LEARNING LOW-DEGREE FUNCTIONS FROM A LOGARITHMIC NUMBER OF RANDOM QUERIES

ALEXANDROS ESKENAZIS AND PAATA IVANISVILI

Abstract. We prove that every bounded function $f:\{-1,1\}^n\to[-1,1]$ of degree at most d can be learned with L_2-accuracy ε and confidence $1-\delta$ from $\log(\frac{n}{\delta})\,\varepsilon^{-d-1}C^{d^{3/2}\sqrt{\log d}}$ random queries, where $C>1$ is a universal finite constant.

2020 Mathematics Subject Classification. Primary: 06E30; Secondary: 42C10, 68Q32.
Key words. Discrete hypercube, learning theory, Bohnenblust–Hille inequality.

1. Introduction

Every function $f:\{-1,1\}^n\to\mathbb{R}$ admits a unique Fourier–Walsh expansion of the form

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

where $w_S(x)=\prod_{i\in S}x_i$ and the Fourier coefficients $\hat{f}(S)$ are given by

$$\forall\, S\subseteq\{1,\ldots,n\},\qquad \hat{f}(S)=\frac{1}{2^n}\sum_{y\in\{-1,1\}^n}f(y)w_S(y). \tag{2}$$

We say that f has degree at most $d\in\{1,\ldots,n\}$ if $\hat{f}(S)=0$ for every subset S with $|S|>d$.

1.1. **Learning functions on the hypercube.** Let $\mathcal{C}$ be a class of functions $f:\{-1,1\}^n\to\mathbb{R}$ on the n-dimensional discrete hypercube. The problem of learning the class $\mathcal{C}$ can be described as follows: given a source of *examples* $(x,f(x))$, where $x\in\{-1,1\}^n$, for an unknown function $f\in\mathcal{C}$, compute a *hypothesis* function $h:\{-1,1\}^n\to\mathbb{R}$ which is a good approximation of f up to a given error in some prescribed metric. In this paper we will be interested in the *random query model* with L_2-error, in which we are given N independent examples $(x,f(x))$, each chosen uniformly at random from the discrete hypercube $\{-1,1\}^n$, and we want to efficiently construct a (random) function $h:\{-1,1\}^n\to\mathbb{R}$ such that $\|h-f\|_{L_2}^2<\varepsilon$ with probability at least $1-\delta$, where $\varepsilon,\delta\in(0,1)$ are given accuracy and confidence parameters. The goal is to construct a randomized algorithm which produces the hypothesis function h from a minimal number N of examples.

The above very general problem has been studied for decades in computational learning theory and many results are known[1], primarily for various classes $\mathcal{C}$ of structured Boolean functions $f:\{-1,1\}^n\to\{-1,1\}$. Already since the late 1980s, researchers used the Fourier–Walsh expansion (1) to design such learning algorithms (see the survey [14]). Perhaps the most classical of these is the *Low-Degree Algorithm* of Linial, Mansour and Nisan [12] who showed that for the class $\mathcal{C}_b^d$ of all *bounded* functions $f:\{-1,1\}^n\to[-1,1]$ of degree at most d there exists an algorithm which produces an ε-approximation of f with probability at least $1-\delta$ using $N=\frac{2n^d}{\varepsilon}\log(\frac{2n^d}{\delta})$ samples. In this generality, the $O_{\varepsilon,\delta,d}(n^d\log n)$ estimate of [12] was the state of the art until the recent work [11] of Iyer, Rao, Reis, Rothvoss and Yehudayoff who employed analytic techniques to derive new bounds on the ℓ_1-size of the Fourier spectrum of bounded

A. E. was 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.

[1]We will by no means attempt to survey this (vast) field, so we refer the interested reader to the relevant chapters of O'Donnell's book [15] and the references therein.

functions (see also Section 3) and used these estimates to show that $N = O_{\varepsilon,\delta,d}(n^{d-1}\log n)$ examples suffice to learn $\mathcal{C}_b^d$. The goal of the present paper is to further improve this result and show that in fact $N = O_{\varepsilon,\delta,d}(\log n)$ samples suffice for this purpose.

