i + \sum_{i \in I} x_i' \Big)^{\odot d} = d! \bigodot_{i=1}^{d} (x_i - x_i') \tag{4.5}$$

for any vectors $x_i, x_i' \in \mathbb{R}^n$. To check this, all we have to do is recall the second proof of Lemma 4.2, which is based on expanding the polynomials on both sides and comparing the resulting monomials. The exact same argument applies for the product $\odot$ since it is commutative.[4]

Furthermore, if $A \in \mathrm{Sym}^d(\mathbb{R}^n)$, then symmetry yields

$$\langle A, x_1 \otimes \cdots \otimes x_d \rangle = \langle A, x_{\sigma(1)} \otimes \cdots \otimes x_{\sigma(d)} \rangle$$

[4] In a more systematic way, one could invoke the equivalence between homogeneous polynomials of degree d and symmetric tensors of order d with the symmetric tensor product; see e.g. [23, Section 3.1].

for any permutation $\sigma \in S_d$ and any vectors $x_i \in \mathbb{R}^n$. By definition of the symmetric tensor product, this gives

$$\langle A, x_1 \otimes \cdots \otimes x_d \rangle = \langle A, x_1 \odot \cdots \odot x_d \rangle. \tag{4.6}$$

Let us take the inner product with A on both sides of (4.5). To simplify the resulting identity, use the linearity of the inner product and then invoke (4.6) to replace all symmetric products $\odot$ by the usual tensor products $\otimes$. This gives the following result.

Lemma 4.3. *Let $d \geq 2$ and $A \in \mathrm{Sym}^d(\mathbb{R}^n)$. Then for any vectors $x_1, \ldots, x_n \in \mathbb{R}^n$ we have*

$$\sum_{I \subset [d]} (-1)^{|I|} \Big\langle A, \Big(\sum_{i \in I^c} x_i + \sum_{i \in I} x_i' \Big)^{\otimes d} \Big\rangle = d! \Big\langle A, \bigotimes_{i=1}^{d} (x_i - x_i') \Big\rangle.$$

The summation is over all 2^d subsets $I \subset [d]$.

4.3. Conditionally independent events are positively correlated.

Let $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^m$ be random vectors, which are possibly correlated. Consider an event $F_{u,v}$ whose validity is determined by these vectors, i.e. an event that can be represented as

$$F_{u,v} = \{(u, v) \in B\}$$

where B is some Borel subset of $\mathbb{R}^{n+m}$. For instance, if $n = m = 1$, an example of such event is $F_{u,v} = \{u + v > 0\}$.

Let v' be an independent copy of v. Then the events $F_{u,v}$ and $F_{u,v'}$ are conditionally independent given u. The following elementary and known lemma (see [21, Lemma 14]) says that the events $F_{u,v}$ and $F_{u,v'}$ are always positively correlated. We include a proof for completeness.

Lemma 4.4 (Positive correlation)**.** *Consider an event $F_{u,v}$ whose validity is determined by a pair of random vectors u, v, and let v' be an independent copy of v. Then*

$$\mathbb{P}(F_{u,v})^2 \leq \mathbb{P}(F_{u,v} \cap F_{u,v'}).$$

Proof. By the law of total expectation, we have

$$\begin{aligned}
\mathbb{P}(F_{u,v} \cap F_{u,v'}) &= \mathbb{E}\, \mathbb{P}(F_{u,v} \cap F_{u,v'} | u) \quad \text{(by conditioning on } u\text{)} \\
&= \mathbb{E}\left[\mathbb{P}(F_{u,v} | u) \cdot \mathbb{P}(F_{u,v'} | u) \right] \quad \text{(by conditional independence)} \\
&= \mathbb{E}\left[\mathbb{P}(F_{u,v} | u)^2 \right] \quad \text{(by identical distribution)} \\
&= \left[\mathbb{E}\, \mathbb{P}(F_{u,v} | u) \right]^2 \quad \text{(by Cauchy-Schwarz inequality)} \\
&= \mathbb{P}(F_{u,v})^2.
\end{aligned}$$

The proof is complete. □

We would like to generalize Lemma 4.4 for multiple events. Such result is convenient to state in terms of *combinations*.

Definition 4.5 (Combinations)**.** *Given two lists of arbitrary objects $x = (x_1, \dots, x_d)$ and $x' = (x'_1, \dots, x'_d)$ and a subset of indices $I \subset [d]$, the combination $y = y(x, x', I)$ is the list whose elements y_i equal either x_i or x'_i depending on whether $i \in I$ or not. Precisely, we set*

$$y_i := \begin{cases} x_i, & i \in I^c \\ x'_i, & i \in I. \end{cases}$$

For example, if $x = (1, 2, 3)$, $x' = (a, b, c)$ and $I = \{1, 3\}$, then $y = (a, 2, c)$.

Lemma 4.6 (Positive correlation for multiple events)**.** *Consider an event E_x whose validity is determined by a random vector $x = (x_1, \dots, x_d)$, and let x' be an independent copy of x. Then*

$$\mathbb{P}(E_x)^{2^d} \le \mathbb{P}\Big(\bigcap_{I \subset [d]} E_{y(x,x',I)} \Big).$$

The intersection is over all 2^d subsets $I \subset [d]$.

Before we prove Lemma 4.6, let us mention that the partial case of this inequality where $d = 2$ is known [22, Lemma 4.7]. In this case, the lemma states that

$$\mathbb{P}(E_{x_1,x_2})^4 \le \mathbb{P}\big(E_{x_1,x_2} \cap E_{x_1,x'_2} \cap E_{x'_1,x_2} \cap E_{x'_1,x'_2}\big).$$

As is noted in [22], this partial case is easily seen to be equivalent to the following well known result in extreme combinatorics. If a bipartite graph connects n and m vertices and contains at least cnm edges for some $0 \le c \le 1$, then it must contain at least $c^4 n^2 m^2$ copies of the four-cycle C_4. (The count includes degenerate four-cycles, too.)

Proof of Lemma 4.6. We have

$$\mathbb{P}(E_x) = \mathbb{P}(E_{x_1,\dots,x_{d-1},x_d}) \le \mathbb{P}\big(E_{x_1,\dots,x_{d-1},x_d} \cap E_{x_1,\dots,x_{d-1},x'_d}\big)^{1/2}. \tag{4.7}$$

Here we applied Lemma 4.4 to "breed" x_d, thus we used it for $u = (x_1, \dots, x_{d-1})$ and $v = x_d$. Next, we apply Lemma 4.4 again to "breed" x_{d-1}, which bounds the right hand side of (4.7) by

$$\mathbb{P}\big(E_{x_1,\dots,x_{d-1},x_d} \cap E_{x_1,\dots,x_{d-1},x'_d} \cap E_{x_1,\dots,x'_{d-1},x'_d} \cap E_{x_1,\dots,x'_{d-1},x'_d}\big)^{1/4}.$$

Here we used Lemma 4.4 for the event $F_{u,v} = E_{x_1,\dots,x_{d-1},x_d} \cap E_{x_1,\dots,x_{d-1},x'_d}$ which is determined by the vectors $u = (x_1, \dots, x_{d-2}, x_d, x'_d)$ and $v = x_{d-1}$. Continuing in the same way to breed x_{d-2}, then x_{d-1}, and all the way down to x_1, we complete the proof. □

Remark 4.7. While we stated Lemma 4.6 for random vectors $x = (x_1, \dots, x_d)$, the same result holds if the coordinates x_k themselves are random vectors (rather than random variables). The proof is the same.

4.4. **Gaussian decoupling.** We are now ready to prove Theorem 4.1. Let x be a standard normal random vector in $\mathbb{R}^n$. By the rotation invariance of the normal distribution, x is distributed identically with

$$\frac{x_1 + \cdots + x_d}{\sqrt{d}},$$

where x_k are independent standard normal random vectors in $\mathbb{R}^n$. Thus we have

$$p := \mathbb{P}\left\{ |\langle A, x^{\otimes d}\rangle - u| \le \varepsilon \right\} = \mathbb{P}\left\{ |\langle A, (x_1 + \cdots + x_d)^{\otimes d}\rangle - u| \le d^{d/2}\varepsilon \right\}.$$

Denote the event in the right hand side by $E_{x_1,\ldots,x_d}$ and apply Lemma 4.6 (see also Remark 4.7). We conclude that

$$p^{2^d} \le \mathbb{P}\left\{ \forall I \subset [d] : |\langle A, (y_1 + \cdots + y_d)^{\otimes d}\rangle - u| \le d^{d/2}\varepsilon \right\}$$

where $y = y\left((x_1, \ldots, x_d), (x'_1, \ldots, x'_d), I\right)$ is the combination of the two lists of vectors. By definition of a combination, we can express this probability as

$$\mathbb{P}\left\{ \forall I \subset [d] : -d^{d/2}\varepsilon \le \Big\langle A, \Big(\sum_{i \in I^c} x_i + \sum_{i \in I} x'_i \Big)^{\otimes d} \Big\rangle - u \le d^{d/2}\varepsilon \right\}. \qquad (4.8)$$

Suppose the event in (4.8) holds. Multiply the middle part of the inequality in this event by $(-1)^{|I|}$ (the inequality will still hold), then take the sum over all 2^d subsets $I \subset [d]$ and finally apply Lemma 4.3. Since $\sum_{I \subset [d]} (-1)^{|I|} = 0$, the vector u disappears and we conclude that the probability in (4.8) is bounded by

$$\mathbb{P}\left\{ \forall I \subset [d] : 2^d d^{d/2}\varepsilon \le d! \Big\langle A, \bigotimes_{i=1}^{d} (x_i - x'_i) \Big\rangle \le 2^d d^{d/2}\varepsilon \right\}. \qquad (4.9)$$

By the rotation invariance of the normal distribution, the random vectors $x_i - x'_i$ are jointly distributed in the same way as $\sqrt{2}\, x_i$. Thus, dividing all sides of the inequality by $d!(\sqrt{2})^d$, we see that the probability in (4.9) is equal to

$$\mathbb{P}\left\{ \forall I \subset [d] : |\langle A, x_1 \otimes \cdots \otimes x_d\rangle| \le (2d)^{d/2}(d!)^{-1}\varepsilon \right\}.$$

To finish the proof of Theorem 4.1, use the inequality $d! \ge (d/e)^d$ to check that $(2d)^{d/2}(d!)^{-1} \le 3$ for all $d \ge 2$. □

4.5. **General decoupling.** Next, we extend Theorem 4.1 to general, possibly non-Gaussian, distributions.

