# Low-degree learning and the metric entropy of polynomials

Alexandros Eskenazis, Paata Ivanisvili, Lauritz Streck

# LOW-DEGREE LEARNING AND THE METRIC ENTROPY OF POLYNOMIALS

ALEXANDROS ESKENAZIS, PAATA IVANISVILI, AND LAURITZ STRECK

Abstract. Let $\mathcal{F}_{n,d}$ be the class of all functions $f : \{-1,1\}^n \to [-1,1]$ on the $n$-dimensional discrete hypercube of degree at most $d$. In the first part of this paper, we prove that any (deterministic or randomized) algorithm which learns $\mathcal{F}_{n,d}$ with $L_2$-accuracy $\varepsilon$ requires at least $\Omega((1-\sqrt{\varepsilon})2^d \log n)$ queries for large enough $n$, thus establishing the sharpness as $n \to \infty$ of a recent upper bound of Eskenazis and Ivanisvili (2021). To do this, we show that the $L_2$-packing numbers $\mathsf{M}(\mathcal{F}_{n,d}, \|\cdot\|_{L_2}, \varepsilon)$ of the concept class $\mathcal{F}_{n,d}$ satisfy the two-sided estimate

$$c(1-\varepsilon)2^d \log n \le \log \mathsf{M}(\mathcal{F}_{n,d}, \|\cdot\|_{L_2}, \varepsilon) \le \frac{2^{Cd}\log n}{\varepsilon^4}$$

for large enough $n$, where $c, C > 0$ are universal constants. In the second part of the paper, we present a logarithmic upper bound for the randomized query complexity of classes of bounded *approximate* polynomials whose Fourier spectra are concentrated on few subsets. As an application, we prove new estimates for the number of random queries required to learn approximate juntas of a given degree, functions with rapidly decaying Fourier tails and constant depth circuits of given size. Finally, we obtain bounds for the number of queries required to learn the polynomial class $\mathcal{F}_{n,d}$ without error in the query and random example models.

*2020 Mathematics Subject Classification.* Primary: 06E30; Secondary: 42C10, 68Q32.
*Key words.* Discrete hypercube, learning theory, query complexity, metric entropy, junta, constant depth circuit.

## 1. Introduction

For any function $f : \{-1,1\}^n \to \mathbb{R}$, there exist unique real coefficients $\{\hat{f}(S)\}_{S\subseteq\{1,\dots,n\}}$ such that

$$\forall\, x \in \{-1,1\}^n, \qquad f(x) = \sum_{S\subseteq\{1,\dots,n\}} \hat{f}(S) w_S(x), \tag{1}$$

where the Walsh functions $w_S : \{-1,1\}^n \to \{-1,1\}$ are defined by $w_S(x) = \prod_{i\in S} x_i$. We say that the function $f$ has degree at most $d \in \{0,1,\dots,n\}$ if $\hat{f}(S) = 0$ for all sets $S$ with $|S| > d$.

1.1. **Learning functions on the hypercube.** Let $\mathcal{F}$ be a class of functions on the discrete hypercube of dimension $n$. The learning problem for the class $\mathcal{F}$ can be described as follows. Consider an unknown function $f \in \mathcal{F}$. Given access to *examples* $(X_1, f(X_1)), \dots, (X_Q, f(X_Q))$, the goal is to algorithmically construct a *hypothesis function* $h : \{-1,1\}^n \to \mathbb{R}$ which effectively approximates $f$. Different access models to examples give rise to concrete versions of the learning problem. The two most standard such models are the *query* model, in which the algorithm can sequentially request the values of $f$ at any $Q$-tuple of points $X_1, \dots, X_Q$ from $\{-1,1\}^n$, and the *random example* model, in which the data points $X_1, \dots, X_Q$ are generated uniformly and independently from $\{-1,1\}^n$. In the query model, the goal is to construct a function $h$ satisfying $\|h-f\|_{L_2}^2 \le \varepsilon$ whereas in the random example model, the desired output is a *random* function $h$ satisfying $\|h-f\|_{L_2}^2 \le \varepsilon$ with probability at least $1-\delta$, where $\varepsilon, \delta \in [0,1)$ are pre-fixed accuracy and confidence parameters respectively. The least number $Q$ of examples required to solve the learning problem in each case is called the *query complexity* of the model and shall be denoted by $\mathsf{Q}(\mathcal{F}, \varepsilon)$ for the query model and by $\mathsf{Q}_r(\mathcal{F}, \varepsilon, \delta)$ for the random example model.

A. E. was partially supported by a Junior Research Fellowship from Trinity College, Cambridge. P. I. was partially supported by the NSF grants DMS-2152346 and CAREER-DMS-2152401. L. S. has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 803711).

The query complexity of learning problems has been studied extensively for various classes $\mathcal{F}$ of functions on the discrete hypercube (see [32, 36]). One of the first rigorous results of this kind is the *Low-Degree Algorithm* of Linial, Mansour and Nisan [30], who considered the class

$$\mathcal{F}_{n,d} \stackrel{\text{def}}{=} \Big\{ f : \{-1,1\}^n \to [-1,1] : \ f \text{ has degree at most } d \Big\} \tag{2}$$

and showed the estimate $\mathrm{Q}_r(\mathcal{F}_{n,d}, \varepsilon, \delta) \le \frac{2n^d}{\varepsilon} \log(\frac{2n^d}{\delta})$ for any $\varepsilon, \delta \in (0,1)$. In the recent work [16], which followed an intermediate $O_{d,\varepsilon,\delta}(n^{d-1} \log n)$ asymptotic[1] improvement in [21], it was shown that this classical estimate is largely suboptimal as $n \to \infty$ and in fact

$$\mathrm{Q}_r(\mathcal{F}_{n,d}, \varepsilon, \delta) \le \min\left\{ \frac{\exp(Cd^{3/2}\sqrt{\log d})}{\varepsilon^{d+1}}, \frac{4dn^d}{\varepsilon} \right\} \log\left(\frac{n}{\delta}\right) \tag{3}$$

The first goal of the present paper is to investigate lower bounds for the query complexity, which in particular imply that (3) is asymptotically optimal as $n \to \infty$, that is

$$\mathrm{Q}_r(\mathcal{F}_{n,d}, \varepsilon, \delta) = \Theta_{d,\varepsilon,\delta}(\log n). \tag{4}$$

Consider the class

$$\mathcal{B}_{n,d} \stackrel{\text{def}}{=} \Big\{ f : \{-1,1\}^n \to \{-1,1\} : \ f \text{ has degree at most } d \Big\} \subset \mathcal{F}_{n,d}. \tag{5}$$

We will prove the following lower estimate for the complexities of this class.

**Theorem 1.** *For any $n \in \mathbb{N}$, $d \in \{1,\dots,n\}$, $\varepsilon \in [0,1)$ and $\delta \in (0,1)$, we have*

$$\mathrm{Q}(\mathcal{B}_{n,d}, \varepsilon) \ge \max\left\{ (1-\sqrt{\varepsilon}) 2^{d-2} \log_2 n - (d+1)2^{d-2}, d \log_2\left(\frac{n}{d}\right) \right\} \tag{6}$$

*and*

$$\mathrm{Q}_r(\mathcal{B}_{n,d}, \varepsilon, \delta) \ge \max\left\{ (1-\sqrt{\varepsilon}) 2^{d-2} \log_2 n - (d+1)2^{d-2}, d \log_2\left(\frac{n}{d}\right) \right\} + \log_2(1-\delta). \tag{7}$$

The equivalence (4) now follows due to the inequality $\mathrm{Q}_r(\mathcal{B}_{n,d}, \varepsilon, \delta) \le \mathrm{Q}_r(\mathcal{F}_{n,d}, \varepsilon, \delta)$. The tools used in the proof of Theorem 1 will be described in Section 1.3.

The second goal of this paper is to show a more robust version of the upper bound (3) that applies to different concept classes $\mathcal{F}$ which are not necessarily of bounded degree. In order to present this result we shall need some terminology (see also [36, Chapter 3]). If $t \ge 0$, we say that the Fourier spectrum of a function $f : \{-1,1\}^n \to \mathbb{R}$ is $t$-concentrated up to degree $d$ if

$$\sum_{|S|>d} \hat{f}(S)^2 \le t. \tag{8}$$

More generally, given a family $\mathcal{S}$ of subsets of $\{1,\dots,n\}$ we say that the spectrum of $f$ is $\eta$-concentrated on $\mathcal{S}$ if

$$\sum_{S \notin \mathcal{S}} \hat{f}(S)^2 \le \eta. \tag{9}$$

Our main upper bound for learning is the following theorem.

**Theorem 2.** *Fix $n, m \in \mathbb{N}$, $d \in \{1,\dots,n\}$ and $t, \eta \in [0,1)$. Let $\mathcal{F}$ be a class of bounded functions $f : \{-1,1\}^n \to [-1,1]$ such that the Fourier spectrum of any $f \in \mathcal{F}$ is $t$-concentrated up to degree $d$ and is $\eta$-concentrated on a family $\mathcal{S}(f)$ of subsets of $\{1,\dots,n\}$ satisfying $\#\mathcal{S}(f) \le m$. Then,*

$$\forall\ \varepsilon, \delta \in (0,1), \qquad \mathrm{Q}_r(\mathcal{F}, \eta + t + \varepsilon, \delta) \le \left\lceil \frac{18m}{\varepsilon} \log\left( \frac{2}{\delta} \sum_{r=0}^{d} \binom{n}{r} \right) \right\rceil. \tag{10}$$

In Remark 13 below we will see how this statement implies the estimate (3) of [16].

[1] We shall use the standard asymptotic notation throughout the article. For $a, b > 0$, we write $a = O(b)$ if there exists a universal constant $c > 0$ such that $a \le cb$. Moreover, we shall write $a = \Theta(b)$ for $a = O(b)$ and $b = O(a)$. We shall also write $O_\xi(\cdot)$, $\Theta_\psi(\cdot)$ to indicate that the implicit constants depend on $\xi$ or $\psi$ respectively.