Theorem 1. *Fix $\varepsilon,\delta \in (0,1)$, $n \in \mathbb{N}$, $d \in \{1,\ldots,n\}$ and a bounded function $f : \{-1,1\}^n \to [-1,1]$ of degree at most d. If $N \in \mathbb{N}$ satisfies*

$$N \geq \min\left\{\frac{\exp(Cd^{3/2}\sqrt{\log d})}{\varepsilon^{d+1}}, \frac{4dn^d}{\varepsilon}\right\}\log\left(\frac{n}{\delta}\right), \tag{3}$$

where $C \in (0,\infty)$ is a large numerical constant, then N uniformly random independent queries of pairs $(x, f(x))$, where $x \in \{-1,1\}^n$, suffice for the construction of a random function $h : \{-1,1\}^n \to \mathbb{R}$ satisfying the condition $\|h - f\|_{L_2}^2 < \varepsilon$ with probability at least $1-\delta$.

The proof of Theorem 1 relies on some important approximation theoretic estimates going back to the 1930s which we shall now describe (see also [9]). To the best of our knowledge, these tools had not yet been exploited in the computational learning theory literature.

1.2. **The Fourier growth of Walsh polynomials in $\ell_{\frac{2d}{d+1}}$.** Estimates for the growth of coefficients of polynomials as a function of their degree and their maximum on compact sets go back to the early days of approximation theory (see [5]). A seminal result of this nature is Littlewood's celebrated $\frac{4}{3}$-inequality [13] for bilinear forms which was later generalized by Bohnenblust and Hille [4] for multilinear forms on the torus $\mathbb{T}^n$ or the unit square $[-1,1]^n$. By means of polarization, one can use this multilinear estimate to derive an inequality for polynomials which reads as follows[2]. For every $\mathbb{K} \in \{\mathbb{R},\mathbb{C}\}$ and $d \in \mathbb{N}$, there exists $B_d^{\mathbb{K}} \in (0,\infty)$ such that for every $n \in \mathbb{N}$ and every coefficients $c_\alpha \in \mathbb{K}$, where $\alpha \in (\mathbb{N}\cup\{0\})^n$ with $|\alpha| \leq d$, we have

$$\left(\sum_{|\alpha|\leq d} |c_\alpha|^{\frac{2d}{d+1}}\right)^{\frac{d+1}{2d}} \leq B_d^{\mathbb{K}} \max\left\{\left|\sum_{|\alpha|\leq d} c_\alpha x^\alpha\right| : x \in \mathbb{K}^n \text{ with } \|x\|_{\ell_\infty^n(\mathbb{K})} \leq 1\right\}. \tag{4}$$

Moreover, $\frac{2d}{d+1}$ is the smallest exponent for which the optimal constant in (4) is independent of the number of variables n of the polynomial. The exact asymptotics of the constants $B_d^{\mathbb{R}}$ and $B_d^{\mathbb{C}}$ remain unknown, however it is known that there is a significant gap between $B_d^{\mathbb{R}}$ and $B_d^{\mathbb{C}}$, namely that $\limsup_{d\to\infty}(B_d^{\mathbb{R}})^{1/d} = 1+\sqrt{2}$ whereas $B_d^{\mathbb{C}} \leq C^{\sqrt{d\ln d}}$ for a finite constant $C > 1$ (see [7, 1, 9, 6, 8] for these and other important advances of the last decade). Restricting inequality (4) to real *multilinear* polynomials, convexity shows that the maximum on the right-hand side is attained at a point $x \in \{-1,1\}^n$, which, in view of (1), makes (4) an estimate for the Fourier–Walsh growth of functions on the discrete hypercube. We shall denote by $B_d^{\{\pm 1\}}$ the corresponding optimal constant (first explicitly investigated by Blei in [3, p. 175]), that is, the least constant such that for every $n \in \mathbb{N}$ and every function $f : \{-1,1\}^n \to \mathbb{R}$ of degree at most d,

$$\left(\sum_{S\subseteq\{1,\ldots,n\}} |\hat{f}(S)|^{\frac{2d}{d+1}}\right)^{\frac{d+1}{2d}} \leq B_d^{\{\pm 1\}} \|f\|_{L_\infty}. \tag{5}$$

