On Agnostic PAC Learning using $\mathcal{L}_2$-polynomial Regression and Fourier-based Algorithms

Mohsen Heidari and Wojciech Szpankowski,
Department of Computer Science, Purdue University,
{mheidari, szpan}@purdue.edu

Abstract

We develop a framework using Hilbert spaces as a proxy to analyze PAC learning problems with structural properties. We consider a joint Hilbert space incorporating the relation between the true label and the predictor under a joint distribution D. We demonstrate that agnostic PAC learning with 0-1 loss is equivalent to an optimization in the Hilbert space domain. With our model, we revisit the PAC learning problem using methods based on *least-squares* such as $\mathcal{L}_2$ polynomial regression and Linial's low-degree algorithm. We study learning with respect to several hypothesis classes such as half-spaces and polynomial-approximated classes (i.e., functions approximated by a fixed-degree polynomial). We prove that (under some distributional assumptions) such methods obtain generalization error up to $2\mathsf{P}_{opt}$ with P_{opt} being the optimal error of the class. Hence, we show the tightest bound on generalization error when $\mathsf{P}_{opt} \leqslant 0.2$.

I. Introduction

We study binary classification using polynomial regression from the agnostic PAC learning perspective [1], [2]. In this problem, multiple training instances are generated IID according to an underlying distribution D on the feature-label sets $\mathcal{X} \times \{-1, 1\}$. In addition, we are given a hypothesis class with respect to which the learning process takes place. If P_{opt} is the minimum error attained using the given class, then the objective of the learning algorithm is to output, with high probability, a classifier whose generalization error is not greater than $\mathsf{P}_{opt} + \epsilon$.

To gain computational efficiency or analytical tractability, many conventional learning methods such as support-vector machine (SVM) rely on intermediate loss functions other than the natural $0-1$ loss. Square loss is an example that is a basis for $\mathcal{L}_2$-polynomial regression or another variant of SVM known

as LS-SVM [3]. The well-known "low-degree" algorithm [4] is also known to be in this category of algorithms [5]. Such methods have been analyzed for many PAC learning problems. Under the *realizability* assumption where $\mathsf{P}_{opt} = 0$, the $\mathcal{L}_2$-polynomial regression and the low-degree algorithm are PAC learners for a variety of hypothesis classes [6]–[8]. Under the agnostic setting where $\mathsf{P}_{opt} > 0$, the current results are not that promising. The best known results for $\mathcal{L}_2$-polynomial regression (and the low-degree algorithm under the uniform distribution) are $8\mathsf{P}_{opt}$ and $\frac{1}{4} + \mathsf{P}_{opt}(1 - \mathsf{P}_{opt})$ for classes such as half-spaces or polynomial-approximated classes [2], [5].

In this paper, we develop a framework using Hilbert spaces as a proxy to analyze such problems. We consider a joint Hilbert space incorporating the relation between the true label and the predictor under the joint distribution D. This is unlike conventional analysis using Hilbert spaces that focus only on the predictors with marginal $D_{\mathbf{x}}$ on the features. As a byproduct, we improve the above mentioned bounds and show that the generalization error of $\mathcal{L}_2$-polynomial regression and the low-degree algorithm is less than $2\mathsf{P}_{opt}$. This bound the improves upon the previous bounds when $\mathsf{P}_{opt} \leqslant 0.2$. We show that methods based on square loss are suitable for learning classes with appropriate geometrical properties.