1.2. **Fourier concentration and learning upper bounds.** In this section we shall present concrete applications of Theorem 2 for various concept classes $\mathcal{F}$. We start with the class of approximate juntas of a given degree. Recall that a function $f : \{-1,1\}^n \to \mathbb{R}$ is called a $(k,\eta)$-junta if there exists a subset $\sigma \subseteq \{1,\ldots,n\}$ with $|\sigma| \leq k$ and a map $g : \{-1,1\}^n \to \mathbb{R}$ depending only on the variables $(x_i)_{i\in\sigma}$ such that $\|f-g\|_{L_2}^2 \leq \eta$. Consider the class

$$\mathcal{J}_{n,k,\eta} = \Big\{ f : \{-1,1\}^n \to [-1,1] : \ f \text{ is a } (k,\eta)\text{-junta} \Big\}. \tag{11}$$

We shall prove the following estimate on the randomized query complexity of $\mathcal{F}_{n,d} \cap \mathcal{J}_{n,k,\eta}$.

**Corollary 3.** *In the setting above, for $\varepsilon,\delta \in (0,1)$ we have*

$$\mathrm{Q}_r\Big(\mathcal{F}_{n,d} \cap \mathcal{J}_{n,k,\eta}, 2\eta+\varepsilon, \delta\Big) \leq \left\lceil \frac{18}{\varepsilon} \sum_{r=0}^{\min\{d,k\}} \binom{k}{r} \log\left( \frac{2}{\delta} \sum_{r=0}^{\min\{d,k\}} \binom{n}{r} \right) \right\rceil. \tag{12}$$

*In particular, choosing $d = n$, we get*

$$\mathrm{Q}_r\Big(\mathcal{J}_{n,k,\eta}, 2\eta+\varepsilon, \delta\Big) \leq \frac{2^{k+5}}{\varepsilon} \log\left( \frac{2}{\delta} \sum_{r=0}^{k} \binom{n}{r} \right). \tag{13}$$

Corollary 3 can be concretely applied in view of the large available literature on junta theorems in Boolean analysis. To motivate a first application along these lines, observe that the upper bound (3) of [16] differs from that of [30] in its dependence on $\varepsilon$ as $\varepsilon \to 0^+$. While we do not know whether the $\varepsilon^{-d-1}$ asymptotic behavior is needed to learn $\mathcal{F}_{n,d}$, Corollary 3 combined with a structural result of Nisan and Szegedy [35] gives the following upper bound for the complexity of $\mathcal{B}_{n,d}$, alas with a somewhat worse dependence on $d$.

**Corollary 4.** *For any $n \in \mathbb{N}$, $d \in \{1,\ldots,n\}$ and $\varepsilon,\delta \in (0,1)$, we have*

$$\mathrm{Q}_r(\mathcal{B}_{n,d}, \varepsilon, \delta) \leq \frac{36 \cdot d 2^{d^2}}{\varepsilon} \log\left(\frac{n}{\delta}\right). \tag{14}$$

Combining Theorem 2 with a deep junta theorem of Dinur, Friedgut, Kindler and O'Donnell [13], we will deduce that bounded functions which are sufficiently close to polynomials of degree $d$ can be learned from $O_d(\log n)$ samples. For $t \geq 0$, consider the class $\mathcal{F}_{n,d}(t)$ consisting of functions $f : \{-1,1\}^n \to [-1,1]$ whose spectra are $t$-concentrated up to degree $d$. In other words, $\mathcal{F}_{n,d}(t)$ consists of all bounded functions which are $\sqrt{t}$-close (in $L_2$) to a polynomial of degree at most $d$. Corollary 3 has the following consequence.

**Corollary 5.** *There exists a universal constant $C > 0$ such that the following holds. For any $n \in \mathbb{N}$, $d \in \{1,\ldots,n\}$, $t \in [0,1)$ and $\eta \geq \frac{Cd^2 \log d}{\log(1/t)}$, we have*

$$\forall\ \varepsilon,\delta \in (0,1), \qquad \mathrm{Q}_r\Big(\mathcal{F}_{n,d}(t), \eta+\varepsilon, \delta\Big) \leq \frac{2^{Cd^2}}{\eta^{2d}\varepsilon} \log\left(\frac{n}{\delta}\right). \tag{15}$$

It is worth emphasizing that as $t \to 0^+$, we can also take $\eta = \eta(t) \to 0^+$ in the statement above. Corollary 5 is a robust version of the main theorem of [16]. On the one hand, the method of [16] seems unfit to provide estimates for the complexity of $\mathcal{F}_{n,d}(t)$ as it uses the Bohnenblust–Hille inequality [10], which heavily relies on the fact that the unknown function is a bounded polynomial. On the other hand (see also Remark 13), Corollary 5 gives a worse estimate on the complexity of $\mathcal{F}_{n,d} = \mathcal{F}_{n,d}(0)$ in terms of $d, \varepsilon$ than the bound (3) of [16].

To the best of our knowledge, Corollary 5 is the best known upper bound for the randomized query complexity of the class $\mathcal{F}_{n,d}(t)$ for $t > 0$ after the Low-Degree Algorithm of [30] which gives the estimate $\mathrm{Q}_r(\mathcal{F}_{n,d}(t), t+\varepsilon, \delta) \leq \frac{2n^d}{\varepsilon} \log\left(\frac{2n^d}{\delta}\right)$. It remains an interesting problem to understand whether one can sharpen the dependence on $d$ of the lower bound for $\eta$ in Corollary 5. Specifically for the case of Boolean functions, the implicit dependence of $\eta$ on $t, d$ can be exponentially improved due to an important junta theorem of Bourgain [8]. For $t > 0$, denote by $\mathcal{B}_{n,d}(t)$ the class of all Boolean functions $f : \{-1,1\}^n \to \{-1,1\}$ satisfying (8).

**Corollary 6.** *There exists a universal constant $C > 0$ such that the following holds. For any $n \in \mathbb{N}$, $d \in \{1, \ldots, n\}$, $t \in [0,1)$ and $\eta \geq t^{1+o(1)} d^{\frac{1}{2}+o(1)}$, we have*[2]

$$\forall\ \varepsilon, \delta \in (0,1), \qquad \mathsf{Q}_r\Big(\mathcal{B}_{n,d}(t), \eta + \varepsilon, \delta\Big) \leq \frac{2^{Cd^2}}{\varepsilon} \log\Big(\frac{n}{\delta}\Big). \tag{16}$$

Finally, we present a concrete application of Corollary 6 to Boolean functions which can be represented by constant depth circuits. We refer to [36, Chapter 4] for the relevant definitions. For readers which are unfamiliar with this class, we just point out that DNF formulas, i.e. functions which are representable as logical $\vee$ of terms, each of which is a logical $\wedge$ of variables $x_i$ or their negations $\neg x_i$, are circuits of depth 2. Similarly, CNF formulas, in which the roles of $\vee$ and $\wedge$ are reversed, are also circuits of depth 2. Corollary 6 combined with estimates on the Fourier concentration of constant depth circuits [19, 30, 20] has the following consequence.

**Corollary 7.** *Let $\mathcal{C}_{n,d,s}$ be the class of all Boolean functions on $\{-1,1\}^n$ computable by a depth-$d$ circuit of size $s > 1$. Then, for every $\varepsilon, \delta \in (0,1)$, we have*

$$\mathsf{Q}_r(\mathcal{C}_{n,d,s}, \varepsilon, \delta) \leq \exp\Big(O(\log(s/\varepsilon))^d\Big) \log\Big(\frac{n}{\delta}\Big). \tag{17}$$

Learning constant depth circuits (also known as $\mathsf{AC}^0$ circuits) in quasi-polynomial time is the main focus of the seminal work [30] of Linial, Mansour and Nisan which prompted them to design the Low-Degree Algorithm. Moreover, it is known (see [22]) that quasi-polynomial time is also *necessary* to learn this class, conditionally on some standard cryptographic assumptions. The contribution of Corollary 7 is the fact that the query complexity of this learning problem is (exponentially) smaller than the corresponding running time [30] for large enough $n$. It is worth emphasizing that the reason Corollary 7 follows from Corollary 6 is that $\mathsf{AC}^0$ circuits have strong enough Fourier concentration. It remains an interesting problem to understand whether Corollary 6 can be boosted to encapsulate classes of Boolean functions with weaker concentration such as linear threshold functions or functions of many hyperplanes [2, 38].

1.3. **Metric entropy and learning lower bounds.** Let $(\mathcal{X}, d_{\mathcal{X}})$ be a metric space and $\varepsilon > 0$. A subset $\mathcal{P} \subseteq \mathcal{X}$ is an $\varepsilon$-packing of $\mathcal{X}$ if for any $p \neq p' \in \mathcal{P}$, we have $d_{\mathcal{X}}(p,p') > \varepsilon$. The largest size of an $\varepsilon$-packing is called the *packing number* of $\mathcal{X}$ and is denoted by $\mathsf{M}(\mathcal{X}, d_{\mathcal{X}}, \varepsilon)$. A subset $\mathcal{C} \subseteq \mathcal{X}$ is an $\varepsilon$-cover of $\mathcal{X}$ if for any $q \in \mathcal{X}$, there exists some $p \in \mathcal{C}$ with $d_{\mathcal{X}}(p,q) \leq \varepsilon$. The smallest size of an $\varepsilon$-cover is called the *covering number* of $\mathcal{X}$ and is denoted by $\mathsf{N}(\mathcal{X}, d_{\mathcal{X}}, \varepsilon)$. The quantity $\log_2 \mathsf{N}(\mathcal{X}, d_{\mathcal{X}}, \varepsilon)$ is called the *$\varepsilon$-metric entropy* of $\mathcal{X}$. It is well known (see [42, Lemma 4.2.8]) that packing and covering numbers are closely related via the elementary inequalities

$$\forall\ \varepsilon > 0, \qquad \mathsf{M}(\mathcal{X}, d_{\mathcal{X}}, 2\varepsilon) \leq \mathsf{N}(\mathcal{X}, d_{\mathcal{X}}, \varepsilon) \leq \mathsf{M}(\mathcal{X}, d_{\mathcal{X}}, \varepsilon). \tag{18}$$

The pertinence of metric entropy in the context of learning lower bounds stems from the classical observation that concept classes with large covering (or packing) numbers cannot be efficiently learned from few queries (see, for instance, the works [4, 31, 14]). In our setting, we shall need the following concrete estimate which we could not locate in the literature.