Theorem 4.8 (Decoupling small ball probabilities for random tensors)**.** *Let $d \ge 2$, $A \in \mathrm{Sym}^d(\mathbb{R}^n)$ and let x be a random vector taking values in $\mathbb{R}^n$. Suppose we can represent x as*

$$x = x_1 + \cdots + x_d$$

where x_i are independent random vectors. Then for any $u \in \mathbb{R}$ and $\varepsilon \geq 0$ we have

$$\mathbb{P}\left\{|\langle A, x \otimes \cdots \otimes x\rangle - u| \leq \varepsilon\right\} \leq \mathbb{P}\left\{|\langle A, y_1 \otimes \cdots \otimes y_d\rangle| \leq 2\varepsilon\right\}^{2^{-d}}, \quad (4.10)$$

where $y_i = x_i - x_i'$ and x_i' are independent copies of x_i.

Proof. We simply follow the proof of Theorem 4.1 except in the two places where rotation invariance was used. This argument shows that the left hand side of (4.10) is bounded by

$$\mathbb{P}\left\{|\langle A, (x_1 - x_1') \otimes \cdots \otimes (x_d - x_d')\rangle| \leq 2^d (d!)^{-1} \varepsilon\right\}^{2^{-d}}.$$

This is slightly stronger than the desired conclusion, since $2^d (d!)^{-1} \leq 2$ for all $d \in \mathbb{N}$. □

Remark 4.9 (Uniform decoupling inequalities). The decoupling inequalities we developed here are stated for a fixed tensor $A \in \mathrm{Sym}^d(\mathbb{R}^n)$. Occasionally one needs to work with multiple tensors simultaneously. In particular, later in this paper we will need to control all tensors A in a given subspace of $\mathrm{Sym}^d(\mathbb{R}^n)$ at once. Fortunately, our proofs of decoupling inequalities can be trivially extended to multiple tensors, as follows. Let $\mathcal{A} \subset \mathrm{Sym}^d(\mathbb{R}^n)$ be an arbitrary fixed subset. Then the conclusion of Theorem 4.8 changes to

$$\begin{aligned} &\mathbb{P}\left\{\forall A \in \mathcal{A}:\ |\langle A, x \otimes \cdots \otimes x\rangle - u| \leq \varepsilon\right\} \\ &\qquad\qquad \leq \mathbb{P}\left\{\forall A \in \mathcal{A}:\ |\langle A, y_1 \otimes \cdots \otimes y_d\rangle| \leq 2\varepsilon\right\}^{2^{-d}}. \end{aligned}$$

The proof holds with obvious changes.[5] A similar extension can be stated for Theorem 4.1.

5. Random contractions of tensor subspaces

Let us take a close look at the operation of tensor-vector multiplication, which we introduced in Section 3.3.3. For a given vector x, the multiplication

$$A \mapsto Ax$$

is a linear transformation that maps $\mathbb{R}^{n_1 \times \cdots \times n_d}$ to $\mathbb{R}^{n_1 \times \cdots \times n_{d-1}}$, thus contracting the dimension. We may wonder how this contraction affects various subsets. The crucial question for the purpose of our paper is what the contraction does to a given *linear subspace* $E \subset \mathbb{R}^{n_1 \times \cdots \times n_d}$. Of course, the kernel of the contraction can be huge and E may fall entirely in it, thus E may collapse to zero under the contraction. However, if we choose the multiplier x *at random*, the dramatic collapse is unlikely to happen, and a lot of E may survive the contraction. We are going to state and prove this fact now.

[5] All one needs to do is to include $\forall A \in \mathcal{A}$ in the events throughout the proof. Note that no measurability issues arise here. Indeed, the set of tensors $Z \in \mathrm{Sym}^d(\mathbb{R}^n)$ that satisfy $|\langle A, Z\rangle - u| \leq \varepsilon$ for all $A \in \mathcal{A}$ is a closed convex set and thus automatically measurable.

The survival phenomenon is non-trivial even for tensors of rank $d = 2$, i.e. for matrices, and we shall examine this case first. Given a subspace of matrices $E \subset \mathbb{R}^{m \times n}$ and a random vector x taking values in $\mathbb{R}^n$, we are interested in the dimension of the subspace $Ex \subset \mathbb{R}^m$ defined by

$$Ex := \{Ax : \; A \in E\}.$$

Theorem 5.1 (Random contractions of matrix subspaces)**.** *Consider a subspace $E \subset \mathbb{R}^{m \times n}$, and let y be a random vector uniformly distributed in $\{-1, 1\}^n$. Fix $\varepsilon \in (0, 1)$. Then*

$$\dim(E) \geq \varepsilon mn \quad \textit{implies} \quad \dim(Ey) \geq \frac{\varepsilon m}{2}$$

with probability at least $1 - 2\exp(-c\varepsilon n)$.

Before we pass to the proof of Theorem 5.1, let us pause quickly to explain why its conclusion is nearly optimal. Consider the subspace E consisting of the $m \times n$ matrices whose first εm rows can be arbitrary and the last $(1 - \varepsilon)m$ rows are completely zero. (For simplicity, assume that εm is an integer.) Then, for any vector $y \in \{-1, 1\}^n$, the subspace Ey consists of the vectors in $\mathbb{R}^m$ whose first εm coordinates are arbitrary and the last $(1-\varepsilon)m$ coordinates are zero. This shows that

$$\dim(E) = \varepsilon mn \quad \text{and} \quad \dim(Ey) = \varepsilon m.$$

Therefore, the bound on the dimension of Ey we get in Theorem 5.1 is optimal up to a factor of 2 (which can be improved by a more careful analysis).

5.1. **Dimension identities.**

To prepare for the proof of Theorem 5.1, let us fix y and look for a more convenient expression for the dimension of Ey. Consider the matrix subspace

$$R_y := \{A \in \mathbb{R}^{m \times n} : \; Ay = 0\}.$$

Lemma 5.2. *We have*

$$\dim(Ey) = \dim(E) - \dim(E \cap R_y).$$

Proof. Consider the map $A \mapsto Ay$ as a linear map from E to $\mathbb{R}^m$. Its image is Ey and the kernel is $E \cap R_y$. Then the rank-nullity theorem yields

$$\dim(E) = \dim(Ey) + \dim(E \cap R_y).$$

This implies the conclusion of the lemma. □

The quantity $\dim(E) - \dim(E \cap R_y)$ is still not very convenient for the analysis. Let us try to rewrite it in terms of the orthogonal complements of the subspaces. The following lemma will help us.

Lemma 5.3. *For any subspaces $U, V \subset \mathbb{R}^N$, we have*

$$\dim(U) - \dim(U \cap V) = \dim(V^\perp) - \dim(V^\perp - U^\perp).$$

Proof. A version of the inclusion-exclusion principle for the dimensions of subspaces states that

$$\dim(U+V) = \dim(U) + \dim(V) - \dim(U \cap V), \tag{5.1}$$

see e.g. [57, p.22]. On the other hand, we have

$$\dim(U+V) = N - \dim((U+V)^\perp) = \dim(U^\perp \cap V^\perp)$$

and $\dim(V) = N - \dim(V^\perp)$. Substitute these two identities into (5.1) and simplify to complete the proof. □

Using Lemmas 5.2 and 5.3, we can express the dimension of Ey as follows:

$$\dim(Ey) = \dim(R_y^\perp) - \dim(R_y^\perp \cap E^\perp). \tag{5.2}$$

5.2. **Bases of subspaces.** Let us take a closer look at the subspaces $R_y^\perp$ and $R_y^\perp \cap E^\perp$. The subspace $R_y^\perp$ is very simple; it consists of rank-one matrices of the form xy^T:

Lemma 5.4. *We have*

$$R_y^\perp = \{xy^\perp : \; x \in \mathbb{R}^m\}.$$

Proof. If $A \in R_y$ then $Ay = 0$ and thus $\langle A, xy^\perp \rangle = \langle Ay, x \rangle = 0$. This shows that

$$\{xy^\perp : \; x \in \mathbb{R}^m\} \subset R_y^\perp. \tag{5.3}$$

To complete the proof, suppose for contradiction that the inclusion in (5.3) is strict, i.e. $\{xy^\perp : \; x \in \mathbb{R}^m\}$ is a proper subspace of $R_y^\perp$. Then there exists a nonzero $A \in R_y^\perp$ that is orthogonal to that subspace, which yields

$$0 = \langle A, xy^\mathsf{T} \rangle = \langle Ay, x \rangle$$

for every $x \in \mathbb{R}^m$. This implies that $Ay = 0$, which means that $A \in R_y$. But we assumed that $A \in R_y^\perp$, so this means that $A = 0$. This contradiction completes the proof. □

Lemma 5.4 implies that if $e_1, \ldots, e_m$ denotes the standard basis of $\mathbb{R}^m$, then $e_1 y^\perp, \ldots, e_m y^\perp$ is a basis of $R_y^\perp$. In particular, if follows that

$$\dim(R_y^\perp) = m. \tag{5.4}$$

In a similar way, we can find a basis of $R_y^\perp \cap E^\perp$.

Lemma 5.5. *The subspace $R_y^\perp \cap E^\perp$ has a basis of the form $x_i y^\mathsf{T}$ where $x_i \in \mathbb{R}^m$ are linearly independent vectors.*

Proof. According to Lemma 5.4, all elements of $R_y^\perp$ and thus also of $R_y^\perp \cap E^\perp$ have the form xy^T. Thus there exists a basis of $R_y^\perp \cap E^\perp$ of the form $x_i y^\mathsf{T}$. Linear independence of $x_i y^\mathsf{T}$ trivially implies linear independence of x_i. □

We are almost ready to prove Theorem 5.1. But before we do this, let us check one more helpful fact.