The best known quantitative result in this setting is due to Defant, Mastyło and Pérez [8] who showed that $B_d^{\{\pm 1\}} \leq \exp(\kappa\sqrt{d\log d})$ for a universal constant $\kappa \in (0,\infty)$. The main contribution of this work is the following theorem relating the growth of the constant $B_d^{\{\pm 1\}}$ and learning.

Theorem 2. *Fix $\varepsilon,\delta \in (0,1)$, $n \in \mathbb{N}$, $d \in \{1,\ldots,n\}$ and a bounded function $f : \{-1,1\}^n \to [-1,1]$ of degree at most d. If $N \in \mathbb{N}$ satisfies*

$$N \geq \frac{e^8 d^2}{\varepsilon^{d+1}}(B_d^{\{\pm 1\}})^{2d}\log\left(\frac{n}{\delta}\right), \tag{6}$$

[2] For $\alpha = (\alpha_1,\ldots,\alpha_n) \in (\mathbb{N}\cup\{0\})^n$, we use the standard notations $|\alpha| = \alpha_1 + \cdots + \alpha_n$ and $x^\alpha = x_1^{\alpha_1}\cdots x_n^{\alpha_n}$.

then given N uniformly random independent queries of pairs $(x, f(x))$, where $x \in \{-1,1\}^n$, one can construct a random function $h: \{-1,1\}^n \to \mathbb{R}$ satisfying $\|h-f\|_{L_2}^2 < \varepsilon$ with probability at least $1-\delta$.

In Section 2 we will prove Theorem 2 and use it to derive Theorem 1. In Section 3 we will present some additional remarks on Boolean analysis and learning, in particular showing that the dependence on n in Theorem 1 is optimal for $\delta \asymp \frac{1}{n}$. Moreover, we shall improve the recent bounds of [11] on the ℓ_1-Fourier growth of bounded functions of low degree.

Acknowledgements. We are very grateful to Assaf Naor for constructive feedback and to Lauritz Streck for useful discussions which led to Proposition 4.

2. Proofs

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

$$N_b \stackrel{\text{def}}{=} \left\lceil \frac{2}{b^2} \log\left(\frac{2}{\delta} \sum_{k=0}^{d} \binom{n}{k}\right)\right\rceil. \tag{7}$$

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

$$\alpha_S = \frac{1}{N_b} \sum_{j=1}^{N_b} f(X_j) w_S(X_j). \tag{8}$$

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

$$\forall\ S \subseteq \{1,\ldots,n\}, \qquad \mathbb{P}\Big\{|\alpha_S - \hat{f}(S)| > b\Big\} \le 2\exp(-N_b b^2/2). \tag{9}$$

Therefore, using the union bound and taking into account that f has degree at most d, we get

$$\mathbb{P}\underbrace{\Big\{|\alpha_S - \hat{f}(S)| \le b, \text{ for every } S \subseteq \{1,\ldots,n\} \text{ with } |S| \le d\Big\}}_{G_b} \ge 1 - 2\sum_{k=0}^{d}\binom{n}{k}\exp(-N_b b^2/2) \stackrel{(7)}{\ge} 1-\delta. \tag{10}$$

Fix an additional parameter $a \in (b, \infty)$ and consider the random collection of sets given by

$$\mathcal{S}_a \stackrel{\text{def}}{=} \Big\{S \subseteq \{1,\ldots,n\} :\ |\alpha_S| \ge a\Big\}. \tag{11}$$