**Proposition 8.** *Fix $n \in \mathbb{N}$ and let $\mathcal{B}$ be a class of Boolean functions on $\{-1,1\}^n$. Then,*

$$\forall\ \varepsilon > 0, \qquad \mathsf{Q}(\mathcal{B}, \varepsilon) \geq \log_2 \mathsf{M}(\mathcal{B}, \|\cdot\|_{L_2}, 2\sqrt{\varepsilon}) \tag{19}$$

*and*

$$\forall\ \varepsilon > 0, \forall\ \delta \in (0,1), \qquad \mathsf{Q}_r(\mathcal{B}, \varepsilon, \delta) \geq \log_2 \mathsf{M}(\mathcal{B}, \|\cdot\|_{L_2}, 2\sqrt{\varepsilon}) + \log_2(1-\delta), \tag{20}$$

*where $\|\phi - \psi\|_{L_2} = \sqrt{\mathbb{E}_x(\phi(x) - \psi(x))^2}$ is the $L_2$-norm with respect to the uniform probability measure.*

The class $\mathcal{B}_{n,d}$ defined in (5) contains all Walsh functions $\{w_S\}_{|S| \leq d}$ and thus

$$\forall\ \varepsilon \in (0, \sqrt{2}), \qquad \mathsf{M}(\mathcal{B}_{n,d}, \|\cdot\|_{L_2}, \varepsilon) \geq \sum_{k=0}^{d} \binom{n}{k} \geq \frac{n^d}{d^d}, \tag{21}$$

[2] The explicit nature of the $o(1)$-terms in the exponents will be made precise in Section 3.

as $\|w_S - w_T\|_{L_2} = \sqrt{2}$ for any $S \neq T$. Combining this simple lower bound with Proposition 8 we already deduce the asymptotic sharpness of (3) as $n \to \infty$. In order to derive a sharper estimate for $\mathsf{Q}(\mathcal{F}_{n,d}, \varepsilon)$ as a function of the degree $d$, we shall prove the following improved lower bound packing numbers of $\mathcal{B}_{n,d}$ along with a matching upper bound for the metric entropy of $\mathcal{F}_{n,d}$.

**Theorem 9.** *Fix $n \in \mathbb{N}$, $d \in \mathbb{N}$ with $d \leq \log_2 n$ and $\varepsilon \in (0,1)$. Then, we have*

$$\log_2 \mathsf{M}(\mathcal{B}_{n,d}, \|\cdot\|_{L_2}, 2\varepsilon) \geq (1-\varepsilon)2^{d-2}\log_2 n - (d+1)2^{d-2} \tag{22}$$

*Moreover,*

$$\log_2 \mathsf{N}(\mathcal{F}_{n,d}, \|\cdot\|_{L_2}, \varepsilon) \leq \frac{2^{Cd}}{\varepsilon^4}\log n + \kappa(d,\varepsilon), \tag{23}$$

*where $C > 0$ is a universal constant and $\kappa(d,\varepsilon) > 0$ depends only on $d$ and $\varepsilon$.*

*Proof of Theorem 1.* Inequality (6) is a direct consequence of (19), (21) and (22), while (7) follows from (20), (21) and (22). □

Having presented Theorem 9, some observations related to the implicit dependencies in (4) are in order. In [16] it was shown that

$$\mathsf{Q}_r(\mathcal{F}_{n,d}, \varepsilon, \delta) \leq \frac{e^8 d^2}{\varepsilon^{d+1}}(B_d^{\{\pm 1\}})^{2d}\log\left(\frac{n}{\delta}\right), \tag{24}$$

where $B_d^{\{\pm 1\}}$ is an important approximation theoretic parameter called the *Bohnenblust–Hille* constant of the hypercube (see [5, 11]). While it is widely believed that $B_d^{\{\pm 1\}}$ grows at most polynomially in $d$, the best known upper bound due to [10] states that $B_d^{\{\pm 1\}} \leq \exp(C\sqrt{d\log d})$ for some universal constant $C > 0$ which, combined with (24), leads to (3). A polynomial bound on $B_d^{\{\pm 1\}}$ combined with (24) would almost match, up to a logarithmic term in the exponent, the asymptotic behavior as $d \to \infty$ of the lower bound (7) of Theorem 1. We mention at this point that we are not aware of any non-constant (as $d \to \infty$) lower bound for the constant $B_d^{\{\pm 1\}}$.

The existence of a large separated set attaining the lower bound (22) is proven via a probabilistic construction of random decision trees with prescribed depth. On the other hand, the upper bound (23) is a consequence of the deep junta theorem of [13] but, to the extent of our knowledge, had not been previously observed in the literature. It is quite surprising that while $\mathcal{F}_{n,d}$ lies in a $O_d(n^d)$-dimensional space, its metric entropy is logarithmic in the dimension of this space rather than polynomial. The reason for this is the strong restriction that $\mathcal{F}_{n,d}$ consists of functions which are bounded in $L_\infty$-norm yet it is endowed with the Hilbertian $L_2$-metric. The existence of such *small nets* is often useful in theoretical computer science and probability theory, in particular in the derandomization literature [33, 39, 28] and in the study of suprema of stochastic processes [42, Chapters 7-8].

1.4. **Exact learning.** Having established reasonable bounds on the number of queries required to learn a function in $\mathcal{F}_{n,d}$ up to error $\varepsilon > 0$, we proceed to investigate the exact case $\varepsilon = 0$. As it turns out, the number of random queries required to learn a function $f \in \mathcal{F}_{n,d}$ up to a constant error $\varepsilon \in (0,1)$ using the classical Low-Degree Algorithm [30] is in fact the same (up to constants depending only on $d$) as the number of queries required to *exactly* learn the concept class $\mathcal{F}_{n,d}$. Formally, our results is this setting are summarized in the following theorem.

**Theorem 10.** *Fix $n \in \mathbb{N}$, $d \in \{1,\ldots,n\}$ and $\delta \in (0,1)$. Then,*

$$\mathsf{Q}(\mathcal{F}_{n,d}, 0) = \sum_{j=0}^{d}\binom{n}{j} \tag{25}$$

*and there exists a universal constant $C > 0$ such that*

$$\mathsf{Q}_r(\mathcal{F}_{n,d}, 0, \delta) \leq Cd2^d n^d \log\left(\frac{n}{\delta}\right). \tag{26}$$

**Structure of the paper.** In Section 2, we prove our main lower bounds for learning, namely Proposition 8 and Theorem 9. In Section 3, we prove Theorem 2 and deduce from it the Corollaries of Section 1.2. Finally, in Section 4, we prove Theorem 10 on exact learning.

**Acknowledgements.** We are grateful to Roman Vershynin for providing many helpful pointers to the literature.

## 2. Metric entropy and query complexity

We start by formalizing the concepts introduced earlier. A *learning algorithm* on $\{-1,1\}^n$ using $Q$ queries is a mapping $H : (\{-1,1\}^n \times \mathbb{R})^Q \to L_2(\{-1,1\}^n)$ which, given input of the form $(X_1, f(X_1)),\dots,(X_Q, f(X_Q))$ produces a hypothesis function for $f$. In this terminology, the randomized query complexity $\mathsf{Q}_r(\mathcal{F},\varepsilon,\delta)$ of a class of functions $\mathcal{F}$ on the hypercube is the smallest $Q \in \mathbb{N}$ for which there exists a learning algorithm $H$ with the following property:

$$\forall\, f \in \mathcal{F}, \qquad \mathbb{P}_{X_1,\dots,X_Q \in \{-1,1\}^n}\left\{ \left\| H\Big((X_1, f(X_1)),\dots,(X_Q, f(X_Q))\Big) - f \right\|_{L_2}^2 \le \varepsilon \right\} \ge 1-\delta. \tag{27}$$

In the case of non-randomized algorithms, we need to ensure that the query points are chosen consistently with respect to the previous data. In other words, $X_1$ is always a fixed point on the hypercube and for any $q \ge 2$, there exists a function $\varphi_q : (\{-1,1\}^n \times \mathbb{R})^{q-1} \to \{-1,1\}^n$ associated to $H$ determining the $q$-th query point as a function of the previous data $X_1,\dots,X_{q-1}$ and the values $y_1,\dots,y_{q-1}$ of the unknown function on these points. Given a learning algorithm $H$ and an unknown function $f \in \mathcal{F}$, we shall denote by $X_1[f], X_2[f],\dots$ the sequence of points that $H$ queries in order to construct a hypothesis function for $f$. In this terminology, the deterministic query complexity of the class $\mathcal{F}$ is the least $Q \in \mathbb{N}$ for which there exists a learning algorithm $H$ using $Q$ queries satisfying the following property:

$$\forall\, f \in \mathcal{F}, \qquad \left\| H\Big(\big(X_1[f], f(X_1[f])\big),\dots,\big(X_Q[f], f(X_Q[f])\big)\Big) - f \right\|_{L_2}^2 \le \varepsilon. \tag{28}$$

Having properly defined these notions, we may proceed to the proof of Proposition 8. The argument relies on an information-theoretic consideration: given samples $X_1,\dots,X_Q$, the outputs $f(X_1),\dots,f(X_Q)$ provide $Q$ bits of information for $f$ and thus cannot distinguish more than $\log_2 Q$ functions which are reasonably far apart.

*Proof of Proposition 8.* Let $M = \mathsf{M}(\mathcal{B}, \|\cdot\|_{L_2}, 2\sqrt{\varepsilon})$ and consider $f_1,\dots,f_M \in \mathcal{B}$ with $\|f_i - f_j\|_{L_2} > 2\sqrt{\varepsilon}$ for all $i \ne j$. We start with the lower bound (19) in the deterministic case. Denote by $Q = \mathsf{Q}(\mathcal{B},\varepsilon)$ and let $X_1[f], X_2[f],\dots,X_Q[f]$ be samples satisfying (28) for some learning algorithm $H$ and all functions $f$ in the class $\mathcal{B}$. Consider the set

$$\Sigma \stackrel{\text{def}}{=} \Big\{ \big(f_i(X_1[f_i]),\dots,f_i(X_Q[f_i])\big) :\ i = 1,\dots,M \Big\}. \tag{29}$$

*Claim.* $|\Sigma| = M$.