Lemma 5.6. *Let $(y_j)_{j\in J}$ be a set of linearly independent vectors in $\mathbb{R}^n$. For each $j \in J$, let $(x_{ij})_{i\in I_j}$ be a set of linearly independent vectors in $\mathbb{R}^m$. Then the set of matrices $(x_{ij}y_j^\mathsf{T})_{j\in J,\, i\in I_j}$ is linearly independent.*

Proof. Suppose some linear combination of $x_{ij}y_j^\mathsf{T}$ vanishes, i.e.

$$0 = \sum_{ij} a_{ij}x_{ij}y_j^\mathsf{T} = \sum_j \Big(\sum_i a_{ij}x_{ij}\Big)y_j^\mathsf{T}.$$

We want to show that all coefficients a_{ij} must vanish.

Fix any vector $\theta \in \mathbb{R}^m$, multiply the equation on the left by $\theta^\perp$, and simplify. This gives

$$0 = \sum_j \Big(\sum_i a_{ij}\langle x_{ij}, \theta\rangle\Big)y_j^\mathsf{T}.$$

Linear independence of y_i implies that for each j we have

$$0 = \sum_i a_{ij}\langle x_{ij}, \theta\rangle = \Big\langle \sum_i a_{ij}x_{ij}, \theta\Big\rangle$$

Since this holds for arbitrary θ, it follows that

$$\sum_i a_{ij}x_{ij} = 0.$$

Finally, linear independence of x_{ij} for each j implies that all $a_{ij} = 0$. The proof is complete. □

5.3. **Proof of Theorem 5.1.** Fix a subspace $E \subset \mathbb{R}^{m\times n}$ such that

$$\dim(E) \geq \varepsilon mn. \tag{5.5}$$

Suppose the conclusion of Theorem 5.1 fails. Thus

$$\mathbb{P}\Big\{\dim(Ey) < \frac{\varepsilon m}{2}\Big\} > 2\exp(-c\varepsilon n).$$

Recall that by (5.2) and (5.4), $\dim(Ey) = m - \dim(R_y^\perp \cap E^\perp)$. Thus

$$\mathbb{P}\Big\{\dim(R_y^\perp \cap E^\perp) > \Big(1 - \frac{\varepsilon}{2}\Big)m\Big\} > 2\exp(-c\varepsilon n). \tag{5.6}$$

Define the (non-random) subspace $L \subset \mathbb{R}^n$ as the linear span of all vectors that define the event in (5.6), i.e. we set

$$L := \operatorname{span}\Big\{v \in \mathbb{R}^n :\ \dim(R_v^\perp \cap E^\perp) > \Big(1 - \frac{\varepsilon}{2}\Big)m\Big\}.$$

Then the bound (5.6) trivially yields

$$\mathbb{P}\{y \in L\} > 2\exp(-c\varepsilon n). \tag{5.7}$$

This in turn implies that

$$\dim(L) > \Big(1 - \frac{\varepsilon}{2}\Big)n. \tag{5.8}$$

Indeed, if this were not the case and $\dim(L) \le (1-\varepsilon/2)n$, then Lemma 3.8 would give $\mathbb{P}\{y \in L\} \le 2\exp(-c\varepsilon n)$, which would contradict (5.7), provided the absolute constant $c > 0$ is chosen appropriately.

Due to (5.8), we can find in L more than $(1-\varepsilon/2)n$ linearly independent vectors y_j. By definition of L, each of these vectors satisfies

$$\dim(R_{y_j}^{\perp} \cap E^{\perp}) > \Big(1 - \frac{\varepsilon}{2}\Big)m.$$

For each j, we can use Lemma 5.5 to find more than $(1-\varepsilon/2)m$ linearly independent vectors $x_{ij} \in \mathbb{R}^m$ such that

$$x_{ij}y_j^{\mathsf{T}} \in R_{y_j}^{\perp} \cap E^{\perp}. \tag{5.9}$$

Now consider the set of matrices $x_{ij}y_j^{\mathsf{T}}$ we constructed for all i and j. This set is linearly independent by Lemma 5.6, it is contained in $E^{\perp}$ by (5.9), and its cardinality is larger than $(1-\varepsilon/2)n \cdot (1-\varepsilon/2)m$ by construction. This implies that

$$\dim(E^{\perp}) > (1-\varepsilon/2)n \cdot (1-\varepsilon/2)m \ge (1-\varepsilon)mn.$$

But this contradicts our assumption (5.5). The proof of Theorem 5.1 is complete. □

5.4. **Random contractions of tensor subspaces.** Theorem 5.1 on matrix contractions can be easily extended to tensors, with the canonical tensor-vector multiplication defined in Section 3.3.3. Given a subspace $E \subset \mathbb{R}^{n_1 \times \cdots \times n_d}$ and a random vector x taking values in $\mathbb{R}^{n_d}$, we are interested in the dimension of the subspace $E_x \subset \mathbb{R}^{n_1 \times \cdots \times n_{d-1}}$ defined by

$$Ex := \{Ax : \; A \in E\}.$$

Theorem 5.7 (Random contractions of tensor subspaces)**.** *Let $d \ge 2$. Consider a subspace $E \subset \mathbb{R}^{n_1 \times \cdots \times n_d}$, and let x be a random vector uniformly distributed in $\{-1,1\}^{n_d}$. Fix $\varepsilon \in (0,1)$. Then*

$$\dim(E) \ge \varepsilon n_1 \ldots n_d \quad \textit{implies} \quad \dim(Ex) \ge \frac{\varepsilon n_1 \cdots n_{d-1}}{2}$$

with probability at least $1 - \exp(-c\varepsilon n_d)$.

Proof. We can reshape a tensor $A \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ into an $m \times n_d$ matrix $\bar{A}$, where $m = n_1 \cdots n_{d-1}$, by vectorizing all $n_1 \times \cdots \times n_{d-1}$ fibers of A. It is easy to check that such reshaping converts the tensor-matrix multiplication into the matrix-vector multiplication. Moreover, because the reshaping is an isomorphism, it preserves the dimension of any subspaces as well. Thus, all we have to do in order to complete the proof is apply Theorem 5.1 for subspaces consisting of $m \times n_d$ matrices. □

Finally, we can extend our results to contractions of many dimensions. In the results below, we continue to use the canonical tensor-vector multiplication defined in Section 3.3.3. Consider a subspace $E \subset \mathbb{R}^{n_1 \times \cdots \times n_d}$ and random vectors $x_2 \in \mathbb{R}^{n_2}, \ldots, x_d \in \mathbb{R}^{n_d}$. We are interested in the dimension of the subspace $Ex_d \cdots x_2 \subset \mathbb{R}^{n_1}$ defined as

$$Ex_d \cdots x_2 := \{Ax_d \cdots x_2 : \; A \in E\}. \tag{5.10}$$

Corollary 5.8 (Random contractions of tensor subspaces)**.** *Let $d \geq 2$. Consider a subspace $E \subset \mathbb{R}^{n_1 \times \cdots \times n_d}$, and let x_k be independent random vectors uniformly distributed in $\{-1,1\}^{n_k}$ for $k = 2, \ldots, d$. Fix $\varepsilon \in (0,1)$. Then*

$$\dim(E) \geq \varepsilon n_1 \ldots n_d \quad \textit{implies} \quad \dim(Ex_d \cdots x_2) \geq \frac{\varepsilon n_1}{2^{d-1}}$$

with probability at least $1 - 2\exp(-c\varepsilon n_0/2^d)$, where $n_0 = \min(n_2, \ldots, n_d)$.

Proof. Apply Theorem 5.7 for $x = x_d$. It states that

$$\dim(Ex_d) \geq \frac{\varepsilon n_1 \cdots n_{d-1}}{2}$$

with probability at least $1 - \exp(-c'\varepsilon n_d)$. We now condition on x_d that satisfies this inequality and apply Theorem 5.7 for the subspace Ex_d instead of E, for $\varepsilon/2$ instead of ε, and for $x = x_{d-1}$. We get

$$\dim(Ex_d x_{d-1}) \geq \frac{\varepsilon n_1 \cdots n_{d-2}}{4}$$

with conditional probability at least $1 - \exp(-c'\varepsilon n_{d-1}/2)$. We now condition on x_{d-1} that satisfies this inequality, and continue as above. After $d-1$ applications of Theorem 5.7, we get

$$\dim(Ex_d \cdots x_2) \geq \frac{\varepsilon n_1}{2^{d-1}} \tag{5.11}$$

with conditional probability at least $1 - 2\exp(-c'\varepsilon n_2/2^{d-2})$. Summing up the failure probabilities for each step, we conclude that (5.11) holds with (unconditional) probability of at least

$$1 - \sum_{k=0}^{d-2} 2\exp\Big(-\frac{c'\varepsilon n_{d-k}}{2^k}\Big) \geq 1 - 2\exp\Big(-\frac{c\varepsilon n_0}{2^d}\Big).$$

The proof is complete. □

Remark 5.9 (General distributions). Like almost everywhere else in this paper, all the results of this section can be trivially extended to general distributions of x_k. Indeed, the randomness of x_k was used just once in the argument, namely where we applied Lemma 3.8 in the proof of Theorem 5.1. As we noted in Section 3.6, this lemma holds for general mean zero sub-gaussian random vectors x_k. Later in this paper, this observation will become useful as we will apply the contraction results to the random vectors $x_k - x'_k$, instead of x_k, where x'_k are independent copies of x_k generated as a byproduct of decoupling.

6. Restrictions of tensor subspaces

In the previous section we examined a certain canonical operation on tensors, namely tensor-vector multiplication. We looked at it as a dimension-reduction map, and we studied how this operation affects the dimension of tensor subspaces. In this section we examine another natural dimension-reduction operation, namely the restriction of tensors onto a given subset of indices. Consider a symmetric tensor $A = (A_{i_1,\ldots,i_d})$ in $\mathrm{Sym}^d(\mathbb{R}^n)$. Imagine that we restrict each index i_k to lie in a given subset $I_k \subset [n]$, where the subsets I_k are disjoint and form a partition of $[n]$. Such restriction is a linear transformation that maps a tensor A into a different tensor, not symmetric anymore, and with smaller dimensions. The main question we are going to address in this section is the following: given a subspace $E \subset \mathrm{Sym}^d(\mathbb{R}^n)$, can we always find a restriction (i.e. the partition of $[n]$ into sets I_k) that approximately preserves the dimension of E?

To provide a formal definition of tensor restriction, for technical reasons it is more convenient for us to assume that a restriction zeroes out, rather than drops, the entries of A whose indices fall outside I_k.