Observe that if the event G_b of equation (10) holds, then

$$\forall\ S \notin \mathcal{S}_a, \qquad |\hat{f}(S)| \le |\alpha_S - \hat{f}(S)| + |\alpha_S| < a + b \tag{12}$$

and

$$\forall\ S \in \mathcal{S}_a, \qquad |\hat{f}(S)| \ge |\alpha_S| - |\alpha_S - \hat{f}(S)| \ge a - b. \tag{13}$$

Finally, consider the random function $h_{a,b}: \{-1,1\}^n \to \mathbb{R}$ given by

$$\forall\ x \in \{-1,1\}^n, \qquad h_{a,b}(x) \stackrel{\text{def}}{=} \sum_{S \in \mathcal{S}_a} \alpha_S w_S(x). \tag{14}$$

Combining (13) with inequality (5), we deduce that

$$|\mathcal{S}_a| \stackrel{(13)}{\le} (a-b)^{-\frac{2d}{d+1}} \sum_{S \in \mathcal{S}_a} |\hat{f}(S)|^{\frac{2d}{d+1}} \le (a-b)^{-\frac{2d}{d+1}} \sum_{S \subseteq \{1,\ldots,n\}} |\hat{f}(S)|^{\frac{2d}{d+1}} \stackrel{(5)}{\le} (a-b)^{-\frac{2d}{d+1}} (B_d^{\{\pm1\}})^{\frac{2d}{d+1}}. \tag{15}$$

Therefore, on the event G_b we have

$$\begin{aligned} \|h_{a,b} - f\|_{L_2}^2 &= \sum_{S \subseteq \{1,\ldots,n\}} \Big|\hat{h}_{a,b}(S) - \hat{f}(S)\Big|^2 = \sum_{S \in \mathcal{S}_a} |\alpha_S - \hat{f}(S)|^2 + \sum_{S \notin \mathcal{S}_a} |\hat{f}(S)|^2 \\ &\stackrel{(12)}{<} |\mathcal{S}_a| b^2 + (a+b)^{\frac{2}{d+1}} \sum_{S \notin \mathcal{S}_a} |\hat{f}(S)|^{\frac{2d}{d+1}} \stackrel{(5)\wedge(15)}{\le} (B_d^{\{\pm1\}})^{\frac{2d}{d+1}} \Big((a-b)^{-\frac{2d}{d+1}} b^2 + (a+b)^{\frac{2}{d+1}}\Big). \end{aligned} \tag{16}$$

Choosing $a = b(1+\sqrt{d+1})$, we deduce that

$$\|h_{b(1+\sqrt{d+1}),b} - f\|_{L_2}^2 < (B_d^{\{\pm 1\}})^{\frac{2d}{d+1}} b^{\frac{2}{d+1}}\big((d+1)^{-\frac{d}{d+1}} + (2+\sqrt{d+1})^{\frac{2}{d+1}}\big). \tag{17}$$

Next, we need the technical inequality

$$(d+1)^{-\frac{d}{d+1}} + (2+\sqrt{d+1})^{\frac{2}{d+1}} \le (e^4(d+1))^{\frac{1}{d+1}} \quad \text{for all} \quad d \ge 1. \tag{18}$$

Rearranging the terms, it suffices to show that $(2+\sqrt{d+1})^{\frac{2}{d+1}} \le (d+1)^{\frac{1}{d+1}}\big(e^{\frac{4}{d+1}} - \frac{1}{d+1}\big)$, which is equivalent to $\big(\frac{2}{\sqrt{d+1}}+1\big)^{\frac{2}{d+1}} \le e^{\frac{4}{d+1}} - \frac{1}{d+1}$. We have

$$\left(\frac{2}{\sqrt{d+1}}+1\right)^{\frac{2}{d+1}} \le \left(\sqrt{2}+1\right)^{\frac{2}{d+1}} \overset{(*)}{\le} 1 + \frac{3}{d+1} \le e^{\frac{4}{d+1}} - \frac{1}{d+1}, \tag{19}$$

where inequality $(*)$ holds because the left hand side is convex in the variable $\lambda \overset{\text{def}}{=} \frac{2}{d+1}$ whereas the right hand side is linear and since $(*)$ holds at the endpoints $\lambda = 0, 1$.