*Proof.* Clearly $|\Sigma| \le M$. If $|\Sigma| < M$, then there exist $i \ne j \in \{1,\dots,M\}$ for which we have

$$\forall\, k \in \{1,\dots,Q\}, \qquad f_i(X_k[f_i]) = f_j(X_k[f_j]). \tag{30}$$

As $X_1[f_i] = X_1[f_j] \stackrel{\text{def}}{=} X_1$ by definition of $H$, (30) gives $f_i(X_1) = f_j(X_1)$ which then, by consistency of the algorithm, implies that $X_2[f_i] = X_2[f_j] \stackrel{\text{def}}{=} X_2$. Continuing iteratively, we deduce that $X_k[f_i] = X_k[f_j] \stackrel{\text{def}}{=} X_k$ for every $k \in \{1,\dots,Q\}$ and thus the common output function

$$h \stackrel{\text{def}}{=} H\Big((X_1, f_i(X_1)),\dots,(X_Q, f_i(X_Q))\Big) = H\Big((X_1, f_j(X_1)),\dots,(X_Q, f_j(X_Q))\Big) \tag{31}$$

satisfies $\|h - f_i\|_{L_2}^2 \le \varepsilon$ and $\|h - f_j\|_{L_2}^2 \le \varepsilon$ which is a contradiction as $\|f_i - f_j\|_{L_2} > 2\sqrt{\varepsilon}$. □

Finally, observe that as the class $\mathcal{B}$ consists of Boolean functions, we have the trivial inclusion $\Sigma \subseteq \{-1,1\}^Q$ which implies that $M = |\Sigma| \le 2^Q$ and the proof is complete.

In the random case, denote by $Q = \mathsf{Q}_r(\mathcal{F}, \varepsilon, \delta)$ and let $X = (X_1, \dots, X_Q)$ where $X_1, X_2, \dots$ are independent random vectors, each uniformly distributed on $\{-1,1\}^n$, satisfying (27) for some learning algorithm $H$. For every $i \in \{1, \dots, M\}$, consider the event

$$B_i \stackrel{\text{def}}{=} \left\{ \left\| H\Big( (X_1, f_i(X_1)), \dots, (X_Q, f_i(X_Q)) \Big) - f_i \right\|_{L_2}^2 > \varepsilon \right\}, \tag{32}$$

which has probability $\mathbb{P}\{B_i\} \le \delta$ by (27) and, as before, consider the (random) set

$$\Sigma(X) \stackrel{\text{def}}{=} \Big\{ \big( f_i(X_1), \dots, f_i(X_Q) \big) : \ i = 1, \dots, M \Big\}. \tag{33}$$

*Claim.* $\mathbb{E}|\Sigma(X)| \ge (1-\delta)M$.

*Proof.* Consider the partition $\{1, \dots, M\} = \sigma_1 \sqcup \dots \sqcup \sigma_{|\Sigma(X)|}$ depending on $X$ such that for every $r \in \{1, \dots, |\Sigma(X)|\}$ and all $i, j \in \sigma_r$, we have $f_i \equiv f_j$ on $\{X_1, \dots, X_Q\}$. Now, suppose that there exist two distinct $i \ne j \in \sigma_r$ such that $X \notin B_i$ and $X \notin B_j$. Then, the function

$$h \stackrel{\text{def}}{=} H\Big( (X_1, f_i(X_1)), \dots, (X_Q, f_i(X_Q)) \Big) = H\Big( (X_1, f_j(X_1)), \dots, (X_Q, f_j(X_Q)) \Big) \tag{34}$$

satisfies $\|h - f_i\|_{L_2}^2 \le \varepsilon$ and $\|h - f_j\|_{L_2}^2 \le \varepsilon$ which contradicts $\|f_i - f_j\|_{L_2} > 2\sqrt{\varepsilon}$. Therefore, for any $r$ and any $X = (X_1, \dots, X_Q)$, there exists a subset $\tau_r \subseteq \sigma_r$ with $|\tau_r| \ge |\sigma_r| - 1$ such that $X \in B_i$ for all $i \in \tau_r$. Adding up these inequalities and taking the expectation, we deduce that

$$M - \mathbb{E}|\Sigma(X)| = \mathbb{E}\Big[ \sum_{r=1}^{|\Sigma(X)|} \big( |\sigma_r| - 1 \big) \Big] \le \mathbb{E}\Big[ \sum_{r=1}^{|\Sigma(X)|} |\tau_r| \Big] \le \mathbb{E}\Big[ \sum_{r=1}^{|\Sigma(X)|} \sum_{i \in \sigma_r} \mathbf{1}_{B_i}(X) \Big] = \sum_{i=1}^{M} \mathbb{P}\{B_i\} \le \delta M, \tag{35}$$

which is the desired inequality. □

As the class $\mathcal{B}$ consists of Boolean functions, we have $(1-\delta)M \le \mathbb{E}|\Sigma(X)| \le 2^Q$. □

2.1. **Decision trees.** Before we estimate the metric entropy of the classes $\mathcal{B}_{n,d}$ and $\mathcal{F}_{n,d}$, we introduce some necessary background. Following [36, §3.2], we define a decision tree $T$ to be a representation of a function $f : \{-1,1\}^n \to \mathbb{R}$ as a rooted binary tree in which the internal nodes are labeled by Boolean variables $x_i$, $i \in \{1, \dots, n\}$, the edges are labeled by -1 and 1 and the leaves are labeled by real numbers. It is required that no Boolean variable $x_i$ appears more than once on any root-leaf path. On input $y \in \{-1,1\}^n$, the tree $T$ computes the value $f(y)$ in the following way. Starting from the root, when the computation path reaches a node labeled by $x_i$, it follows the unique edge labeled by the value $y_i \in \{-1,1\}$. The output $f(y)$ of $T$ is the label of the leaf reached by this path. It is a classical fact (see [36, Proposition 3.16]) that if a function $f$ can be represented by a decision tree of depth $d$, then $f$ has degree at most $d$.

In order to prove the lower bound (22) on the packing number of $\mathcal{B}_{n,d}$, we shall need the following combinatorial lemma on large families of sets with pairwise small intersections.

**Lemma 11.** *Fix $m, k \in \mathbb{N}$ with $k < m$ and $\varepsilon \in (0,1)$. Then, there exists $t \ge (2k)^{-k/2} m^{(1-\varepsilon)k/2}$ and subsets $\sigma_1, \dots, \sigma_t \subset \{1, \dots, m\}$ of size $k$ satisfying*

$$\forall\ i \ne j \in \{1, \dots, t\}, \qquad |\sigma_i \cap \sigma_j| < (1-\varepsilon)k. \tag{36}$$

*Proof.* We shall use the probabilistic method. Suppose that $\sigma$ is a uniformly chosen random subset of $\{1, \dots, m\}$ of cardinality $k$. Then, we have

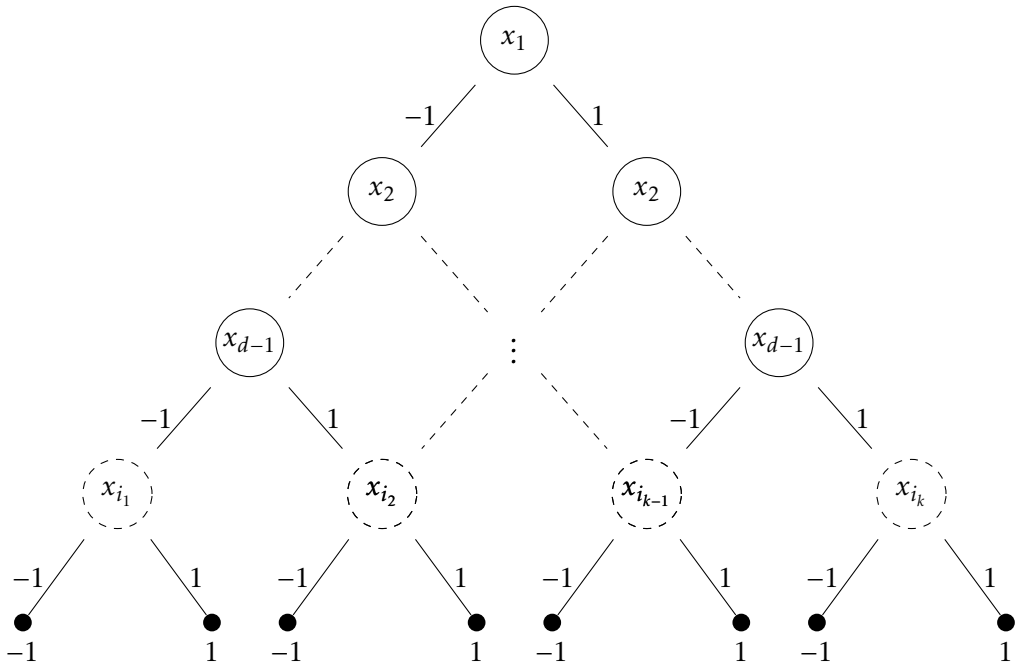
$$\mathbb{P}\Big\{ |\sigma \cap \{1, \dots, k\}| \ge (1-\varepsilon)k \Big\} = \frac{1}{\binom{m}{k}} \sum_{j \le \varepsilon k} \binom{k}{k-j} \binom{m-k}{j} \le \frac{k^k}{m^k} m^{\varepsilon k} \sum_{j \le \varepsilon k} \binom{k}{k-j} \le (2k)^k m^{-(1-\varepsilon)k}, \tag{37}$$

where we used the fact that $\frac{r^s}{s^s} \le \binom{r}{s} \le r^s$. If $\sigma_1, \dots, \sigma_t$ are i.i.d. copies of $\sigma$, then by homogeneity

$$\forall\ i \ne j \in \{1, \dots, t\}, \qquad \mathbb{P}\Big\{ |\sigma_i \cap \sigma_j| \ge (1-\varepsilon)k \Big\} \le (2k)^k m^{-(1-\varepsilon)k} \tag{38}$$

and thus

$$\mathbb{E}\Big[ \#\{ i \ne j : |\sigma_i \cap \sigma_j| \ge (1-\varepsilon)k \} \Big] \le \binom{t}{2} (2k)^k m^{-(1-\varepsilon)k} < t^2 (2k)^k m^{-(1-\varepsilon)k}. \tag{39}$$




The decision tree $T_\sigma$ corresponding to $\sigma = \{i_1 < i_2 < \cdots < i_k\}$

Therefore, if $t \le (2k)^{-k/2} m^{(1-\varepsilon)k/2}$, there exist $\sigma_1, \ldots, \sigma_t$ with the desired property. □

Equipped with Lemma 11, we proceed to the proof of the lower bound in Theorem 9.