Definition 6.1 (Restrictions of tensors)**.** *Consider a tensor $A \in (\mathbb{R}^n)^{\otimes d} = \mathbb{R}^{n\times\cdots\times n}$ and a subset of multi-indices $S \subset [n]^d$. The* restriction of A onto S *is the tensor $A_S \in (\mathbb{R}^n)^{\otimes d}$ obtained from A by zeroing the coordinates that fall outside S.*

Let E be a subspace of $(\mathbb{R}^n)^{\otimes d}$. The *restriction of E onto S* is the subspace of $(\mathbb{R}^n)^{\otimes d}$ obtained by restricting every tensor of E onto S, i.e.

$$E_S := \{A_S : \ A \in E\}.$$

Theorem 6.2 (Restrictions of tensor subspaces)**.** *Let $d \ge 2$. Consider a subspace $E \subset \mathrm{Sym}^d(\mathbb{R}^n)$, and assume that $\varepsilon \in (0,1)$ is such that*

$$\dim(E) \ge \varepsilon n^d.$$

Then there exists a decomposition $[n] = I_1 \cup \cdots \cup I_d$ where $|I_k| \ge \varepsilon n/4$ for all k, and such that

$$\dim(E_{I_1\times\cdots\times I_d}) \ge \Big(\frac{\varepsilon n}{4}\Big)^d.$$

Our proof of Theorem 6.2 is based on a pigeonhole principle. Let us call a multi-index $(i_1,\ldots,i_d) \in [n]^d$ *ordered* if it consists of non-decreasing indices, i.e.

$$i_1 \le i_2 \le \cdots \le i_d.$$

Fix a real number $\delta \in (10/n, 1)$ whose value we will chose later. We say that an ordered multi-index $(i_1,\ldots,i_d)$ is *δn-separated* if all of its indices are more than δn apart, i.e.

$$i_{k+1} - i_k > \delta n \quad \text{for each } k = 1,\ldots,d-1.$$

Lemma 6.3 (Most multi-indices are separated)**.** *Denote by O the set of ordered multi-indices and by S the subset of δn-separated ordered multi-indices. Then*

$$|O \setminus S| \le \frac{\delta n^d}{(d-2)!}.$$

Proof. If an ordered multi-index $(i_1, \dots, i_d)$ is not δn-separated, there exists a $k_0 \in \{1, \dots, d-1\}$ such that

$$i_{k_0+1} - i_{k_0} \le \delta n. \tag{6.1}$$

There are $d-1$ ways to choose k_0. Once k_0 is chosen, there are at most $\binom{n}{d-1}$ ways to choose the indices i_k for $k \ne k_0+1$. Finally, once these are chosen, there are δn ways to choose i_{k_0+1}, since it must satisfy (6.1). Thus, the total number of ways to choose a multi-index from $O \setminus S$ is at most

$$(d-1) \cdot \binom{n}{d-1} \cdot \delta n \le \frac{\delta n^d}{(d-2)!}.$$

The proof is complete. □

We call a product set $I_1 \times \cdots \times I_d \subset [n]^d$ a *grid box* if $I_k \subset [n]$ are consecutive non-empty intervals whose lengths[6] are multiples of $\delta n/2$, and which form a decomposition of $[n]$:

$$[n] = I_1 \cup I_2 \cup \ldots \cup I_d. \tag{6.2}$$

Lemma 6.4. *Every δn-separated ordered multi-index belongs to some grid box.*

Proof. Let $(i_1, \dots, i_d)$ be a δn-separated ordered multi-index. Thus

$$1 \le i_1 \le i_2 \le i_3 \le \cdots \le i_d \le n$$

and $i_{k+1} - i_k > \delta n > 10$ for every $k = 1, \dots, d-1$. Then we can find a sequence j_k that strictly interleaves with the sequence i_k, i.e.

$$1 \le i_1 < j_1 < i_2 < j_2 < i_3 < \cdots < i_{d-1} < j_{d-1} < i_d \le n.$$

Because of the separation, we can arrange for j_k to be multiples of $\delta n/2$, and so that $n - j_{d-1} > \delta n/2$. Define $I_1 = [1, j_1]$, $I_2 = (j_1, j_2]$, $\dots$, $I_d = (j_{d-1}, n]$. Then $I_1 \times \cdots \times I_d$ is a grid box which contains the multi-index $(i_1, \dots, i_d)$. □

Lemma 6.5 (There are few grid boxes)**.** *Let $\mathcal{G}$ denote the family of all grid boxes. Then*

$$|\mathcal{G}| \le (2/\delta)^{d-1}.$$

Proof. Any grid box can be identified with the decomposition (6.2) of $[n]$ into d consecutive intervals whose endpoints are multiples of $\delta n/2$. Such decomposition is determined by the $d-1$ break points – the endpoints of the intervals I_k that are different from 1 or n. Each break point must be a multiple of $\delta n/2$. Thus, each endpoint can be chosen in at most $n/(\delta n/2) =$

[6] For simplicity of presentation, we assume that $\delta n/2$ is an integer.

$2/\delta$ ways. Thus, the total number of ways to choose all break points, and thus a grid box, is at most $(2/\delta)^{d-1}$. □

The following elementary lemma will help us set up a version of the pigeonhole principle for dimension counting.

Lemma 6.6 (Dimension and covering)**.** *Let E be a subspace of $(\mathbb{R}^n)^{\otimes d}$ and $S \subset [n]^d$ be a subset of multi-indices. Then:*

1. $\dim(E_S) \le |S|$.
2. *Suppose S is covered by a family of subsets $S_1, \ldots, S_N \subset [n]^d$, i.e. $S \subset S_1 \cup \cdots \cup S_N$. Then*
$$\dim(E_S) \le \dim(E_{S_1}) + \cdots + \dim(E_{S_N}).$$

Proof. The first part is trivial since E_S lies in $\mathbb{R}^S$, the space of tensors supported on S. To verify the second part, note that we can represent each tensor from E_S as a sum of tensors from $E_{S_1}, \ldots, E_{S_N}$. Thus, the subspace E_S lies in the direct sum of the subspaces $E_{S_1}, \ldots, E_{S_N}$. This yields the dimension bound. □

Proof of Theorem 6.2. We continue to use the notation of the Lemma 6.3, calling O the set of ordered multi-indices and S the subset of δn-separated ordered multi-indices. We have

$$\varepsilon n^d \le \dim(E) = \dim(E_O) \le \dim(E_{O \setminus S}) + \dim(E_S). \tag{6.3}$$

The first bound is the assumption of the theorem, the equality is due to the symmetry of the tensors, and the last bound follows from part 2 of Lemma 6.6 and the decomposition $O = (O \setminus S) \cup S$.

By part 1 of Lemma 6.6 and Lemma 6.3, we have

$$\dim(E_{O \setminus S}) \le |O \setminus S| \le \frac{\delta n^d}{(d-2)!}.$$

If we choose $\delta := \varepsilon/2$, this bound yields $\dim(E_{O\setminus S}) \le \varepsilon n^d/2$. Putting it into (6.3), we get

$$\dim(E_S) \ge \frac{\varepsilon n^d}{2}.$$

Lemma 6.4 states that

$$S \subset \bigcup_{B \in \mathcal{G}} B$$

where the union is over all grid boxes $B = I_1 \times \cdots \times I_d$, whose family we denoted by $\mathcal{G}$ in Lemma 6.5. Using part 2 of Lemma 6.6, we get

$$\frac{\varepsilon n^d}{2} \le \dim(E_S) \le \sum_{B \in \mathcal{G}} \dim(E_B).$$

By the pigeonhole principle, there exists a grid box $B \in \mathcal{G}$ such that

$$\dim(E_B) \ge \frac{\varepsilon n^d}{2|\mathcal{G}|} \ge \frac{\varepsilon n^d}{(2/\delta)^{d-1}} \ge \Big(\frac{\varepsilon n}{4}\Big)^d.$$

In the second inequality we used the bound on $|\mathcal{G}|$ given by Lemma 6.5, and in the third, the choice of the value $\delta = \varepsilon/2$ we made above. This completes the proof of Theorem 6.2. □

7. Linear independence of random tensors

In this section we use all the tools developed earlier – decoupling, contractions, and restrictions for random tensors – to prove a crucial result about linear independence of random tensors. We announced a simplified version of this result in Theorem 2.2, and we will prove a more precise version in Theorem 7.3 below.

7.1. A random tensor is unlikely to fall in a subspace.

We start by showing that a random tensor of order $d \geq 1$ is unlikely to fall in a fixed tensor subspace. In the partial case of random vectors ($d = 1$), this fact is well known as was proved in Lemma 3.8 for the case of vectors. For random tensors of any higher order $d > 1$, and in particular for random matrices ($d = 2$), this fact is new and non-trivial. We prove two versions: first for non-symmetric tensors and then for symmetric tensors.

Proposition 7.1 (A random tensor is unlikely to fall in a subspace)**.** *Let $d \geq 1$. Consider a subspace $L \subset \mathbb{R}^{n_1 \times \cdots \times n_d}$, and assume that $\varepsilon \in (0, 1)$ is such that*

$$\operatorname{codim}(L) \geq \varepsilon n_1 \cdots n_d.$$

Let x_k be independent random vectors uniformly distributed in $\{-1, 1\}^{n_k}$ where $k = 1, \ldots, d$. Then

$$\mathbb{P}\left\{x_1 \otimes \cdots \otimes x_d \in L\right\} \leq 4 \exp\Big(-\frac{c\varepsilon n_0}{2^d}\Big)$$

where $n_0 = \min(n_1, \ldots, n_d)$.