Combining (17) and (18) we see that $\|h_{b(1+\sqrt{d+1}),b} - f\|_{L_2}^2 < \varepsilon$ holds for $b^2 \le e^{-5} d^{-1} \varepsilon^{d+1} (B_d^{\{\pm 1\}})^{-2d}$. Plugging this choice of b in (7) shows that given N random queries, where

$$N = \left\lceil \frac{e^6 d (B_d^{\{\pm 1\}})^{2d}}{\varepsilon^{d+1}} \log\left(\frac{2}{\delta}\sum_{k=0}^{d}\binom{n}{k}\right)\right\rceil, \tag{20}$$

the random function $h_{b(1+\sqrt{d+1}),b}$ satisfies $\|h_{b(1+\sqrt{d+1}),b} - f\|_{L_2}^2 < \varepsilon$ with probability at least $1-\delta$ and the conclusion of the theorem follows from elementary estimates, such as

$$\sum_{k=0}^{d}\binom{n}{k} \le \sum_{k=0}^{d} \frac{n^k}{k!} = \sum_{k=0}^{d} \frac{d^k}{k!}\left(\frac{n}{d}\right)^k \le \left(\frac{en}{d}\right)^d. \qquad \square$$

Theorem 1 is a straightforward consequence of Theorem 2.

Proof of Theorem 1. Theorem 2 combined with the bound $B_d^{\{\pm 1\}} \le \exp(\kappa\sqrt{d\log d})$ of [8] imply the conclusion of Theorem 1 for $\varepsilon \ge \frac{\exp(C\sqrt{d\log d})}{n}$, where $C \in (0,\infty)$ is a large universal constant. The case $\varepsilon < \frac{\exp(C\sqrt{d\log d})}{n}$ follows from the Low-Degree Algorithm of [12]. $\square$

3. Concluding remarks

We conclude with a few additional remarks on the spectrum of bounded functions defined on the hypercube and corresponding learning algorithms. For a function $f : \{-1,1\}^n \to \mathbb{R}$, its Rademacher projection on level $\ell \in \{1,\ldots,n\}$ is defined as

$$\forall\, x \in \{-1,1\}^n, \qquad \mathrm{Rad}_\ell f(x) = \sum_{\substack{S \subseteq \{1,\ldots,n\} \\ |S| = \ell}} \hat{f}(S) w_S(x). \tag{21}$$

1. The first main theorem of [11] asserts that if $f : \{-1,1\}^n \to \mathbb{R}$ is a function of degree d, then

$$\forall\, \ell \in \{1,\ldots,d\}, \qquad \left\|\mathrm{Rad}_\ell f\right\|_{L_\infty} \le \begin{cases} \frac{|T_d^{(\ell)}(0)|}{\ell!}\cdot\|f\|_{L_\infty}, & \text{if } (d-\ell) \text{ is even} \\ \frac{|T_{d-1}^{(\ell)}(0)|}{\ell!}\cdot\|f\|_{L_\infty}, & \text{if } (d-\ell) \text{ is odd} \end{cases}, \tag{22}$$

where $T_d(t)$ is the d-th Chebyshev polynomial of the first kind, that is, the unique real polynomial of degree d such that $\cos(d\theta) = T_d(\cos\theta)$ for every $\theta \in \mathbb{R}$. Moreover, Iyer, Rao, Reis, Rothvoss and Yehudayoff observed in [11, Proposition 2] that this estimate is asymptotically sharp. We present a simple proof of their inequality (22) (see also [10] for related arguments).