*Proof of* (22). Let $\sigma$ be a subset of $\{d, d+1, \ldots, n\}$ of cardinality $k = 2^{d-1}$. We shall associate to $\sigma$ a Boolean function $f_\sigma : \{-1,1\}^n \to \{-1,1\}$ of degree at most $d$ represented by the decision tree $T_\sigma$ which is constructed as follows. The root of $T_\sigma$ is labeled by $x_1$ and every node which is at distance $i$ from the root is labeled by $x_{i+1}$ for $i \in \{1, \ldots, d-2\}$. If $\sigma = \{i_1, \ldots, i_k\}$, then the nodes at distance $d-1$ from the root are labeled by the distinct variables $x_{i_1}, \ldots, x_{i_k}$ in accordance with the lexicographic ordering $\le_L$, meaning that if $(\varepsilon_r(1), \ldots, \varepsilon_r(d-1)), (\varepsilon_s(1), \ldots, \varepsilon_s(d-1)) \in \{-1,1\}^{d-1}$ are the labels of the edges joining the root with the nodes labeled by $x_{i_r}$ and $x_{i_s}$, then

$$i_r \le i_s \quad \Longleftrightarrow \quad (\varepsilon_r(1), \ldots, \varepsilon_r(d-1)) \le_L (\varepsilon_s(1), \ldots, \varepsilon_s(d-1)). \tag{40}$$

Finally, if given an input $y \in \{-1,1\}^n$ the tree $T_\sigma$ queries the variable $x_{i_j}$ on the $d$-th level, then its output is $y_{i_j}$. This construction is depicted pictorially in the figure above. Observe that in this picture, the restriction (40) is equivalent to $i_1 < i_2 < \ldots < i_k$.

Using Lemma 11, we can find $t \ge 2^{-d2^{d-2}} (n-d+1)^{(1-\varepsilon)2^{d-2}} \ge 2^{-(d+1)2^{d-2}} n^{(1-\varepsilon)2^{d-2}}$ and subsets $\sigma_1, \ldots, \sigma_t$ of $\{d, d+1, \ldots, n\}$ with cardinality $d$ satisfying (36). We will show that the family of functions $f_{\sigma_1}, \ldots, f_{\sigma_t} \in \mathcal{B}_{n,d}$ is well-separated. Indeed, let $r \ne s$ and suppose that $\sigma_r = \{i_1, \ldots, i_k\}$ and $\sigma_s = \{j_1, \ldots, j_k\}$ with $i_1 < \ldots < i_k$ and $j_1 < \ldots < j_k$. Then, we have

$$\|f_{\sigma_r} - f_{\sigma_s}\|_{L_2}^2 = \frac{1}{2^{d-1}} \sum_{\ell=1}^{k} \mathbb{E}(x_{i_\ell} - x_{j_\ell})^2 = \frac{1}{2^{d-2}} |\{\ell : \ i_\ell \ne j_\ell\}| \ge \frac{2^{d-1} - |\sigma_r \cap \sigma_s|}{2^{d-2}} \overset{(36)}{\ge} 2\varepsilon \tag{41}$$

and the proof is complete. □

While Theorem 1 provides a logarithmic lower bound for the query complexity of learning $\mathcal{F}_{n,d}$ in both the query and the random example models, the upper bound (3) of [16] is currently known to hold only in the random case. Derandomizing the algorithm used there or finding a different deterministic algorithm whose query complexity is logarithmic in the dimension (and which, ideally, has reasonable running time) remains an interesting problem.

2.2. **Juntas.** A general principle in analysis on the hypercube asserts that functions whose spectrum is not spread out, effectively depend only on few variables. Many concrete instances of this phenomenon have been studied for Boolean functions, such as the important works [17, 8, 18, 25]. The definitive junta theorem for general *bounded* functions, is the following deep result of Dinur, Friedgut, Kindler and O'Donnell (DFKO) [13] (see also [37] for a quantitatively sharp statement in terms of the dependence on $d$).

**Theorem 12.** *Fix $n, d \in \mathbb{N}$, $\varepsilon > 0$ and let $f : \{-1,1\}^n \to [-1,1]$ be a function satisfying*

$$\sum_{|S|>d} \hat{f}(S)^2 \leq \exp\Big(-C(d^2 \log d)/\varepsilon^2\Big) \tag{42}$$

*for a large enough universal constant $C > 0$. Then, there exists a subset $\sigma \subseteq \{1,\ldots,n\}$ with $|\sigma| \leq \frac{2^{Cd}}{\varepsilon^4}$ and a function $g : \{-1,1\}^n \to \mathbb{R}$ depending only on the variables $(x_i)_{i\in\sigma}$ such that $\|f-g\|_{L_2} \leq \varepsilon$.*

*Proof of* (23). Let $\mathsf{m}_{d,\varepsilon}$ be the size of the smallest $\frac{\varepsilon}{4}$-net on the space of all bounded functions $h : \{-1,1\}^{k_{d,\varepsilon}} \to [-1,1]$, where $k_{d,\varepsilon} = \frac{2^{Cd+2}}{\varepsilon^4}$, equipped with the $L_2$-metric and let $\{h_1,\ldots,h_{\mathsf{m}_{d,\varepsilon}}\}$ be such a net. For a subset $\sigma \subseteq \{1,\ldots,n\}$ of cardinality $k_{d,\varepsilon}$ and $s \in \{1,\ldots,\mathsf{m}_{d,\varepsilon}\}$, denote by

$$\forall\, x \in \{-1,1\}^n, \qquad h_s^\sigma(x) = h_s\Big((x_i)_{i\in\sigma}\Big). \tag{43}$$

*Claim.* The set $\{h_s^\sigma :\ s = 1,\ldots,\mathsf{m}_{d,\varepsilon} \text{ and } \sigma \subseteq \{1,\ldots,n\} \text{ with } |\sigma| = k_{d,\varepsilon}\}$ is an $\frac{\varepsilon}{2}$-covering of $\mathcal{F}_{n,d}$.

*Proof.* Indeed, let $f : \{-1,1\}^n \to [-1,1]$ be a function of degree at most $d$. By Theorem 12, there exists a subset $\sigma \subseteq \{1,\ldots,n\}$ with $|\sigma| \leq k_{d,\varepsilon}$ and a function $g : \{-1,1\}^n \to \mathbb{R}$ depending only on the variables $(x_i)_{i\in\sigma}$ such that $\|f-g\|_{L_2} \leq \frac{\varepsilon}{4}$. Notice that without loss of generality we can assume that $g$ takes values in $[-1,1]$ as we can otherwise define $\tilde{g} : \{-1,1\}^n \to [-1,1]$ by

$$\forall\, x \in \{-1,1\}^n, \qquad \tilde{g}(x) = \begin{cases} g(x), & g(x) \in [-1,1] \\ \operatorname{sign}(g(x)), & g(x) \notin [-1,1] \end{cases} \tag{44}$$

and observe that $\|\tilde{g}-f\|_{L_2} \leq \|g-f\|_{L_2} \leq \frac{\varepsilon}{4}$. Therefore, by definition of the covering $\{h_1,\ldots,h_{\mathsf{m}_{d,\varepsilon}}\}$ there exists $s \in \{1,\ldots,\mathsf{m}_{d,\varepsilon}\}$ such that $\|g-h_s^\sigma\|_{L_2} \leq \frac{\varepsilon}{4}$ and hence $\|f-h_s^\sigma\|_{L_2} \leq \frac{\varepsilon}{2}$. □

To conclude, for every $s$ and $\sigma$, choose an arbitrary point $p_s^\sigma \in \mathcal{F}_{n,d}$ satisfying $\|p_s^\sigma - h_s^\sigma\|_{L_2} \leq \frac{\varepsilon}{2}$, provided that such exists (in the opposite case the corresponding ball can be omitted from the cover). Then,

$$\mathcal{F}_{n,d} \subseteq \bigcup_{s,\sigma} \mathrm{Ball}(h_s^\sigma, \tfrac{\varepsilon}{2}) \subseteq \bigcup_{s,\sigma} \mathrm{Ball}(p_s^\sigma, \varepsilon), \tag{45}$$

thus proving that $\mathsf{N}(\mathcal{F}_{n,d}, \|\cdot\|_{L_2}, \varepsilon) \leq \mathsf{m}_{d,\varepsilon}\binom{n}{k_{d,\varepsilon}} \leq \mathsf{m}_{d,\varepsilon} n^{\frac{2^{Cd+2}}{\varepsilon^4}}$. This concludes the proof. □

## 3. Fourier concentration and learning

To prove Theorem 2, we shall employ an algorithm of [29] with one important twist from the analysis of [16]. We include the argument in full detail for completeness.

*Proof of Theorem 2.* Fix a parameter $b \in (0,\infty)$ to be determined later and denote by

$$Q_b \stackrel{\text{def}}{=} \left\lceil \frac{2}{b^2} \log\left( \frac{2}{\delta} \sum_{r=0}^{d} \binom{n}{r} \right) \right\rceil. \tag{46}$$

Let $X_1,\ldots,X_{Q_b}$ be independent random vectors, each uniformly distributed on $\{-1,1\}^n$. For a subset $S \subseteq \{1,\ldots,n\}$ with $|S| \leq d$, consider the empirical Walsh coefficient of $f$ given by

$$\alpha_S = \frac{1}{Q_b} \sum_{j=1}^{Q_b} f(X_j) w_S(X_j). \tag{47}$$

As $\alpha_S$ is a sum of bounded i.i.d. random variables and $\mathbb{E}[\alpha_S] = \hat{f}(S)$, the Chernoff bound gives

$$\forall\ S \subseteq \{1,\dots,n\} \text{ with } |S| \le d, \qquad \mathbb{P}\Big\{|\alpha_S - \hat{f}(S)| > b\Big\} \le 2\exp(-Q_b b^2/2). \tag{48}$$

Therefore, using the union bound, we get

$$\mathbb{P}\underbrace{\Big\{|\alpha_S - \hat{f}(S)| \le b, \text{ for every subset } S \text{ with } |S| \le d\Big\}}_{G_b} \ge 1 - 2\sum_{r=0}^{d}\binom{n}{r}\exp(-Q_b b^2/2) \overset{(46)}{\ge} 1-\delta.$$

Consider the random collection of sets given by

$$\mathcal{T}_b \overset{\text{def}}{=} \Big\{S \subseteq \{1,\dots,n\} :\ |S| \le d \text{ and } |\alpha_S| \ge 2b\Big\}. \tag{49}$$

Observe that if the event $G_b$ holds, then

$$\forall\ S \notin \mathcal{T}_b \text{ with } |S| \le d, \qquad |\hat{f}(S)| \le |\alpha_S - \hat{f}(S)| + |\alpha_S| \le 3b \tag{50}$$

and

$$\forall\ S \in \mathcal{T}_b, \qquad |\hat{f}(S)| \ge |\alpha_S| - |\alpha_S - \hat{f}(S)| \ge b. \tag{51}$$