Proof. As we noted, for $d = 1$ the proposition reduces to Lemma 3.8, so we can assume that $d \geq 2$. The orthogonal complement $E := L^\perp$ has dimension

$$\dim(E) \geq \varepsilon n_1 \cdots n_d. \tag{7.1}$$

Suppose the event $x_1 \otimes \cdots \otimes x_d \in L$ happens. Then $\langle A, x_1 \otimes \cdots \otimes x_d \rangle = 0$ for all $A \in E$. By the cyclic identity (3.7), this is the same as $\langle A x_d x_{d-1} \cdots x_2, x_1 \rangle = 0$ for all $A \in E$. This in turn is equivalent to $x_1 \in (E x_d x_{d-1} \cdots x_2)^\perp$ where we use the notion of contraction of subspaces introduced in (5.10). Summarizing, we have

$$p := \mathbb{P}\left\{x_1 \otimes \cdots \otimes x_d \in L\right\} \leq \mathbb{P}\left\{x_1 \in (E x_d x_{d-1} \cdots x_2)^\perp\right\}.$$

In view of (7.1), Corollary 5.8 gives

$$\dim(E x_d \cdots x_2) \geq \frac{\varepsilon n_1}{2^d} \tag{7.2}$$

with probability at least $1 - \exp(-c\varepsilon n_0/2^d)$. We now condition on a realization of $x_2, \dots, x_d$ satisfying (7.7) and apply Lemma 3.8 for the random vector $x = x_1$ and fixed subspace $F := (Ex_d x_{d-1} \cdots x_2)^\perp$. This gives

$$\mathbb{P}\left\{x_1 \in (Ex_d x_{d-1} \cdots x_2)^\perp \mid x_2, \dots, x_d\right\} \le 2\exp\Big(-\frac{c\varepsilon n_1}{2^d}\Big).$$

Then the (unconditional) probability satisfies

$$\begin{aligned} p &\le \mathbb{P}\left\{x_1 \in (Ex_d x_{d-1} \cdots x_2)^\perp\right\} \\ &\le 2\exp\Big(-\frac{c\varepsilon n_0}{2^d}\Big) + 2\exp\Big(-\frac{c\varepsilon n_1}{2^d}\Big) \le 4\exp\Big(-\frac{c\varepsilon n_0}{2^d}\Big). \end{aligned}$$

The proposition is proved. □

Proposition 7.2 (A symmetric random tensor is unlikely to fall in a subspace)**.** *Let $d \ge 1$. Consider a subspace $L \subset \mathrm{Sym}^d(\mathbb{R}^n)$, and assume that $\varepsilon \in (0,1)$ is such that*

$$\mathrm{codim}(L) \ge \varepsilon n^d.$$

Let x be a random vector uniformly distributed in $\{-1,1\}^n$. Then

$$\mathbb{P}\left\{x^{\otimes d} \in L\right\} \le 4\exp\left(-(c\varepsilon)^{d+1}\, n\right).$$

Proof. **Step 1: Restriction.** The orthogonal complement $E := L^\perp$ has dimension

$$\dim(E) \ge \varepsilon n^d. \tag{7.3}$$

Rewrite the probability in question as

$$p := \mathbb{P}\left\{x^{\otimes d} \in L\right\} = \mathbb{P}\left\{\forall A \in E :\ \langle A, x^{\otimes d}\rangle = 0\right\}.$$

Theorem 6.2 provides for us a convenient restriction of the subspace E. It says that there exists a decomposition $[n] = I_1 \cup \cdots \cup I_d$ where

$$n_k := |I_k| \ge \frac{\varepsilon n}{4} \quad \text{for all } k, \tag{7.4}$$

and such that

$$\dim(E_{I_1 \times \cdots \times I_d}) \ge \Big(\frac{\varepsilon n}{4}\Big)^d. \tag{7.5}$$

Step 2: Decoupling. Next we apply the decoupling inequality for small ball probabilities (Theorem 4.8 and Remark 4.9). We do this for the natural decomposition

$$x = x_{I_1} + \cdots + x_{I_d},$$

where x_{I_k} denotes the vector obtained from x by zeroing the coordinates outside I_k. This gives

$$p^{2^d} \le \mathbb{P}\left\{\forall A \in E :\ \langle A, y_{I_1} \otimes \cdots \otimes y_{I_d}\rangle = 0\right\}$$

where

$$y = x - x'$$

and x' is an independent copy of x.

Denote $\bar{A} \in \mathbb{R}^{n_1 \times \cdots n_d}$ the matrix obtained from A by considering only the entries indexed by $I_1 \times \cdots \times I_d$, and denote by $y_k \in \mathbb{R}^{n_k}$ the vectors obtained from y by considering only the coefficients indexed by I_k.[7] Note that y_k are independent random vectors all of whose coordinates y_{kj} are independent and have distribution

$$\mathbb{P}\left\{y_{kj} = -2\right\} = \frac{1}{4}, \quad \mathbb{P}\left\{y_{kj} = 0\right\} = \frac{1}{2}, \quad \mathbb{P}\left\{y_{kj} = 2\right\} = \frac{1}{4}.$$

The definition (3.5) of the inner product of tensors obviously gives

$$\langle A, y_{I_1} \otimes \cdots \otimes y_{I_d} \rangle = \langle \bar{A}, y_1 \otimes \cdots \otimes y_d \rangle.$$

Thus

$$p^{2^d} \le \mathbb{P}\left\{\forall \bar{A} \in \bar{E} : \; \langle \bar{A}, y_1 \otimes \cdots \otimes y_d \rangle = 0\right\},$$

where $\bar{E}$ consists of all matrices $\bar{A} \in \mathbb{R}^{n_1 \times \cdots n_d}$ obtained from $A \in E$ as we explained above. Obviously[8] the subspace $\bar{E}$ is isomorphic to $E_{I_1 \times \cdots \times I_d}$, so (7.5) gives

$$\dim(\bar{E}) \ge \left(\frac{\varepsilon n}{4}\right)^d \ge \left(\frac{\varepsilon}{4}\right)^d n_1 \cdots n_d. \tag{7.6}$$

From now on, we can essentially follow the proof of Proposition 7.1. Suppose $\langle \bar{A}, y_1 \otimes \cdots \otimes y_d \rangle = 0$ for all $\bar{A} \in \bar{E}$. By (3.7), this is the same as $\langle \bar{A} y_d y_{d-1} \cdots y_2, y_1 \rangle = 0$ for all $\bar{A} \in \bar{E}$. This in turn is equivalent to $y_1 \in (\bar{E} y_d y_{d-1} \cdots y_2)^\perp$ where we use the notion of contraction of subspaces introduced in (5.10). Summarizing, we have

$$p^{2^d} \le \mathbb{P}\left\{y_1 \in (\bar{E} y_d y_{d-1} \cdots y_2)^\perp\right\}.$$

Step 3: Contraction. In view of (7.6), Corollary 5.8 gives

$$\dim(\bar{E} y_d y_{d-1} \cdots y_2) \ge \left(\frac{\varepsilon}{8}\right)^d n_1 \tag{7.7}$$

with probability at least[9]

$$1 - \exp\Big(-c\Big(\frac{\varepsilon}{8}\Big)^d n_0\Big),$$

where $n_0 = \min(n_2, \ldots, n_d)$.

Now condition on a realization of $y_2, \ldots, y_d$ satisfying (7.7) and apply Lemma 3.8 for the random vector $x = y_1$ and fixed subspace $F := (\bar{E} y_d y_{d-1} \cdots y_2)^\perp$. This gives

$$\mathbb{P}\left\{y_1 \in (\bar{E} y_d y_{d-1} \cdots y_2)^\perp \mid y_2, \ldots, y_d\right\} \le 2\exp\Big(-c\Big(\frac{\varepsilon}{8}\Big)^d n_1\Big).$$

[7] Note the difference between what we call the restriction $y_{I_k} \in \mathbb{R}^n$, which a vector with zero coefficients outside I_k, and the vector $y_k \in \mathbb{R}^{n_k}$ which has no coordinates at all outside I_k. It is more convenient to consider the latter vectors at this stage of the argument.

[8] As we noted above, the only difference between these two subspaces is whether the restriction is defined as dropping some entries or zeroing them.

[9] Remark 5.9 explains why we can use Corollary 5.8 for random vectors y_k although they are not uniformly distributed in the hypercube.

Then the (unconditional) probability satisfies

$$\begin{aligned} p^{2^d} &\le \mathbb{P}\left\{y_1 \in (\bar{E} y_d y_{d-1} \cdots y_2)^\perp\right\} \\ &\le 2\exp\Big(-c\Big(\frac{\varepsilon}{8}\Big)^d n_0\Big) + 2\exp\Big(-c\Big(\frac{\varepsilon}{8}\Big)^d n_1\Big). \end{aligned}$$

Recall that $n_0 = \min(n_2, \ldots, n_d) \ge \varepsilon n/4$ and $n_1 \ge \varepsilon n/4$ by (7.4). Thus

$$p^{2^d} \le 4\exp\Big(-c\Big(\frac{\varepsilon}{8}\Big)^d \varepsilon n\Big)$$

Simplifying the bound completes the proof. □

7.2. **Random tensors are linearly independent.** We are ready to state and prove a precise, quantitative version of the result we announced in Theorem 2.2.

Theorem 7.3 (Random tensors are linearly independent)**.** *Let $x_1, \ldots, x_m$ be independent random vectors uniformly distributed in $\{-1,1\}^n$. Let $d \ge 1$ and assume that*

$$m \le N(n,d) - \varepsilon n^d$$

for some $\varepsilon \in (0,1)$. Then the random tensors $x_1^{\otimes d}, \ldots, x_m^{\otimes d}$ are linearly independent with probability at least $1 - 4n^d \exp\big(-(c\varepsilon)^{d+1} n\big)$.

Proof. The random tensors are linearly independent if and only if each tensor $x_k^{\otimes d}$ does not lie in the linear span of the others, which we denote

$$L_k = \operatorname{span}(x_j^{\otimes d})_{j \ne k}.$$

The dimension of L_k is trivially less than the total number of vectors m. Thus, by assumption, the codimension on L_k is greater than εn^d. Applying Proposition 7.2, we get

$$\mathbb{P}\left\{x_k^{\otimes d} \in L_k\right\} \le 4\exp\Big(-(c\varepsilon)^{d+1} n\Big) =: p.$$

for each k. Taking the union bound over all $k = 1, \ldots, m$, we see that

$$\mathbb{P}\{\text{linear dependence}\} = \mathbb{P}\left\{\exists k :\ x_k^{\otimes d} \in L_k\right\} \le mp. \tag{7.8}$$

Due to our assumption on m and the elementary bound (3.4) on $N(n,d)$, we have $m \le N(n,d) \le n^d$. Substituting this into (7.8) completes the proof. □

8. Random tensors span unique subspaces

In this section we prove a key result on the uniqueness of spans of random tensors. We announced a simplified version of this result in Theorem 8, and here we will prove a sharper, quantitative version. For random vectors ($d = 1$) this result was first proved by A. Oldyzko [62]; for matrices ($d = 2$) as well as higher-order tensors ($d > 2$) the following theorem is new.