Proof of (22). For any $f:\{-1,1\}^n\to\mathbb{R}$ consider its harmonic extension on $[-1,1]^n$,

$$\forall\ (x_1,\ldots,x_n)\in[-1,1]^n,\qquad \tilde{f}(x_1,\ldots,x_n)=\sum_{S\subseteq\{1,\ldots,n\}}\hat{f}(S)\prod_{j\in S}x_j. \tag{23}$$

By convexity $\|\tilde{f}\|_{L^\infty([-1,1]^n)}=\|f\|_{L^\infty(\{-1,1\}^n)}$. In particular, the restriction of $\tilde{f}$ on the ray $t(x_1,\ldots,x_n)$, $t\in[-1,1]$, i.e.

$$\forall\ t\in\mathbb{R},\qquad h_x(t)\stackrel{\mathrm{def}}{=}\sum_{S\subseteq\{1,\ldots,n\}}\hat{f}(S)w_S(x)t^{|S|} \tag{24}$$

satisfies $\max_{t\in[-1,1]}|h_x(t)|\le\|f\|_{L_\infty}$ for all $(x_1,\ldots,x_n)\in\{-1,1\}^n$. Therefore, since $\deg h_x\le d$, a classical inequality of Markov (see e.g. [5, p. 248]) gives

$$\left|\mathrm{Rad}_\ell f(x)\right|=\frac{|h_x^{(\ell)}(0)|}{\ell!}\le\begin{cases}\frac{|T_d^{(\ell)}(0)|}{\ell!}\cdot\|f\|_{L_\infty}, & \text{if } (d-\ell) \text{ is even}\\ \frac{|T_{d-1}^{(\ell)}(0)|}{\ell!}\cdot\|f\|_{L_\infty}, & \text{if } (d-\ell) \text{ is odd}\end{cases} \tag{25}$$

and (22) follows by taking a maximum over all $x\in\{-1,1\}^n$. □

In particular, as observed in [11], inequality (22) implies that if f has degree at most d then

$$\forall\ \ell\in\{1,\ldots,d\},\qquad \left\|\mathrm{Rad}_\ell f\right\|_{L_\infty}\le\frac{d^\ell}{\ell!}\cdot\|f\|_{L_\infty}. \tag{26}$$

2. The second main theorem of [11] asserts that if $f:\{-1,1\}^n\to[-1,1]$ is a bounded function of degree at most d, then for every $\ell\in\{1,\ldots,d\}$ we have

$$\sum_{S\subseteq\{1,\ldots,n\}}|\widehat{\mathrm{Rad}_\ell f}(S)|=\sum_{\substack{S\subseteq\{1,\ldots,n\}\\|S|=\ell}}|\hat{f}(S)|\le n^{\frac{\ell-1}{2}}d^\ell e^{\binom{\ell+1}{2}}. \tag{27}$$

The Bohnenblust–Hille-type inequality of [8] implies the following improved bound.

Corollary 3. *Let $n\in\mathbb{N}$ and $d\in\{1,\ldots,n\}$. Then, every bounded function $f:\{-1,1\}^n\to[-1,1]$ of degree at most d satisfies*

$$\forall\ \ell\in\{1,\ldots,d\},\qquad \sum_{\substack{S\subseteq\{1,\ldots,n\}\\|S|=\ell}}|\hat{f}(S)|\le\binom{n}{\ell}^{\frac{\ell-1}{2\ell}}e^{\kappa\sqrt{\ell\log\ell}}\frac{d^\ell}{\ell!}\le n^{\frac{\ell-1}{2}}d^\ell\ell^{-c\ell}, \tag{28}$$

for some universal constant $c\in(0,1)$.