Now, consider the random function $h_b : \{-1,1\}^n \to \mathbb{R}$, given by

$$\forall\ x \in \{-1,1\}^n, \qquad h_b(x) \overset{\text{def}}{=} \sum_{S\in\mathcal{T}_b} \alpha_S w_S(x) \tag{52}$$

and write

$$\begin{aligned}\|h_b - f\|_{L_2}^2 &= \sum_{S\subseteq\{1,\dots,n\}} \left|\hat{h}_b(S) - \hat{f}(S)\right|^2 = \sum_{S\in\mathcal{T}_b} |\alpha_S - \hat{f}(S)|^2 + \sum_{S\notin\mathcal{T}_b} |\hat{f}(S)|^2 \\ &= \sum_{\substack{S\in\mathcal{T}_b\\ S\in\mathcal{S}(f)}} |\alpha_S - \hat{f}(S)|^2 + \sum_{\substack{S\in\mathcal{T}_b\\ S\notin\mathcal{S}(f)}} |\alpha_S - \hat{f}(S)|^2 + \sum_{\substack{S\notin\mathcal{T}_b\\ S\in\mathcal{S}(f),\ |S|\le d}} |\hat{f}(S)|^2 + \sum_{\substack{S\notin\mathcal{T}_b\\ S\notin\mathcal{S}(f),\ |S|\le d}} |\hat{f}(S)|^2 + \sum_{|S|>d} \hat{f}(S)^2.\end{aligned} \tag{53}$$

On the event $G_b$ we then have

$$\sum_{\substack{S\in\mathcal{T}_b\\ S\in\mathcal{S}(f)}} |\alpha_S - \hat{f}(S)|^2 + \sum_{\substack{S\notin\mathcal{T}_b\\ S\in\mathcal{S}(f),\ |S|\le d}} |\hat{f}(S)|^2 \overset{(50)}{\le} (9b^2)\cdot \#\mathcal{S}(f) \le 9b^2 m. \tag{54}$$

On the other hand, as $|\alpha_S - \hat{f}(S)| \le b \le |\hat{f}(S)|$ for $S \in \mathcal{T}_b$, we get

$$\sum_{\substack{S\in\mathcal{T}_b\\ S\notin\mathcal{S}(f)}} |\alpha_S - \hat{f}(S)|^2 + \sum_{\substack{S\notin\mathcal{T}_b\\ S\notin\mathcal{S}(f),\ |S|\le d}} |\hat{f}(S)|^2 \overset{(51)}{\le} \sum_{S\notin\mathcal{S}(f)} |\hat{f}(S)|^2 \le \eta \tag{55}$$

by the Fourier concentration property. Combining the above with the assumption that the spectrum of $f$ is $t$-concentrated up to degree $d$, we conclude that

$$\|h_b - f\|_{L_2}^2 \le \eta + t + 9b^2 m \le \eta + t + \varepsilon \tag{56}$$

for $b^2 \le \varepsilon/9m$. Plugging this choice of $b$ in (46), we get the conclusion. □

We are now well-equippped to prove Corollary 3.

*Proof of Corollary 3.* Let $f \in \mathcal{F}_{n,d} \cap \mathcal{J}_{n,k,\eta}$. Then, there exists a subset $\sigma \subseteq \{1,\dots,n\}$ with $|\sigma| \le k$ and a function $g : \{-1,1\}^n \to \mathbb{R}$ depending only on the variables $(x_i)_{i\in\sigma}$ such that $\|f-g\|_{L_2}^2 \le \eta$. Then, $f$ is $\eta$-concentrated on the collection $\mathcal{S}(f) = \{S \subseteq \sigma :\ |S| \le d\}$, as

$$\sum_{S\not\subseteq\sigma} \hat{f}(S)^2 \le \|f-g\|_{L_2}^2 \le \eta. \tag{57}$$

Similarly, the spectrum of $f$ is $\eta$-concentrated up to degree $\min\{d,k\}$ and the conclusion of the corollary follows from Theorem 2 since $\#\mathcal{S}(f) = \sum_{r=0}^{\min\{d,k\}}\binom{k}{r}$. □

We emphasize that Corollary 3 does not make any claim about the *running time* required to learn approximate juntas. This is a notoriously difficult problem even for actual juntas that has been investigated in a series of important works, see for instance [34, 26, 41].

Corollary 4 follows from Corollary 3 combined with a classical theorem of Nisan and Szegedy [35], asserting that a Boolean function of degree $d$ depends on at most $d2^{d-1}$ variables. We note in passing that this result has recently been improved in important work of Chiarelli, Hatami and Saks [9] (see also [43] for the best known value of the implicit constant) who derived the optimal conclusion that such a function only depends on $O(2^d)$ variables, but this refinement will be immaterial for our considerations.

*Proof of Corollary 4.* By [35, Theorem 1.2], we have the set inclusion $\mathcal{B}_{n,d} \subseteq \mathcal{F}_{n,d} \cap \mathcal{J}_{n,k,0}$ where $k = d2^{d-1}$. Therefore, by Corollary 3, we conclude that

$$\mathrm{Q}_r(\mathcal{B}_{n,d}, \varepsilon, \delta) \leq \frac{18}{\varepsilon} \sum_{r=0}^{d} \binom{d2^{d-1}}{r} \log\left(\frac{2}{\delta} \sum_{r=0}^{d} \binom{n}{r}\right) \leq \frac{36 \cdot d2^{d^2}}{\varepsilon} \log\left(\frac{n}{\delta}\right) \tag{58}$$

where the last inequality follows by elementary estimates. □

We now proceed to prove Corollary 5, which relies on Theorem 2 and [13].

*Proof of Corollary 5.* Let $f \in \mathcal{F}_{n,d}(t)$ and $\eta \geq \frac{Cd^2 \log d}{\log(1/t)}$ so that

$$\sum_{|S|>d} \hat{f}(S)^2 \leq t \leq \exp\Big(-C(d^2 \log d)/\eta\Big). \tag{59}$$

Instead of using Theorem 12 directly, we will use a stronger statement from its proof. In [13, p. 405], it was shown that there exists a function $h$ of degree at most $d$ which depends only on the variables $(x_i)_{i\in\sigma}$ for a subset $\sigma \subseteq \{1,\dots,n\}$ with $|\sigma| \leq 2^{O(d)}/\eta^2$ such that $\|f-h\|_{L_2}^2 \leq \eta$. Choosing $\mathcal{S}(f) = \{S \subseteq \sigma : \ |S| \leq d\}$, we deduce that $f$ is $\eta$-concentrated up to degree $d$ and on the collection $\mathcal{S}(f)$. The conclusion follows from Theorem 2 as $\#\mathcal{S}(f) \leq 2^{O(d^2)}/\eta^{2d}$. □

We note in passing that one can replace the use of the DFKO theorem with a result of O'Donnell and Zhao [37, Corollary 3.5] to improve the dependence of $\eta$ on $d$ to $\eta \geq \frac{Cd^2}{\log(1/t)}$ at the expense of an exponentially worse dependence of the complexity on $d$ and $\varepsilon$.

**Remark 13.** *Choosing $t = 0$ in Corollary 5 provides a different proof of the main result of* [16]*, i.e. that $\mathrm{Q}_r(\mathcal{F}_{n,d}, \varepsilon, \delta) = O_{d,\varepsilon,\delta}(\log n)$, using the DFKO theorem. Indeed, by Theorem 12, we have $\mathcal{F}_{n,d} = \mathcal{F}_{n,d} \cap \mathcal{J}_{n,k(d,\eta),\eta}$ for any $\eta > 0$, where $k(d,\eta) = \lceil 2^{Cd}/\eta^2 \rceil$. Plugging this in the bound* (12) *and optimizing over $\eta$, we deduce that there exists a universal constant $C > 0$ such that*

$$\forall\ \varepsilon, \delta \in (0,1), \qquad \mathrm{Q}_r(\mathcal{F}_{n,d}, \varepsilon, \delta) \leq \frac{2^{Cd^2}}{\varepsilon^{2d+1}} \log\left(\frac{n}{\delta}\right). \tag{60}$$

*It is worth emphasizing that, while* (60) *captures the correct dependence on the dimension $n$, it is asymptotically worse than the bounds* (3), (24) *both as $d \to \infty$ and as $\varepsilon \to 0^+$.*

*On the other hand, inequality* (24) *is a special case of the bound* (10)*, up to lower order terms depending only on the degree $d$. Indeed, if for a function $f \in \mathcal{F}_{n,d}$ we define $\mathcal{S}(f)$ to be the collection of subsets $S$ of $\{1,\dots,n\}$ for which $|\hat{f}(S)| \geq \varepsilon^{\frac{d+1}{2}} B_d^{-d}$, then we have*

$$\#\mathcal{S}(f) \leq \varepsilon^{-d} B_d^{\frac{2d^2}{d+1}} \sum_{S\in\mathcal{S}(f)} |\hat{f}(S)|^{\frac{2d}{d+1}} \leq \frac{B_d^{2d}}{\varepsilon^d} \tag{61}$$

*and*

$$\sum_{S\notin\mathcal{S}(f)} \hat{f}(S)^2 \leq \varepsilon B_d^{-\frac{2d}{d+1}} \sum_{S\notin\mathcal{S}(f)} |\hat{f}(S)|^{\frac{2d}{d+1}} \leq \varepsilon \tag{62}$$

*by two applications of the Bohnenblust–Hille inequality. Thus* (24) *follows from* (10) *with $t = 0$.*

To prove Corollary 6, we will use a deep junta theorem of Bourgain [8, Proposition]. The quantitative version which we employ below follows from [23, Theorem 7.1] (see also [24, 12]).

**Theorem 14.** *Fix $n, d \in \mathbb{N}$ and $t \in (0,1)$. For any Boolean function $f \in \mathcal{B}_{n,d}(t)$ and any parameter $\eta \geq \exp\left(C\sqrt{\log(2/t)\log\log d}\right)\left(t\sqrt{d} + \frac{1}{2^d}\right)$ there exists a collection of subsets $\mathcal{S}(f)$ of $\{1,\ldots,n\}$ with $\#\mathcal{S}(f) \leq 2^{O(d^2)}$ such that the spectrum of $f$ is $\eta$-concentrated on $\mathcal{S}(f)$.*

To see how Theorem 14 follows from [23, Theorem 7.1], choose $\beta = 2^{-\Omega(d)}$ in that statement and consider $\mathcal{S}(f)$ to be the collection of subsets $S \subseteq \{1,\ldots,n\}$ with $|S| \leq d$ and $S \subseteq J_\beta$. The fact that $\#\mathcal{S}(f) \leq 2^{O(d^2)}$ then follows since $|J_\beta| \leq 2^{O(d)}$ and the Fourier concentration property on $\mathcal{S}(f)$ follows from the conclusion of [23, Theorem 7.1]. We note that the lower order terms on the size of $\eta$ with respect to $t, d$ can be removed from Theorem 14 in view of a result of Kindler and O'Donnell [24] at the expense of a worse dependence of $\#\mathcal{S}(f)$ on $d$.