Theorem 8.1 (Random tensors span unique subspaces)**.** *Let $x_1, \ldots, x_m$ be independent random vectors uniformly distributed in $\{-1,1\}^n$. Let $1 \le d \le c\sqrt{\log n / \log\log n}$ and assume that*

$$m \le N(n,d) - \frac{Cn^d}{\log n}.$$

Then, with probability at least $1 - 2\exp(-n^{1/3})$, the span of the random tensors $x_1^{\otimes d}, \ldots, x_m^{\otimes d}$ does not contain any simple tensor $u^{\otimes d}$ that is different from $\pm x_k^{\otimes d}$ and such that $u \in \{-1,1\}^n$.

To prove Theorem 8.1, we have to bound 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 of which are non-zero, and such that

$$\sum_{k=1}^m a_k x_k^{\otimes d} = u^{\otimes d}. \tag{8.1}$$

Our argument will be a little different depending on how many coefficients a_k are nonzero. In the next subsection, we bound the probability assuming that there are at least $n^{1/4}$ nonzero coefficients (“long combinations”), and in Section 8.2 we analyze the remaining case of “short combinations”. We combine these two cases in Section 8.3 where we finish the proof of Theorem 8.1.

8.1. **Long combinations.**

Lemma 8.2 (Long combinations)**.** *Let $x_1, \ldots, x_m$ be independent random vectors uniformly distributed in $\{-1,1\}^n$. Let $1 \le d \le n$ and assume that*

$$m \le N(n,d) - \varepsilon n^d$$

for some $\varepsilon \ge C/\log n$. 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 $n^{1/4}$ of which are nonzero, and such that (8.1) holds. Then

$$P_{\text{long}} \le 8n^d \exp\big(-(c\varepsilon)^{d+1}\, n\big).$$

Proof. **Step 1: Extracting two batches of equations.** We can view (8.1) as a system of n^d linear equations in variables $a_1, \ldots, a_m$, and we can write it as

$$\sum_{k=1}^m a_k x_{ki_1} \cdots x_{ki_d} = u_{i_1} \cdots u_{i_d}, \quad 1 \le i_1, \ldots, i_d \le n.$$

We will consider two subsets (“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. Let

$$n_0 := \Big(1 - \frac{\varepsilon}{4}\Big) n \tag{8.2}$$

and define the first batch to be

$$\sum_{k=1}^{m} a_k x_{ki_1} \cdots x_{ki_d} = u_{i_1} \cdots u_{i_d}, \quad 1 \le i_1, \ldots, i_d \le n_0 \tag{8.3}$$

and the second batch to be

$$\sum_{k=1}^{m} a_k (x_{kn})^{d-1} x_{ki} = (u_n)^{d-1} u_i, \quad n_0 < i < n. \tag{8.4}$$

The second batch can be obtained by choosing $i_1 = \cdots = i_{d-1} = n$ and letting i_d vary in the interval (n_0, n). The crucial property is that these two sets of equations are stochastically independent, since the two sets of random variables x_{ki} defining them are disjoint.

To simplify the calculations, let us assume at this time that d is odd; if d is even, the argument below still holds with obvious alterations. Under this assumption, $(x_{kn})^{d-1}$ and $(u_n)^{d-1}$ in the equations (8.4) both equal 1, and these equations become

$$\sum_{k=1}^{m} a_k x_{ki} = u_i, \quad n_0 < i < n. \tag{8.5}$$

Step 2: The first batch has full rank. Rewrite the first batch of equations (8.3) as

$$\sum_{k=1}^{m} a_k \bar{x}_k^{\otimes d} = \bar{u}^{\otimes d} \tag{8.6}$$

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 $\bar{x}_k \in \{-1,1\}^{n_0}$.

In preparing to apply Theorem 7.3, we claim that

$$N(n,d) - \varepsilon n^d \le N(n_0,d) - \frac{\varepsilon n_0^d}{2}. \tag{8.7}$$

To check this inequality, first note that definition (8.2) of n_0 yields

$$n_0 \ge \Big(1 - \frac{\varepsilon d}{4}\Big) n^d$$

by Bernoulli's inequality. This bounds together with Lemma 3.2 gives

$$N(n,d) - N(n_0,d) \le \frac{n^d}{d!}\Big(\frac{2d^2}{n} + \frac{\varepsilon d}{4}\Big). \tag{8.8}$$

On the other hand, since $n_0 \le n$, we have

$$\varepsilon n^d - \frac{\varepsilon n_0^d}{2} \ge \frac{\varepsilon n^d}{2}. \tag{8.9}$$

If $\varepsilon > C/\log n$ with sufficiently large absolute constant C, it is easy to check that the right hand side of (8.8) is bounded by the right hand side of (8.9). In other words, (8.7) holds.

Our condition on the number of vectors $m \le N(n,d) - \varepsilon n^d$ and (8.7) imply that $m \le N(n_0,d) - \varepsilon n_0^d/2$. Thus we can apply Theorem 7.3 in dimension

n_0 instead of n, and with $\varepsilon/2$ instead of ε. It states that the random tensors $\bar{x}_1^{\otimes d}, \ldots, \bar{x}_m^{\otimes d}$ are linearly independent with probability at least $1-p_0$, where

$$p_0 := 4n_0^d \exp\big(-(c'\varepsilon)^{d+1} n_0\big) \leq 4n^d \exp\big(-(c\varepsilon)^{d+1} n\big). \tag{8.10}$$

In other words, the first batch of equations has full rank with high probability $1-p_0$.

Step 3: The second batch bounds the probability. Assume that the event that defines P_{long} holds. We can rule out the case where the random tensors $\bar{x}_1^{\otimes d}, \ldots, \bar{x}_m^{\otimes d}$ are linearly dependent, which happens with small probability (at most p_0). So let us condition on a realization of the random vectors $\bar{x}_1, \ldots, \bar{x}_m$ for which linear independence holds.[10] The vector $u \in \{-1,1\}^n$ in the event that defines P_{long} can be chosen in 2^n ways; let us fix it. Since the tensors $\bar{x}_k^{\otimes d}$ and $\bar{u}^{\otimes d}$ are fixed at this point, linear independence implies that the coefficients (a_k) are uniquely determined by the first batch of equations (8.6). Thus the coefficients (a_k) are now fixed, too. Put them in the second batch of equations (8.5), which is stochastically independent from the first. This reasoning gives

$$P_{\text{long}} \leq p_0 + 2^n \cdot \max_{a,u} \mathbb{P}\left\{\sum_{k=1}^m a_k x_{ki} = u_i,\ n_0 < i < n\right\} \tag{8.11}$$

where the maximum is over all vectors $a = (a_k)$ with at least $n^{1/4}$ nonzero coefficients, and over all vectors $u = (u_i)$ with ± 1 coefficients.

Step 4: Applying Littlewood-Offord Lemma. The Littlewood-Offord Lemma (Lemma 3.10 and Lemma 3.11) can help us bound the probability of each equation in (8.11). Since at least $n^{1/4}$ coefficients a_k are nonzero, we get

$$\mathbb{P}\left\{\sum_{k=1}^m a_k x_{ki} = u_i\right\} \leq P(n^{1/4}) \leq \frac{C'}{n^{1/8}}$$

for each i. Since all $n - n_0 - 1$ such equations in (8.11) are stochastically independent, this implies

$$P_{\text{long}} \leq p_0 + 2^n \cdot \left(\frac{C'}{n^{1/8}}\right)^{n-n_0-1}.$$

Definition (8.2) of n_0 and the assumption $\varepsilon > C/\log n$ give a good bound on the exponent, namely $n - n_0 - 1 \geq \varepsilon n/5$. Thus,

$$\left(\frac{C'}{n^{1/8}}\right)^{n-n_0-1} \leq \exp(-c'\varepsilon n \log n) \leq 4^{-n},$$

where the last bound follows from the assumption $\varepsilon \geq C/\log n$ with sufficiently large absolute constant C. This gives

$$P_{\text{long}} \leq p_0 + 2^{-n} \leq 2p_0$$

[10] If d is even, condition also on the last coordinates of all vectors, i.e. on x_{kn} for all $k = 1, \ldots, m$. This fixes the weights $(x_{kn})^{d-1}$ in (8.4).

if the constant $c > 0$ in the definition (8.10) of p_0 is sufficiently small. The proof is complete. □

8.2. **Short combinations.**

Lemma 8.3 (Short combinations)**.** *Let $x_1, \ldots, x_m$ be independent random vectors uniformly distributed in $\{-1,1\}^n$. Let $1 \le d \le n$ and m and assume that*

$$m \le N(n,d).$$

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 $n^{1/4}$ of which are nonzero, and such that (8.1) holds. Then

$$P_{\text{short}} \le 8(en)^{2dn^{1/4}} \exp\big(-(c/d!)^{d+1}\sqrt{n}\big).$$

Proof. The overall plan of our argument will be similar to the proof of Lemma 8.2 for long combinations. We need an extra step in the beginning though.

Step 0: Fixing the pattern of nonzeros. 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 $m_0 \in [2, n^{1/4}]$ 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 coefficients $a_1, \ldots, a_{m_0} \in \mathbb{R}$ exactly m_0 of which are nonzero, and which satisfy

$$\sum_{k=1}^{m_0} a_k x_k^{\otimes d} = u^{\otimes d}. \tag{8.12}$$

Step 1: Extracting two batches of equations. Like in the proof of Lemma 8.2, we consider two stochastically independent batches of equations, which we extract from the system (8.12). Set

$$n_0 := 2\sqrt{n} \tag{8.13}$$

and define the first batch to be

$$\sum_{k=1}^{m_0} a_k x_{ki_1} \cdots x_{ki_d} = u_{i_1} \cdots u_{i_d}, \quad 1 \le i_1, \ldots, i_d \le n_0, \tag{8.14}$$

and the second batch to be

$$\sum_{k=1}^{m_0} a_k (x_{kn})^{d-1} x_{ki} = (u_n)^{d-1} u_i, \quad n_0 < i < n. \tag{8.15}$$

Like above, in order to simplify the calculations, we assume that d is odd, in which case the second batch becomes

$$\sum_{k=1}^{m_0} a_k x_{ki} = u_i, \quad n_0 < i < n. \tag{8.16}$$

Step 2: The first batch has full rank. Rewrite the first batch of equations (8.14) as

$$\sum_{k=1}^{m_0} a_k \bar{x}_k^{\otimes d} = \bar{u}^{\otimes d} \tag{8.17}$$

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

In preparing to apply Theorem 7.3, we claim that

$$\sqrt{n} \le N(n_0,d) - \frac{n_0^d}{2d!}. \tag{8.18}$$

To check this inequality, note that the coarse bound (3.4) on $N(n,d)$, the definition (8.13) of n_0 and Lemma 3.2 yield

$$N(n_0,d) \ge n_0 = 2\sqrt{n} \quad \text{and} \quad N(n_0,d) \ge \frac{n_0}{d!}.$$

A combination of these two bounds yields (8.18).