A. *Our approach*

We develop our framework by constructing two Hilbert spaces one with respect to the true underlying distribution D and the other with respect to the empirical one. The first one is $\mathcal{L}_2(D)$, that is all real-valued functions f on $\mathcal{X} \times \mathcal{Y}$ such that $\mathbb{E}[f(\mathbf{X}, Y)^2] < \infty$. The second one is $\mathcal{L}_2(\hat{D})$ with $\hat{D}$ being the empirical distribution of the training set. With this formulation, the true label Y and the training labels are understood as a member of these spaces. With this formulation, the generalization error of any classifier c equals $\frac{1}{4}\|Y - c\|^2_{2,D}$. Similarly, when the distance is calculated in the second Hilbert space, we obtain a characterization of the empirical error. Hence, minimizing the generalization (or empirical) error is equivalent to minimizing the distance between Y and the classifier c in the first (or second) Hilbert space. We argue that the mentioned hypothesis classes have appropriate structures using that allows us to drive lower bounds on its minimum error P_{opt}. For instance, given k, the polynomial-approximated class is characterized by the subspace of $\mathcal{L}_2(D)$ spanned by polynomials of degree up to k. With this structure, finding P_{opt} is equivalent to finding the minimum distance of Y to the subspace spanned by polynomials of degree up to k. As for the learning algorithms, we argue the low-degree algorithm and $\mathcal{L}_2$-polynomial regression have suitable structures using which we drive our upper bounds on their generalization errors. For instance, in the case of $\mathcal{L}_2$ polynomial regression, the error of any classifier of the form $\mathsf{sign}[p(x) - \theta]$, with θ chosen appropriately, is bounded from above by $\frac{1}{2}\|Y - p\|^2_2$. Hence, minimizing the squares-loss as in $\mathcal{L}_2$-regression yields an error less than $2\mathsf{P}_{opt}$.

B. Summary of the Results

In this work, we first present a more general version of the low-degree algorithm incorporating non-uniform but product probability distributions. We refer to this generalization as Fourier algorithm. With our framework, we study learning with respect to three well-known hypothesis classes. The first class is half-spaces consisting of all the Boolean-valued functions of the form $c(\mathbf{x}) = \mathsf{sign}[\sum_{j=1}^{d} w_j x_j - \theta]$. The second class is called polynomial-approximated functions. Given a positive integer k and $\epsilon > 0$, it consists of Boolean-valued functions that are approximated by a degree k polynomial with square error up to ϵ^2. The thirst class is a generalization of the second. We use our framework to analyze learning these hypothesis classes using $\mathcal{L}_2$-polynomial regression and the Fourier algorithm. Below, is the summary of our results:

1) The $\mathcal{L}_2$ polynomial regression with degree k outputs a hypothesis $\hat{g}$ whose generalization error has the following properties:

- For learning polynomial-approximated classes, it is less than $2\mathsf{P}_{opt} + 3\epsilon$ (Theorem 1).
- For learning half-spaces, when the marginal $D_{\mathbf{x}}$ is uniform over the unit ball in $\mathbb{R}^d$, it is less than $2\mathsf{P}_{opt} + 3\epsilon$ (Theorem 3).
- For learning *generalized concentrated classes* , under any distribution, it is less than $2\mathsf{P}_{opt} + \epsilon$ (Theorem 4).

2) If the marginal $D_{\mathbf{x}}$ is a product probability distribution on $\{-1, 1\}^d$, then with probability $(1 - \delta)$, the Fourier algorithm outputs a hypothesis such that its generalization error is less than $2\mathsf{P}_{opt} + 2\epsilon$ for learning polynomial-approximated classes.

C. Related Works

The low-degree algorithm is introduced by [4] with PAC learning guarantees under the uniform distribution over $\{-1, 1\}^d$. This algorithm which is based on the Fourier expansion on the Boolean cube has been used for in various problems [6], [8], [9]. The $\mathcal{L}_2$ polynomial regression along with its $\mathcal{L}_1$ counterpart is introduced by [5] for learning with respect to polynomial-approximated classes, k-juntas, and half-spaces. Learning with respect to such classes has been studied extensively in the literature [5], [10]–[12]. Among such classes, learning with respect to half-spaces is the most challenging. In the case of *proper* agnostic PAC learning, where the algorithm's predictor must be a half-space, it is an NP-hard problem [13], [14]. Even without the *proper* restriction, the problem is NP-hard. That said, under distributional assumptions, polynomial time algorithms are introduced [5], [15], [16]. Among them are the *improper* learning algorithms based on regression methods such as $\mathcal{L}_1$ or $\mathcal{L}_2$ polynomial regression