*Proof of Corollary 6.* Let $f \in \mathcal{B}_{n,d}(t)$. If $\eta \geq t^{1+o(1)} d^{\frac{1}{2}+o(1)}$ in the precise sense of Theorem 14, we have that the spectrum of $f$ is $\eta$-concentrated on a collection of subsets with cardinality $2^{O(d^2)}$. Therefore the conclusion follows from Theorem 2 since also $t = o(\eta)$ as $t \to 0^+$. □

**Remark 15.** *It is worth emphasizing that the constraint $\eta \geq t\sqrt{d}$ which follows from* [8, 24] *is in some sense optimal if one wishes to learn the class $\mathcal{B}_{n,d}(t)$ from logarithmically many samples. A linear threshold function (LTF) is a Boolean function of the form $f(x) = \operatorname{sign}\langle x, \theta\rangle$, where $x \in \{-1,1\}^n$ and $\theta \in \mathbb{S}^{n-1}$ is a fixed vector. A well-known theorem of Peres* [38] *(see also* [2]*) asserts that any LTF on $n$ variables belongs in $\mathcal{B}_{n,\Omega(1/t^2)}(t)$ for every $t \in (0,1)$. We shall argue that there exist $2^{\Omega(n)}$ LTFs which are pairwise $\Omega(1)$-apart which, in vue of Proposition 8, will imply that the class of LTFs requires at least $\Omega(n)$ samples to be learned with accuracy $\frac{1}{4}$ and confidence $\frac{3}{4}$. Equivalently, we will show that there exist $N = 2^{\Omega(n)}$ vectors $\theta_1, \ldots, \theta_N \in \mathbb{S}^{n-1}$ such that*

$$\forall\ i \neq j, \qquad \left\|\operatorname{sign}\langle x, \theta_i\rangle - \operatorname{sign}\langle x, \theta_j\rangle\right\|_{L_2}^2 = 8\mathbb{P}\Big\{\Big(\langle x, \theta_i\rangle, \langle x, \theta_j\rangle\Big) \in U\Big\} = \Omega(1), \tag{63}$$

*where $U$ is the second quadrant $\{(s,t) : s \leq 0 \leq t\}$. The corresponding estimate in Gauss space follows from classical computations (see, e.g.,* [27, Lemme 1]*) as*

$$\forall\ u, v \in \mathbb{S}^{n-1}, \qquad \left\|\operatorname{sign}\langle g, u\rangle - \operatorname{sign}\langle g, v\rangle\right\|_{L_2}^2 = 2 - 2\mathbb{E}\Big[\operatorname{sign}\Big(\langle g, u\rangle \cdot \langle g, v\rangle\Big)\Big] = 2 - \frac{2}{\pi}\arcsin\langle u, v\rangle, \tag{64}$$

*where $g \sim N(0, \mathrm{Id}_n)$ is a standard Gaussian random vector, and thus it suffices to choose the vectors $\{\theta_i\}_{i=1}^N$ to form an $\Omega(1)$-net in the unit sphere. To pass from the Gaussian statement implied by* (64) *to the corresponding discrete inequality* (63) *we shall use a classical (multivariate) Berry–Esseen theorem (see, e.g.,* [3, Theorem 1.1]*). In order to apply this result to the random vectors $(\langle x, \theta_i\rangle, \langle x, \theta_j\rangle)$, it suffices to find an $\Omega(1)$-separated set $\{\theta_i\}_{i=1}^N$ in $\mathbb{S}^{n-1}$ with $N = 2^{\Omega(n)}$ points such that $\|\theta_i\|_{\ell_\infty^n} \leq \tau$ for some small enough universal constant $\tau > 0$. The existence of such a set can be proven by the probabilistic method in view of standard concentration estimates of $\ell_p^n$-norms on $\ell_q^n$-spheres, see for instance* [40, Remark 2 in p. 223] *and* [1, Theorem 1].

Finally, we prove Corollary 7 on the complexity of constant depth circuits.

*Proof of Corollary 7.* By [30, Main Lemma] (see also the exposition in [36, Section 4.5] and a slight refinement in [20]), every $f \in \mathcal{C}_{n,d,s}$ also belongs in $\mathcal{B}_{n,m(d,s,t)}(t)$ for every $t > 0$, where $m(d,s,t) = O(\log(s/t))^d$. The conclusion follows by choosing $\eta = \varepsilon$ and $t$ small enough such that $\varepsilon \geq t^{1+o(1)} O(\log(s/t))^{d/2+o(1)}$ and applying Corollary 6. □

## 4. Exact learning

In this section, we shall prove Theorem 10. We start with the deterministic case.

*Proof of* (25). Let $Q = \mathrm{Q}(\mathcal{F}_{n,d}, 0)$. For the upper bound on $Q$, consider an enumeration $X_1, \ldots, X_k$ of the points in the closed Hamming Ball$(\mathbf{1}, d)$, where $\mathbf{1} = (1, \ldots, 1)$ and $k = \sum_{j=0}^{d} \binom{n}{j}$.

*Claim.* If $f \in \mathcal{F}_{n,d}$, then the values $f(X_1), \ldots, f(X_k)$ completely determine $f$.

*Proof.* As usual, for $i \in \{1, \ldots, n\}$ we denote by

$$\forall\ x \in \{-1,1\}^n, \qquad \partial_i f(x) \stackrel{\text{def}}{=} \frac{f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, -x_i, \ldots, x_n)}{2} \tag{65}$$

the discrete partial derivative of $f$. It is straightforward to see that if $f = \sum_S c_s w_S$, then

$$\forall\ x \in \{-1,1\}^n, \qquad \partial_i f(x) = \sum_{S:\ i \in S} c_S w_S(x). \tag{66}$$

In particular, if $f$ has degree at most $d$ and $S = \{i_1, \ldots, i_d\}$ has cardinality $d$, then

$$\partial_{i_1} \circ \cdots \circ \partial_{i_d} f(\mathbf{1}) = c_S. \tag{67}$$

On the other hand, (65) implies that $\partial_{i_1} \circ \cdots \circ \partial_{i_d} f(\mathbf{1})$ is a linear combination of $f(X_1), \ldots, f(X_k)$. In other words, knowing $f(X_1), \ldots, f(X_k)$, we can reconstruct the top-order Walsh coefficients $\{c_S\}_{|S|=d}$. To conclude, we consider the function $f - \sum_{|S|=d} c_S w_S$ and iterate. □

Therefore, the claim implies that if the algorithm queries the values of $f$ at $X_1, \ldots, X_k$, then the function can be fully reconstructed, thus proving that $Q \le k = \sum_{j=0}^{d} \binom{n}{j}$.

The lower bound is a simple dimension counting argument. Assume, for contradiction, that $\mathcal{F}_{n,d}$ can be learned exactly using $Q$ queries where $Q < k = \sum_{j=0}^{d} \binom{n}{j}$. Then, for any fixed points $X_1, \ldots, X_Q \in \{-1,1\}^n$, the linear system

$$\forall\ r = 1, \ldots, Q, \qquad \sum_{|S| \le d} c_S w_S(X_r) = 0 \tag{68}$$

with $k$ unknowns $\{c_S\}_{|S| \le d}$ and $Q$ equations has at least one nonzero solution. In other words, there exists a nonzero function $g \in \mathcal{F}_{n,d}$ which vanishes on $\{X_1, \ldots, X_Q\}$. If $X_1, \ldots, X_Q$ are the points queried by the algorithm in order to learn $g$, then (68) shows the same points need to be queried to learn the zero function $\mathbf{0}$. This is a contradictions as $H$ would produce the same hypothesis function for both and $g$ is not identically zero. □

Finally, we prove the upper bound (26) for the random example model.

*Proof of* (26). For points $X_1, \ldots, X_Q$ on the hypercube, consider the (linear) evaluation operator $\Phi_{X_1,\ldots,X_Q} : \mathcal{F}_{n,d} \to \mathbb{R}^Q$ given by $\Phi_{X_1,\ldots,X_Q}(f) = (f(X_1), \ldots, f(X_Q))$. In order to prove the upper bound on the query complexity of the random example model without error, it suffices to show that if $Q$ is large enough and $X_1, \ldots, X_Q$ are independent and uniformly distributed random vectors on $\{-1,1\}^n$, then the operator $\Phi_{X_1,\ldots,X_Q}$ is injective with high probability. Indeed, if this is the case then the values of any function $f \in \mathcal{F}_{n,d}$ on a random sequence of samples uniquely determine $f$ with high probability and thus the function can be fully reconstructed by solving a system of linear equations with respect to its Walsh coefficients.

To show that $\Phi_{X_1,\ldots,X_Q}$ is injective with high probability, fix points $P_1, \ldots, P_q \in \{-1,1\}^n$ and let $X$ be a uniform random vector on the hypercube. Suppose that $\Phi_{P_1,\ldots,P_q}$ is not injective and choose a nonzero function $g \in \ker \Phi_{P_1,\ldots,P_q}$. Then, we have

$$\mathbb{P}\Big\{\dim\ker \Phi_{P_1,\ldots,P_q,X} < \dim\ker \Phi_{P_1,\ldots,P_q}\Big\} \ge \mathbb{P}\{g(X) \ne 0\} \ge \frac{1}{2^d}, \tag{69}$$

where the last inequality is a classical property satisfied by nonzero functions of degree at most $d$ which can be proven inductively, see [36, Lemma 3.5].

To conclude the proof, consider $Q > k \stackrel{\text{def}}{=} \sum_{j=0}^{d} \binom{n}{j}$ and let $X_1, \ldots, X_Q$ be independent uniformly random points on the hypercube. Suppose that the operator $\Phi_{X_1,\ldots,X_Q}$ is not injective.