Proof. Combining Hölder's inequality with the estimate of [8] and (26) we get

$$\begin{aligned}\sum_{\substack{S\subseteq\{1,\ldots,n\}\\|S|=\ell}}|\hat{f}(S)|&\le\binom{n}{\ell}^{\frac{\ell-1}{2\ell}}\Bigg(\sum_{S\subseteq\{1,\ldots,n\}}|\widehat{\mathrm{Rad}_\ell f}(S)|^{\frac{2\ell}{\ell+1}}\Bigg)^{\frac{\ell+1}{2\ell}}\\&\le\binom{n}{\ell}^{\frac{\ell-1}{2\ell}}\exp(\kappa\sqrt{\ell\log\ell})\left\|\mathrm{Rad}_\ell f\right\|_{L_\infty}\stackrel{(26)}{\le}\binom{n}{\ell}^{\frac{\ell-1}{2\ell}}\exp(\kappa\sqrt{\ell\log\ell})\frac{d^\ell}{\ell!}.\end{aligned} \tag{29}$$

The last inequality of (28) follows from (22) and the elementary bound $\binom{n}{\ell}\le\left(\frac{ne}{\ell}\right)^\ell$. □

We refer to the recent work [2] for a systematic study of inequalities relating the Fourier growth with various well-studied properties of Boolean functions.

3. It is straightforward to observe (see also [15, Proposition 3.31]) that if $f:\{-1,1\}^n\to\{-1,1\}$ is a Boolean function and $h:\{-1,1\}^n\to\mathbb{R}$ is an arbitrary function, then

$$\left\|\mathrm{sign}(h)-f\right\|_{L_2}^2=4\mathbb{P}\{\mathrm{sign}(h)\ne f\}\le4\mathbb{P}\{|h-f|\ge1\}\le4\|h-f\|_{L_2}^2, \tag{30}$$

where we define sign(0) as ±1 arbitrarily. Therefore, applying Theorem 1 to a Boolean function, the above algorithm produces a *Boolean* function $\tilde{h} = \mathrm{sign}(h)$ which is a 4ε-approximation of f.

4. In Theorem 1 we showed that bounded functions $f : \{-1,1\}^n \to [-1,1]$ of degree at most d can be learned with accuracy at most ε and confidence at least $1-\delta$ from $N = O_{\varepsilon,d}\Big(\log(n/\delta)\Big)$ random queries. We will now show that this estimate is sharp for small enough values of δ.

Proposition 4. *Suppose that bounded linear functions $\ell : \{-1,1\}^n \to [-1,1]$ can be learned with accuracy at most $\frac{1}{2}$ and confidence at least $1-\frac{1}{2n}$ from N random queries. Then $N > \log_2 n$.*

Proof. By the assumption, for any input $(X_1,y_1),\ldots,(X_N,y_N) \in \{-1,1\}^n \times [-1,1]$, there exists a function $h_{(X_1,y_1),\ldots,(X_N,y_N)} : \{-1,1\}^n \to \mathbb{R}$ such that if $X_1,\ldots,X_N$ are chosen independently and uniformly from $\{-1,1\}^n$ and there exists a linear function $\ell : \{-1,1\}^n \to [-1,1]$ such that $y_j = \ell(X_j)$ for every $j \in \{1,\ldots,N\}$, then $\mathbb{P}(\Omega_\ell) > 1-\frac{1}{2n}$, where Ω_ℓ is the event

$$\Omega_\ell \stackrel{\mathrm{def}}{=} \Big\{ \mathbb{E}\Big(h_{(X_1,\ell(X_1)),\ldots,(X_N,\ell(X_N))} - \ell\Big)^2 < \frac{1}{2} \Big\}. \tag{31}$$

Let $X_j = (X_j(1),\ldots,X_j(n))$ for $j \in \{1,\ldots,N\}$ and consider the event

$$\mathcal{W} = \Big\{ X_j(1) = X_j(2), \ \forall\, j \in \{1,\ldots,N\} \Big\}. \tag{32}$$