[4], [5]. In particular, [5] proved that $\mathcal{L}_1$ polynomial regression learns a range of hypothesis classes such as half-spaces (under distributional assumptions) and polynomial-approximated classes

II. Preliminaries

Notation: The input set is denoted by $\mathcal{X}$ which is a subset of $\mathbb{R}^d$ for some positive integer d. The output set is denoted by $\mathcal{Y}$ which is a subset of $\mathbb{R}$. In binary classification $\mathcal{Y} = \{-1, 1\}$. For shorthand, the random vectors in $\mathbb{R}^d$ are denoted by $\mathbf{X} = (X_1, X_2, ..., X_d)$. Further, for any ordered subset $\mathcal{J} = \{j_1, j_2, \cdots, j_m\}$, by $X^{\mathcal{J}}$ denote the random vector $(X_{j_1}, X_{j_2}, \cdots, X_{j_m})$. Similarly, by $x^{\mathcal{J}}$ denote the vector $(x_{j_1}, x_{j_2}, \cdots, x_{j_m})$. For a pair of functions f, g on $\mathcal{X}$, the notation $f \equiv g$ means that $f(x) = g(x)$ for all $x \in \mathcal{X}$. Lastly, for any natural number ℓ, the set $\{1, 2, \cdots, \ell\}$ is denoted by $[\ell]$.

A. A Hilbert Space Representation

We first develop a Hilbert Space formulation for the binary classification problem. Let D be a joint probability distribution on the input-output set $\mathcal{X} \times \mathcal{Y}$. In this paper, it is assumed that the marginal $D_{\mathbf{x}}$ of any joint distribution D on $\mathcal{X} \times \mathcal{Y}$ has finite moments. Consider a Hilbert space of all real-valued functions $f : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ which are $\mathcal{L}_2(D)$, that is $\mathbb{E}_D[f(\mathbf{X}, Y)^2] < \infty$. The inner product between two members f, g is defined as

$$\langle f, g \rangle \triangleq \mathbb{E}_D[f(\mathbf{X}, Y) g(\mathbf{X}, Y)].$$

Given any integer $p > 0$ and distribution D, the p-norm of a function f is defined as

$$\|f\|_{p,D} \triangleq \left(\mathbb{E}_D[f(\mathbf{X}, Y)^p]\right)^{1/p}.$$

Given any training sample $\mathcal{S} = \{(\mathbf{x}_i, y_i) : i = 1, 2, ..., n\}$, let $\hat{D}$ denote its empirical distribution, that is a uniform distribution on $\mathcal{S}$ and zero outside of it. Associated with this distribution, we consider the Hilbert space $\mathcal{L}_2(\hat{D})$ with the inner product and norms defined based on the empirical distribution $\hat{D}$. We use this formulation to study the binary classification problem where $\mathcal{Y} = \{-1, 1\}$. Therefore, the generalization error of any predictor $c : \mathcal{X} \mapsto \{-1, 1\}$ can be written in terms of the inner products as

$$\mathbb{P}_D\Big\{Y \neq c(\mathbf{X})\Big\} = \frac{1}{2} - \frac{1}{2}\langle Y, c \rangle_D = \frac{1}{4}\|Y - c\|_{2,D}^2, \tag{1}$$

where, with slight abuse of notation, Y is understood as the mapping $(x, y) \mapsto y$ and c is understood as a mapping on $\mathcal{X} \times \mathcal{Y}$ which depends only on $\mathcal{X}$. Similarly, the empirical error of c is equal to

$$\hat{\mathbb{P}}_{\hat{D}}\Big\{Y \neq c(\mathbf{X})\Big\} = \frac{1}{2} - \frac{1}{2}\langle Y, c \rangle_{\hat{D}} = \frac{1}{4}\|Y - c\|_{2,\hat{D}}^2.$$

The goal now is to derive bounds on the minimum generalization error when learning with respect to various hypothesis classes. In Section III we describe $\mathcal{L}_2$-polynomial regression and the Fourier algorithm,

in Section IV we study polynomial-approximated classes, and finally in Section V we discuss half-spaces, and more general hypothesis classes that have structural properties.