Then, at least $Q-k+1$ steps in the following chain of inequalities are in fact equalities:

$$k \geq \dim\ker\Phi_{X_1} \geq \dim\ker\Phi_{X_1,X_2} \geq \cdots \geq \dim\ker\Phi_{X_1,\ldots,X_Q}. \tag{70}$$

By inequality (69) and the independence of $X_1,\ldots,X_Q$, we deduce that

$$\mathbb{P}\Big\{\Phi_{X_1,\ldots,X_Q} \text{ is not injective}\Big\} \leq \binom{Q}{Q-k+1}\Big(1-\tfrac{1}{2^d}\Big)^{Q-k+1} \leq Q^{k-1}\Big(1-\tfrac{1}{2^d}\Big)^{Q-k+1} \leq (2Q)^{k-1}\Big(1-\tfrac{1}{2^d}\Big)^{Q}.$$

Choosing $Q = C2^d k\log\big(\frac{k}{\delta}\big)$ for a large enough universal constant $C>1$ ensures that $\Phi_{X_1,\ldots,X_Q}$ is injective with probability at least $1-\delta$, thus completing the proof as $k \leq (d+1)n^d$. □

**Remark 16.** *We point out that the query complexity estimate* (26) *can be realized algorithmically. At every step of the algorithm, one has to compute the rank of the matrix $\Phi_{X_1,\ldots,X_q}$ until it becomes full-rank. Then, the unknown function $f$ can be recovered by solving a system of linear equations with respect to its Walsh coefficients.*

**Remark 17.** *Throughout this paper and* [16]*, we have been studying learning algorithms for $\mathcal{F}_{n,d}$ and $\mathcal{B}_{n,d}$ equipped with the Hilbertian $L_2$-metric. This choice allows us to use Parseval's identity and thus exploit properties of individual Walsh coefficients to study the distance between $f$ and the hypothesis function $h$. However as the constructed hypothesis functions $h$ are always of degree at most $d$ themselves, this can be generalized to any $L_p$ norm, where $0<p<\infty$, since these are equivalent to the $L_2$ norm on the space of degree-$d$ polynomials up to constants depending only on $d$ (see* [36, §9.5] *and* [7, 6, 15] *for more on moment comparison of polynomials).*

## References


[1] Juan Arias-de Reyna, Keith Ball, and Rafael Villa. Concentration of the distance in finite-dimensional normed spaces. *Mathematika*, 45(2):245–252, 1998.

[2] Itai Benjamini, Gil Kalai, and Oded Schramm. Noise sensitivity of Boolean functions and applications to percolation. *Inst. Hautes Études Sci. Publ. Math.*, (90):5–43 (2001), 1999.

[3] Vidmantas Bentkus. A Lyapunov type bound in $\mathbf{R}^d$. *Teor. Veroyatn. Primen.*, 49(2):400–410, 2004.

[4] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. Assoc. Comput. Mach.*, 36(4):929–965, 1989.

[5] Henri Frederic Bohnenblust and Einar Hille. On the absolute convergence of Dirichlet series. *Ann. of Math. (2)*, 32(3):600–622, 1931.

[6] Christer Borell. On polynomial chaos and integrability. *Probab. Math. Statist.*, 3(2):191–203, 1984.

[7] Jean Bourgain. Walsh subspaces of $L^p$-product spaces. In *Seminar on Functional Analysis, 1979–1980 (French)*, pages Exp. No. 4A, 9. École Polytech., Palaiseau, 1980.

[8] Jean Bourgain. On the distributions of the Fourier spectrum of Boolean functions. *Israel J. Math.*, 131:269–276, 2002.

[9] John Chiarelli, Pooya Hatami, and Michael Saks. An asymptotically tight bound on the number of relevant variables in a bounded degree Boolean function. *Combinatorica*, 40(2):237–244, 2020.

[10] Andreas Defant, Mieczysław Mastyło, and Antonio Pérez. On the Fourier spectrum of functions on Boolean cubes. *Math. Ann.*, 374(1-2):653–680, 2019.

[11] Andreas Defant and Pablo Sevilla-Peris. The Bohnenblust-Hille cycle of ideas from a modern point of view. *Funct. Approx. Comment. Math.*, 50(1, [2013 on table of contents]):55–127, 2014.

[12] Ilias Diakonikolas, Ragesh Jaiswal, Rocco A. Servedio, Li-Yang Tan, and Andrew Wan. Noise stable halfspaces are close to very small juntas. *Chic. J. Theoret. Comput. Sci.*, pages Article 4, 13, 2015.

[13] Irit Dinur, Ehud Friedgut, Guy Kindler, and Ryan O'Donnell. On the Fourier tails of bounded functions over the discrete cube. *Israel J. Math.*, 160:389–412, 2007.

[14] Richard M. Dudley, Sanjeev R. Kulkarni, Thomas Richardson, and Ofer Zeitouni. A metric entropy bound is not sufficient for learnability. *IEEE Trans. Inform. Theory*, 40(3):883–885, 1994.

[15] Alexandros Eskenazis and Paata Ivanisvili. Polynomial inequalities on the Hamming cube. *Probab. Theory Related Fields*, 178(1-2):235–287, 2020.

[16] Alexandros Eskenazis and Paata Ivanisvili. Learning low-degree functions from a logarithmic number of random queries. To appear in *Proceedings of STOC 2022*. Preprint available at `https://arxiv.org/abs/2109.10162`, 2021.

[17] Ehud Friedgut. Sharp thresholds of graph properties, and the $k$-sat problem. *J. Amer. Math. Soc.*, 12(4):1017–1054, 1999. With an appendix by Jean Bourgain.

[18] Ehud Friedgut, Gil Kalai, and Assaf Naor. Boolean functions whose Fourier transform is concentrated on the first two levels. *Adv. in Appl. Math.*, 29(3):427–437, 2002.

[19] Johan Håstad. Computational limitations of small-depth circuits. MIT Press, 1987.
[20] Johan Håstad. A slight sharpening of LMN. *J. Comput. System Sci.*, 63(3):498–508, 2001.
[21] Siddharth Iyer, Anup Rao, Victor Reis, Thomas Rothvoss, and Amir Yehudayoff. Tight bounds on the Fourier growth of bounded functions on the hypercube. To appear in ECCC 2021. Preprint available at https://arxiv.org/abs/2107.06309, 2021.
[22] Michael Kharitonov. Cryptographic hardness of distribution-specific learning. In *STOC: ACM Symposium on Theory of Computing (STOC)*, 1993.
[23] Subhash Khot and Assaf Naor. Nonembeddability theorems via Fourier analysis. *Math. Ann.*, 334(4):821–852, 2006.
[24] Guy Kindler and Ryan O'Donnell. Gaussian noise sensitivity and Fourier tails. In *2012 IEEE 27th Conference on Computational Complexity—CCC 2012*, pages 137–147. IEEE Computer Soc., Los Alamitos, CA, 2012.
[25] Guy Kindler and Shmuel Safra. Noise-resistant Boolean functions are juntas. Manuscript, 2002.
[26] Mihail N. Kolountzakis, Richard J. Lipton, Evangelos Markakis, Aranyak Mehta, and Nisheeth K. Vishnoi. On the Fourier spectrum of symmetric Boolean functions. *Combinatorica*, 29(3):363–387, 2009.
[27] Jean-Louis Krivine. Constantes de Grothendieck et fonctions de type positif sur les sphères. *Adv. in Math.*, 31(1):16–30, 1979.
[28] Andrey Kupavskii and Nikita Zhivotovskiy. When are epsilon-nets small? *J. Comput. System Sci.*, 110:22–36, 2020.
[29] Eyal Kushilevitz and Yishay Mansour. Learning decision trees using the Fourier spectrum. *SIAM J. Comput.*, 22(6):1331–1348, 1993.
[30] Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, Fourier transform, and learnability. *J. Assoc. Comput. Mach.*, 40(3):607–620, 1993.
[31] Nathan Linial, Yishay Mansour, and Ronald L. Rivest. Results on learnability and the Vapnik-Chervonenkis dimension. *Inform. and Comput.*, 90(1):33–49, 1991.
[32] Yishay Mansour. *Learning Boolean Functions via the Fourier Transform*, pages 391–424. Springer US, Boston, MA, 1994.
[33] Jiří Matoušek. Derandomization in computational geometry. *J. Algorithms*, 20(3):545–580, 1996.
[34] Elchanan Mossel, Ryan O'Donnell, and Rocco P. Servedio. Learning juntas. In *Proceedings of the Thirty-Fifth Annual ACM Symposium on Theory of Computing*, pages 206–212. ACM, New York, 2003.
[35] Noam Nisan and Márió Szegedy. On the degree of Boolean functions as real polynomials. volume 4, pages 301–313. 1994. Special issue on circuit complexity (Barbados, 1992).
[36] Ryan O'Donnell. *Analysis of Boolean functions*. Cambridge University Press, New York, 2014.
[37] Ryan O'Donnell and Yu Zhao. Polynomial bounds for decoupling, with applications. In *31st Conference on Computational Complexity*, volume 50 of *LIPIcs. Leibniz Int. Proc. Inform.*, pages Art. No. 24, 18. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, 2016.
[38] Yuval Peres. Noise stability of weighted majority. In *In and out of equilibrium 3. Celebrating Vladas Sidoravicius*, volume 77 of *Progr. Probab.*, pages 677–682. Birkhäuser/Springer, Cham, [2021] ©2021.
[39] Yuval Rabani and Amir Shpilka. Explicit construction of a small $\epsilon$-net for linear threshold functions. *SIAM J. Comput.*, 39(8):3501–3520, 2010.
[40] Gideon Schechtman and Joel Zinn. On the volume of the intersection of two $L_p^n$ balls. *Proc. Amer. Math. Soc.*, 110(1):217–224, 1990.
[41] Amir Shpilka and Avishay Tal. On the minimal Fourier degree of symmetric Boolean functions. *Combinatorica*, 34(3):359–377, 2014.
[42] Roman Vershynin. *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018. An introduction with applications in data science, With a foreword by Sara van de Geer.
[43] Jake Wellens. A tighter bound on the number of relevant variables in a bounded degree boolean function. Preprint available at https://arxiv.org/abs/1903.08214, 2019.

(A. E.) CNRS, Institut de Mathématiques de Jussieu, Sorbonne Université, France and Trinity College, University of Cambridge, UK.
*Email address*: alexandros.eskenazis@imj-prg.fr, ae466@cam.ac.uk

(P. I.) Department of Mathematics, University of California, Irvine, Irvine, CA 92617, USA.
*Email address*: pivanisv@uci.edu

(L. S.) Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, UK.
*Email address*: ls909@cam.ac.uk