Our condition that $m_0 \le n^{1/4} \le \sqrt{n}$ and (8.18) imply that $m \le N(n_0,d) - n_0^d/2d!$. Thus we can apply Theorem 7.3 in dimension $n_0 = 2\sqrt{n}$ instead of n, and with $\varepsilon := 1/2d!$. It states that the random tensors $\bar{x}_1^{\otimes d}, \dots, \bar{x}_m^{\otimes d}$ are linearly independent with probability at least $1 - p_0$, where

$$p_0 = 4n^{d/2} \exp\big(-(c/d!)^{d+1}\sqrt{n}\big). \tag{8.19}$$

In other words, the first batch of equations has full rank with high probability $1 - p_0$.

Step 3: The second batch bounds the probability. The same reasoning as we had for long combinations (in Step 3 of the proof of Lemma 8.2) gives

$$P_{m_0} \le p_0 + 2^n \cdot \max_{a,u} \mathbb{P}\left\{ \sum_{k=1}^{m_0} a_k x_{ki} = u_i,\ n_0 < i < n \right\} \tag{8.20}$$

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

The Littlewood-Offord Lemma (Lemma 3.12 and Lemma 3.11) can help us bound the probability of each equation in (8.11). Since all coefficients a_k are nonzero, $m_0 \ge 2$ and $u_i = \pm 1 \ne 0$, we have

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

for each i. Since all $n - n_0 - 1$ such equations in (8.20) are stochastically independent, this implies

$$P_{m_0} \le p_0 + 2^n \Big(\frac{3}{8}\Big)^{n-n_0-1}.$$

Recall that we set $n_0 = 2\sqrt{n}$ in (8.13). Then

$$2^n \Big(\frac{3}{8} \Big)^{n-n_0-1} \le 2e^{-c'n}.$$

(For sufficiently large n this holds since $3/8 < 1/2$, and for smaller n the bound is trivial if we choose an absolute constant $c' > 0$ small enough.) This gives

$$P_{m_0} \le p_0 + 2e^{-c'n} \le 2p_0$$

if the constant $c > 0$ in the definition (8.19) of p_0 is sufficiently small.

Step 4: Unfixing the pattern of nonzeros. 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 to choose the support of a, and over all $m_0 \in [2, n^{1/4}]$ allowed sizes of the support, we obtain

$$P_{\text{short}} \le \sum_{m_0=2}^{n^{1/4}} \binom{m}{m_0} P_{m_0} \le (em)^{n^{1/4}} \cdot 2p_0.$$

In the last step we used the bound (3.1) on the binomial sum. To complete the proof, note that our assumption on m and a course bound (3.4) on $N(n,d)$ imply $m \le N(n,d) \le n^d$, and simplify the resulting bound. □

8.3. **Proof of Theorem 8.1.** Apply Lemma 8.2 with $\varepsilon := C/\log n$ for long combinations and Lemma 8.3 for short combinations, and take the union bound. The probability that the conclusion of Theorem 8.1 fails is bounded by $P_{\text{long}} + P_{\text{short}}$. To complete the proof, simplify the bound using the assumption on d. (The assumption $d \le c\sqrt{\log n / \log\log n}$ ensures that $(d!)^{d+1} \le n^{1/6}$ and, as a result, the bound on P_{short} in Lemma 8.3 is nontrivial.) □

9. Lots of unique subspaces and threshold functions

We are about to deduce the main result or this paper – a lower bound on the number of polynomial threshold functions. An asymptotic form of this result was announced in Theorem 1.1. In this section we prove a more precise, quantitative form of the lower bound in Theorem 9.3. (The upper bound was already proved in Theorem 3.6.)

In preparation for the proof, let us note how Theorem 8.1 implies that random tensors $x_k^{\otimes d}$ span a lot of different subspaces in $\mathrm{Sym}^d(\mathbb{R}^n)$. Precisely, we will count different subspaces of the form

$$E(S) := \mathrm{span}\{x^{\otimes d} : \; x \in S\},$$

where S is a subset of the Boolean hypercube $\{-1,1\}^n$.

Lemma 9.1 (Lots of unique subspaces)**.** *Let d and m be positive integers such that $1 \le d \le c\sqrt{\log n/\log\log n}$ and*

$$m \le N(n,d) - \frac{Cn^d}{\log n}.$$

Then there exist at least $\frac{1}{2}\binom{2^n}{m}$ different subspaces $E(S)$ 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 the E_S does not contain any simple tensor $u^{\otimes d}$ that is different from all $\pm x^{\otimes d}$ where $x \in S$. Call S *bad* otherwise. Obviously, if S and T are two different good subsets, then E_S and E_T are two different subspaces. Thus, to complete the proof, it suffices to show that at least half of the m-element subsets of the Boolean hypercube $\{-1,1\}^n$ are good.

Theorem 8.1 implies that if S is a random set obtained by sampling m points from the Boolean hypercube $\{-1,1\}^n$ *with replacement*, then

$$\mathbb{P}\{S \text{ is bad}\} \le \frac{1}{4}. \tag{9.1}$$

Alternatively, we can sample a random m-element subset $S \subset \{-1,1\}^n$ *without replacement*. Denoting the probability corresponding to this model by $\mathbb{P}_0$, we have

$$\mathbb{P}_0\{S \text{ is bad}\} = \mathbb{P}\{S \text{ is bad} \mid \text{no repeat}\} \le \frac{\mathbb{P}\{S \text{ is bad}\}}{\mathbb{P}\{\text{no repeat}\}}, \tag{9.2}$$

where "no repeat" stands for the event where no element is sampled more than once when sampling with replacement. We have

$$\begin{aligned}\mathbb{P}\{\text{no repeat}\} &= \Big(1 - \frac{1}{2^n}\Big)\Big(1 - \frac{2}{2^n}\Big)\cdots\Big(1 - \frac{m-1}{2^n}\Big)\\ &\ge \Big(1 - \frac{m-1}{2^n}\Big)^{m-1} \ge 1 - \frac{(m-1)^2}{2^n}. \end{aligned}\tag{9.3}$$

To simplify this further, note that our assumption on m and the coarse bound (3.4) of $N(n,d)$ imply that $m \le N(n,d) \le n^d$, and simplify the bound. This and the assumption on d then yield $(m-1)^2 \le m^2 \le 2^{n-1}$. Substituting this into (9.3), we conclude that

$$\mathbb{P}\{\text{no repeat}\} \ge \frac{1}{2}.$$

This, (9.1) and (9.2) give

$$\mathbb{P}_0\{S \text{ is bad}\} \le \frac{1/4}{1/2} = \frac{1}{2}.$$

Since S is sampled without replacement here, this bound says that at most half of the m-elements subsets of $\{-1,1\}^n$ are bad. The proof is complete. □

Theorem 9.2 (Lots of unique subspaces)**.** *Let $E(n,d)$ denote the number of different linear subspaces E_S where $S \subset \{-1,1\}^n$ are subsets of the Boolean hypercube. If $1 \le d \le c\sqrt{\log n/\log\log n}$, then*

$$\log_2 E(n,d) \ge \frac{n^{d+1}}{d!} - \frac{Cn^{d+1}}{\log n}. \tag{9.4}$$

Proof. Applying Lemma 9.1 with

$$m := N(n,d) - Cn^d/\log n,$$

we get

$$E(n,d) \ge \frac{1}{2}\binom{2^n}{m} \ge \frac{1}{2}\Big(\frac{2^n}{m}\Big)^m,$$

where the last inequality is obtained using the elementary bound (3.1) on the binomial coefficient. Thus

$$\log_2 E(n,d) \ge m(n - \log_2 m) - 1.$$

Next, the definition of m and the bounds on $N(n,d)$ from Lemma 3.2 and (3.4) allow us to estimate m as follows:

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

Therefore

$$\begin{aligned}\log_2 E(n,d) &\ge \Big(\frac{n^d}{d!} - \frac{Cn^d}{\log n}\Big)(n - d\log_2 n) - 1\\ &\ge \frac{n^{d+1}}{d!} - \frac{Cn^{d+1}}{\log n} - \frac{n^d d\log_2 n}{d!} - 1. \qquad (9.5)\end{aligned}$$

The third and fourth terms in the right side of (9.5), namely $n^d d\log_2 n/d!$ and 1, are trivially bounded by the second term if the absolute constant C is sufficiently large. This gives

$$\log_2 E(n,d) \ge \frac{n^{d+1}}{d!} - \frac{3Cn^{d+1}}{\log n},$$

as claimed (up to renaming the absolutely constant C). □

Theorem 9.3 (Lower bound)**.** *Let $2 \le d \le c\sqrt{\log n/\log\log n}$. Then the number of homogeneous polynomial threshold functions $\bar{T}(n,d)$ satisfies*

$$\log_2 \bar{T}(n,d) \ge \frac{n^{d+1}}{d!} - \frac{Cn^{d+1}}{\log n}.$$

Proof. Recall from the proof of Proposition 3.5 that $\bar{T}(n,d)$ is the same as the number of homogeneous linear threshold functions on the tensor lift $\mathcal{X} \subset \mathrm{Sym}^d(\mathbb{R}^n)$ of the Boolean hypercube, which is defined as

$$\mathcal{X} := \Big\{x^{\otimes d} :\ x \in \{-1,1\}^n\Big\}.$$

Lemmas 3.3 and 3.4 yield that $\bar{T}(n,d)$ is bounded below by the number of all intersection subspaces, which are the linear subspaces generated by intersecting various hyperplanes $z^\perp$, where $z \in \mathcal{X}$. 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 $\mathcal{X}$, which is the number we denoted $E(n,d)$ in Theorem 9.2. An application of that theorem completes the proof. □

10. Further questions

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

10.1. **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. Recall that the upper bound on $T(n,d)$ that we stated in Theorem 3.6 does hold for all degrees d, whether bounded or not. The proof of this result can be sharpened just a little to produce the following tighter bound which is implicit in [8], see [5, Theorem 4.7]:

$$T(n,d) \le 2B\big(2^n - 1, B(n,d) - 1\big) \tag{10.1}$$

where $B(n,d)$ is the binomial sum

$$B(n,d) = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d}.$$

To check (10.1), follow the argument in Section 3.5 but replace Lemma 3.4 by Remark 3.4.1 for central arrangements, and replace the dimension $N(n,d)$ by the smaller dimension $B(n,d)$ of the space of all n-variable polynomials of degree at most d on the Boolean cube $\{-1,1\}^n$.