III. PAC Learning with $\mathcal{L}_2$-Polynomial Regression

We employ a PAC learning algorithm using $\mathcal{L}_2$-polynomial regression. Given a training set, the objective of the polynomial regression is to minimize the empirical square loss over all polynomials of degree up to k. This process can be implemented by stochastic gradient descent or by solving a linear system of equations. We describe how this polynomial regression can be used for PAC learning. Let $\hat{p}$ be the output of the polynomial regression. The idea is to shift the polynomial $\hat{p}$ by a threshold θ and take its sign. This process is demonstrated as Algorithm 1.

Algorithm 1 PAC Learning with $\mathcal{L}_2$-Polynomial Regression

Input: Degree parameter k, and training samples $\{(\mathbf{x}(i), y(i)), i \in [n]\}$.

1: Find a polynomial $\hat{p}$ of degree up to k that minimizes

$$\frac{1}{n}\sum_{i}\big(y(i) - p(\mathbf{x}(i))\big)^2.$$

2: Find $\theta \in [-1, 1]$ such that the empirical error of $\mathsf{sign}[\hat{p}(\mathbf{x}) - \theta]$ is minimized.

3: **return** $\hat{g} \equiv \mathsf{sign}[\hat{p} - \theta]$.

A. *Fourier-Based Learning Algorithm*

We present another variant of $\mathcal{L}_2$ polynomial regression, known as the low-degree (Fourier) algorithm [4]. Although this algorithm is more efficient than the polynomial regression, it requires binary input set $\mathcal{X} = \{-1, 1\}^d$. The low-degree algorithm was originally designed for uniform distribution on the Boolean cube. In this paper, we present a more general version of it for incorporating non-uniform but product probability distributions on $\{-1, 1\}^d$ [17]. In this approach, the objective is to find an estimate of the p^* polynomial that minimizes the square loss $\|Y - p^*\|_{2,D}$ under the true distribution. This method is based on the Fourier expansion on the Boolean cube [18] and is summarized in the following.

Under product probability distribution on $\{-1, 1\}^d$, any bounded real-valued functions can be written as

$$f(\mathbf{x}) = \sum_{\mathcal{S}\subseteq[d]} f_{\mathcal{S}}\ \psi_{\mathcal{S}}(\mathbf{x}),$$

where $f_{\mathcal{S}}$'s are the Fourier coefficients and calculated as $f_{\mathcal{S}} \triangleq \langle f, \psi_{\mathcal{S}} \rangle$ for every subset $\mathcal{S} \subseteq [d]$. Further, the parity $\psi_{\mathcal{S}}$ is a monomial defined as

$$\psi_{\mathcal{S}}(x) = \prod_{j \in \mathcal{S}} \frac{x_j - \mu_j}{\sigma_j},$$

with μ_j and σ_j being the mean and standard deviation of the X_j, respectively. As the distribution is unknown, these quantities are estimated in the algorithm.

As a result, we can write the Fourier decomposition of the optimal polynomial p^*. For that, we have the following statement:

Fact 1. *Let D be a probability distribution with the marginal $D_{\mathbf{x}}$ that is a product probability distribution on $\{-1, 1\}^d$. Then, the optimal polynomial p^* admits the following Fourier decomposition*

$$p^* \equiv \sum_{\mathcal{S} \subseteq [d] : |\mathcal{S}| \leqslant k} \langle Y, \psi_{\mathcal{S}} \rangle \, \psi_{\mathcal{S}}.$$

With that decomposition, the idea behind the Fourier algorithm is to compute an empirical estimate of $\langle Y, \psi_{\mathcal{S}} \rangle$. This is demonstrated as Algorithm 2.

Algorithm 2 Fourier-Based Learning

Input: Training samples $\{(\mathbf{x}(i), y(i)), i \in [n]\}$.