By the independence of the samples, we have $\mathbb{P}(\mathcal{W}) = \frac{1}{2^N}$. Therefore, if $N \le \log_2 n$ and we consider the linear functions $r_i : \{-1,1\}^n \to \{-1,1\}$ given by $r_i(x) = x_i$, then

$$\mathbb{P}(\Omega_{r_1} \cap \Omega_{r_2}) > 1 - \frac{1}{n} \ge 1 - \frac{1}{2^N} = 1 - \mathbb{P}(\mathcal{W}), \tag{33}$$

which implies that $\Omega_{r_1} \cap \Omega_{r_2} \cap \mathcal{W} \neq \emptyset$. Choosing $X_1,\ldots,X_N$ from this event and denoting by $h = h_{(X_1,X_1(1)),\ldots,(X_N,X_N(1))} = h_{(X_1,X_1(2)),\ldots,(X_N,X_N(2))}$, we deduce from the triangle inequality that

$$2 = \mathbb{E}(r_1 - r_2)^2 \le 2\mathbb{E}(h - r_1)^2 + 2\mathbb{E}(h - r_2)^2 \stackrel{(31)}{<} 2 \tag{34}$$

which is clearly a contradiction. Therefore $N > \log_2 n$. □

References

[1] F. Bayart, D. Pellegrino, and J. B. Seoane-Sepúlveda, *The Bohr radius of the n-dimensional polydisk is equivalent to $\sqrt{(\log n)/n}$*, Adv. Math., 264 (2014), pp. 726–746.

[2] J. Błasiok, P. Ivanov, Y. Jin, C. H. Lee, R. Servedio, and E. Viola, *Fourier growth of structured $\mathbb{F}_2$-polynomials and applications*. To appear in RANDOM 2021. Preprint available at https://arxiv.org/abs/2107.10797, 2021.

[3] R. Blei, *Analysis in integer and fractional dimensions*, vol. 71 of Cambridge Studies in Advanced Mathematics, Cambridge University Press, Cambridge, 2001.

[4] H. F. Bohnenblust and E. Hille, *On the absolute convergence of Dirichlet series*, Ann. of Math. (2), 32 (1931), pp. 600–622.

[5] P. Borwein and T. Erdélyi, *Polynomials and polynomial inequalities*, vol. 161 of Graduate Texts in Mathematics, Springer-Verlag, New York, 1995.

[6] J. R. Campos, P. Jiménez-Rodríguez, G. A. Muñoz Fernández, D. Pellegrino, and J. B. Seoane-Sepúlveda, *On the real polynomial Bohnenblust-Hille inequality*, Linear Algebra Appl., 465 (2015), pp. 391–400.

[7] A. Defant, L. Frerick, J. Ortega-Cerdà, M. Ounaïes, and K. Seip, *The Bohnenblust-Hille inequality for homogeneous polynomials is hypercontractive*, Ann. of Math. (2), 174 (2011), pp. 485–497.

[8] A. Defant, M. Mastyło, and A. Pérez, *On the Fourier spectrum of functions on Boolean cubes*, Math. Ann., 374 (2019), pp. 653–680.

[9] A. Defant and P. Sevilla-Peris, *The Bohnenblust-Hille cycle of ideas from a modern point of view*, Funct. Approx. Comment. Math., 50 (2014), pp. 55–127.

[10] A. Eskenazis and P. Ivanisvili, *Polynomial inequalities on the Hamming cube*, Probab. Theory Related Fields, 178 (2020), pp. 235–287.

[11] S. Iyer, A. Rao, V. Reis, T. Rothvoss, and A. 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.

[12] N. Linial, Y. Mansour, and N. Nisan, *Constant depth circuits, Fourier transform, and learnability*, J. Assoc. Comput. Mach., 40 (1993), pp. 607–620.

[13] J. E. Littlewood, *On bounded bilinear forms in an infinite number of variables*, Q. J. Math., os-1 (1930), pp. 164–174.
[14] Y. Mansour, *Learning Boolean Functions via the Fourier Transform*, Springer US, Boston, MA, 1994, pp. 391–424.
[15] R. O'Donnell, *Analysis of Boolean functions*, Cambridge University Press, New York, 2014.

(A. E.) Trinity College and Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, UK.
Email address: ae466@cam.ac.uk

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