It is plausible that the bound (10.1) may be tight. By the argument in Section 3.5, this would be true if most Boolean tensors $x^{\otimes d}$ are in general position not only for bounded degrees d but for higher degrees as well. Thus, the following conjecture mentioned by M. Anthony [4] could hold:

Conjecture 10.1. *The number $T(n,d)$ of n-variable polynomial threshold functions of degree d satisfy*

$$T(n,d) \approx 2B\big(2^n - 1, B(n,d) - 1\big) \tag{10.2}$$

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

The approximation in (10.2) hides a smaller order term. This conjecture might be too strong and it possibly holds only after we take logarithms on both sides of (10.2).

For bounded or mildly growing degrees d, Conjecture 10.1 easily implies the main result of this paper, Theorem 1.1, namely that $\log_2 T(n,d) \approx n^{d+1}/d!$.

For $d = n/2$, Conjecture 10.1 and a careful asymptotic analysis of the bound (10.1) 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 [4]).

Finally, for $d = n$, Conjecture 10.1 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.

10.2. **Random tensors.** The core of this paper is a set of new results about random tensors. It would be interesting to extend and sharpen some of them.

10.2.1. *High degrees.* We demonstrated that stochastically independent random tensors are linearly independent with high probability. Theorem 2.2 proves this for simple random tensors $x_k^{\otimes d}$ whose degrees d are bounded or mildly growing (see also Theorem 7.3). We may ask if the same phenomenon also holds for higher degrees. As we explained above, such result may be a key to proving Conjecture 10.1.

Conjecture 10.2. *Let $x_1, \ldots, x_m$ be independent random vectors uniformly distributed in $\{-1, 1\}^n$. Then, for any degree $1 \le d \le n$, the set of*

$$m = N(n, d)(1 - o(1))$$

random tensors $x_1^{\otimes d}, \ldots, x_m^{\otimes d}$ is linearly independent with high probability.

A similar question can be asked about Theorem 2.3, namely whether this result on the uniqueness of spans of random tensors holds for higher degrees d.

10.2.2. *All the way up to the dimension threshold.* The number of linearly independent tensors is always bounded by the dimension of the ambient space. So Theorem 2.2 holds with *almost* the maximal possible number m of random tensors – "almost" is due to the $o(1)$ term. Can one remove any slack here, and take m in Theorem 2.2 to be equal *exactly* to the dimension of the span of simple tensors $x^{\otimes d}$, $x \in \{-1, 1\}^n$?

The answer to this question is positive for $d = 1$, corresponding to the statement that n random vectors with i.i.d. ±1 coordinates are linearly independent with high probability. This is equivalent to the fact that an $n \times n$ random matrix with i.i.d. ±1 entries is invertible with high probability. Originally proved by G. Halasz [33], this result was a starting point in many recent developments in random matrix theory; some of the relevant references are mentioned in Section 1.1. But already for $d = 2$ the same question is open and it seems to be challenging. It can be stated as follows: are $m = \binom{n}{2} + 1$ random matrices $x_1 x_1^\mathsf{T}, \ldots, x_m x_m^\mathsf{T}$ linearly independent with high probability? Here as usual x_k are i.i.d. ±1 random vectors.

10.3. **Polynomial threshold functions with restricted coefficients.**

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

10.3.1. *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 [34] and N. Alon and V. Vu [3]. For any higher degree $d \geq 2$, determining the optimal bound on the coefficients of $p(x)$ seems to be an open problem.

Furthermore, it may be natural to look for a bound on the coefficients of $p(x)$ that holds for *most* (or many) polynomial threshold functions $f(x) = \mathrm{sign}(p(x))$, 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 where we require the integer coefficients of $p(x)$ to be bounded by a given number M. How many polynomial threshold functions $f(x) = \mathrm{sign}(p(x))$ can be generated with this restriction? What if we consider polynomials $p(x)$ with all ± 1 coefficients? Since each such polynomial consists of $B(n,d) = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{d}$ monomial terms and each term is assigned an ± 1 coefficient, there are at most $2^{B(n,d)}$ polynomials with ± 1 coefficients. Thus the number of corresponding polynomial threshold functions is bounded by $2^{B(n,d)}$. 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.

10.3.2. *Positive coefficients.* In some other situations (e.g. excitatory neurons in neural networks), it is natural to consider polynomial threshold functions $f(x) = \mathrm{sign}(p(x))$ 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 $f(x) = \mathrm{sign}(a_1 x_1 + \cdots a_n x_n)$ 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.

10.4. **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?

10.5. **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. Preliminary results in this direction are described in [10] 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.

10.6. **The geometry of boolean threshold functions.** As we noted in Section 3.4, 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 [19]; see also [68] for random tessellations on the sphere. However, the main new challenge here is to handle the discrete distribution induced by the Boolean cube.

References

[1] 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.

[2] 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).

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

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

[5] 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.

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

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

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

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

[10] P. Baldi, R. Vershynin, *On the capacity of shallow and deep learning architectures,* submitted, 2018.

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

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

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

[14] R. Beigel, N. Reingold, D. Spielman, *PP is closed under intersection,* Journal of Computer & System Sciences 50 (1995), 191–202.

[15] B. Bollobás, *Combinatorics. Set systems, hypergraphs, families of vectors and combinatorial probability.* Cambridge University Press, Cambridge, 1986.

[16] J. Bourgain, V. Vu, P. Wood, *On the singularity probability of discrete random matrices,* J. Funct. Anal. 258 (2010), 559–603.

[17] J. Bruck, *Harmonic analysis of polynomial threshold functions,* SIAM J. Discrete Math. 3 (1990), 168–177.

[18] R. C. Buck, *Partition of space,* Amer. Math. Monthly 50 (1943), 541–544.

[19] P. Calka, *Tessellations.* New perspectives in stochastic geometry, 145–169, Oxford Univ. Press, Oxford, 2010.

[20] N. Cook, *On the singularity of adjacency matrices for random regular digraphs,* Probab. Theory Related Fields 167 (2017), 143–200.

[21] K. Costello, *Bilinear and quadratic variants on the Littlewood-Offord problem,* Israel J. Math. 194 (2013), 359–394.

[22] K. Costello, T. Tao, V. Vu, *Random symmetric matrices are almost surely nonsingular,* Duke Math. J. 135 (2006), 395–413.

[23] P. Comon, G. Golub, L.-H. Lim, B. Mourrain, *Symmetric tensors and symmetric tensor rank,* SIAM J. Matrix Anal. Appl. 30 (2008), 1254–1279.

[24] 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.

[25] V. de la Peña, E. Giné, *Decoupling: from dependence to independence.* Springer Verlag, 1999.

[26] 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.

[27] 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.

[28] P. Erdös, *On a lemma of Littlewood and Offord,* Bull. Amer. Math. Soc. 51 (1945), 898–902.

[29] P. Erdös, *Extremal problems in number theory,* Proc. Sympos. Pure Math., vol.VIII, AMS, Providence, RI, 1965, pp.181–189.

[30] P. Frankl,Z. Füredi, *Solution of the Littlewood-Offord problem in high dimensions,* Ann. of Math. (2) 128 (1988), 259–270.

[31] F. Götze, *Asymptotic expansions for bivariate von Mises functionals,* Z. Wahrsch. Verw. Gebiete 50 (1979), 333–355.

[32] 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.

[33] G. Halász, *Estimates for the concentration function of combinatorial number theory and probability,* Period. Math. Hungar. 8 (1977) 197–211.

[34] J. Håstad, *On the size of weights for threshold gates,* SIAM J. Discrete Math. 7 (1994), 484–492.

[35] A. A. Irmatov, *On the number of threshold functions,* Diskret. Mat. 5 (1993), 40–43; translation in Discrete Math. Appl. 3 (1993), 429–432.

[36] A. A. Irmatov, *Arrangements of hyperplanes and the number of threshold functions,* Acta Appl. Math. 68 (2001), 211–226.

[37] N. Kalton, *Rademacher series and decoupling,* New York J. Math. 11 (2005), 563–595.

[38] 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.

[39] Z. Kovijanić Vukićević, *An enumerative problem in threshold logic,* Publ. Inst. Math. (Beograd) (N.S.) 82(96) (2007), 129–134.

[40] J. Kahn, J. Komlós, E. Szemerédi, *On the probability that a random* ± 1*-matrix is singular,* J. Amer. Math. Soc. 8 (1995), 223–240.

[41] R. Kannan, *Decoupling and partial independence,* Building bridges, 321–331, Bolyai Soc. Math. Stud., 19, Springer, Berlin, 2008.

[42] 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.

[43] A. Klivans, R. Servedio, *Learning DNF in time* $2^{O(n^{1/3})}$, J. Computer and System Sciences 68 (2004), 303–318.

[44] T. Kolda, B. Bader, *Tensor decompositions and applications,* SIAM Rev. 51 (2009), no. 3, 455-?500.

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

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

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

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

[49] 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.

[50] 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[66] 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.

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

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

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

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

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

[72] 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.

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

[74] 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.

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

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

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

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

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

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

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

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

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

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

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

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

[87] 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.

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

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

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

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

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

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

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

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

[96] 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.

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

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

[99] 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.

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

[101] 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.

[102] 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.

[103] 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