1: Compute the empirical mean $\hat{\mu}_j$ and standard deviation $\hat{\sigma}_j$ of each feature.

2: For every $\mathcal{S} \subseteq [d]$ with $|\mathcal{S}| \leqslant k$, construct the empirical parity as $\widehat{\psi}_{\mathcal{S}}(\mathbf{x}) = \prod_{j \in \mathcal{S}} \frac{x_j - \hat{\mu}_j}{\hat{\sigma}_j}$.

3: Compute the empirical Fourier coefficients $a_{\mathcal{S}}$, for every $\mathcal{S}$ with at most k elements, as

$$a_{\mathcal{S}} = \frac{1}{n} \sum_{i=1}^{n} y(i) \widehat{\psi}_{\mathcal{S}}(\mathbf{x}(i)).$$

4: Construct and return the function $\hat{\Pi}_Y$ as

$$\hat{\Pi}_Y(\mathbf{x}) \triangleq \sum_{\mathcal{S} : |\mathcal{S}| \leqslant k} a_{\mathcal{S}} \widehat{\psi}_{\mathcal{S}}(\mathbf{x}).$$

In the following lemma which is proved in Appendix A, we derive bounds for estimating the optimal polynomial p^*.

Lemma 1. *Let D be a probability distribution with the marginal $D_{\mathbf{x}}$ that is a product probability distribution on $\{-1, 1\}^d$. Given $\delta \in (0, 1)$, with probability at least $(1 - \delta)$, the following inequality holds*

$$\|p^* - \hat{\Pi}_Y\|_2 \leqslant O\Big(\sqrt{\frac{d^k c_k}{(k-1)! n} \log \frac{4 d^k}{(k-1)! \delta}}\Big), \tag{2}$$

where $c_k \triangleq \max_{\mathcal{S} \subseteq [d], |\mathcal{S}| \leqslant k} \|\psi_{\mathcal{S}}\|_\infty^2$ and n is the number of samples.

IV. Polynomially Approximated Class

In this section, we study agnostic PAC learning with respect to concept classes whose members are approximated by fixed-degree polynomials. We adopt the Hilbert space representation in Section II-A to analyze PAC learning using Algorithm 1 and 2. We start with the following formulation:

Definition 1. Given $\epsilon \in [0,1]$, $k \in \mathbb{N}$ and any probability distribution $D_{\mathbf{X}}$ on $\mathcal{X}$, a concept class $\mathcal{C}$ of functions $c : \mathcal{X} \mapsto \{-1,1\}$ is (ϵ, k)-approximated if

$$\sup_{c\in\mathcal{C}} \inf_{p\in\mathcal{P}_k} \mathbb{E}\big[\big(c(\mathbf{X}) - p(\mathbf{X})\big)^2\big] \leqslant \epsilon^2,$$

where $\mathcal{P}_k$ is the set of all polynomials of degree up to k.

We consider agnostic PAC learning with respect to $\mathcal{C}$ and under the $0-1$ loss function. The minimum generalization error and empirical error of $\mathcal{C}$ are, respectively, defined as

$$\mathsf{P}_{opt} \triangleq \min_{c\in\mathcal{C}} \mathbb{P}_D\big\{Y \neq c(\mathbf{X})\big\},$$
$$\widehat{\mathsf{P}}_{opt} \triangleq \min_{c\in\mathcal{C}} \hat{\mathbb{P}}_{\hat{D}}\big\{Y \neq c(\mathbf{X})\big\}.$$

We use the Hilbert space representation in Section II-A and provide a lower bound on P_{opt}.

Lemma 2. *The minimum generalization error attainable by any (ϵ, k) concept class $\mathcal{C}$ is bounded from below as*

$$\mathsf{P}_{opt} \geqslant \frac{1}{2} - \frac{1}{2}\|p^*\|_{1,D} - \epsilon,$$

where $p^* = \arg\min_{p\in\mathcal{P}_k} \mathbb{E}_D\big[\big(Y - p(\mathbf{X})\big)^2\big]$.

Proof. From (1) the $0-1$ loss of any function $c \in \mathcal{C}$ can be written as $\mathbb{P}\Big\{Y \neq c(\mathbf{X})\Big\} = \frac{1}{2} - \frac{1}{2}\langle Y, c\rangle$. Let $p \in \mathcal{P}_k$ be such that $\|c - p\|_{2,D} \leqslant \epsilon$. Then, by adding and subtracting p, we obtain that

$$\begin{aligned}\langle Y, c\rangle &= \langle Y, p\rangle + \langle Y, (c-p)\rangle \\ &\leqslant \langle Y, p\rangle + \|Y\|_2\|c-p\|_2 \leqslant \langle Y, p\rangle + \epsilon, \end{aligned} \tag{3}$$

where the first inequality follows from Cauchy–Schwarz inequality and the second inequality follows because $\|Y\|_2 = 1$. Note that $\mathcal{P}_k$, the set of all polynomials on $\mathcal{X}$ with degree upto k, is a (finite dimensional) subspace inside the Hilbert space $\mathcal{L}_2(D)$. Therefore, it has an orthonormal basis denoted by $\{\Psi_1, \Psi_2, ..., \Psi_m\}$, where m is less than $O(d^k)$. As a result, the polynomial p can be written as $p \equiv \sum_{j=1}^m \langle p, \Psi_j\rangle \Psi_j$. Hence,

$$\langle Y, p\rangle = \sum_{j=1}^m \langle p, \Psi_j\rangle \, \langle Y, \Psi_j\rangle = \langle \Pi_Y, p\rangle,$$

where $\Pi_Y \equiv \sum_{j=1}^{m} \langle Y, \Psi_j \rangle \Psi_j$ is the *projection* of Y onto this subspace. Consequently, from the above equality and (3), we obtain that

$$\begin{aligned}
\langle Y, c \rangle &\leqslant \langle \Pi_Y, p \rangle + \epsilon = \langle \Pi_Y, c \rangle + \langle \Pi_Y, (p - c) \rangle + \epsilon \\
&\leqslant \langle \Pi_Y, c \rangle + \|\Pi_Y\|_2 \|p - c\|_2 + \epsilon \\
&\leqslant \|\Pi_Y\|_1 + \|\Pi_Y\|_2 \|p - c\|_2 + \epsilon \\
&\leqslant \|\Pi_Y\|_1 + 2\epsilon,
\end{aligned}$$

where the second inequality follows from Cauchy–Schwarz inequality, the third one holds as $|c(\mathbf{x})| = 1$ and the last inequality follows from Bessel's inequality, implying $\|\Pi_Y\|_2 \leqslant 1$, and the assumption that $\|p - c\|_2 \leqslant \epsilon$. Next, we proceed with the following fact about the projection.

Fact 2. *Π_Y the projection of Y onto $\mathcal{P}_k$ is the polynomial minimizing $\mathbb{E}\big[\big(Y - p(\mathbf{X})\big)^2\big]$ over all $p \in \mathcal{P}_k$.*

The proof is complete by the following fact implying that $\Pi_Y \equiv p^*$. □

We show in Section III-A that the lower-bound in Lemma 2 helps to prove our results for the low-degree algorithm.

A. PAC Learning Bounds

Next, we analyze Algorithm 1 and 2 for this class and prove the first main result of the paper.

Theorem 1. *Given $\epsilon > 0$ and $k \in \mathbb{N}$, the degree k $\mathcal{L}_2$ polynomial regression agnostically PAC learns any (ϵ, k)-approximated concept class with expected error up to*

$$2\mathsf{P}_{opt} + 3\epsilon + \sqrt{\frac{2\ d^{k+1}}{n} \log \frac{en}{d^{k+1}}},$$

where d is the number of input variables and n is the sample size.

Proof. To derive an upper bound on the empirical error of $\hat{g}$, we first consider a weaker version of the algorithm. The idea is to select θ randomly instead of optimizing it as in the algorithm. For that, we establish the following lemma.

Lemma 3. *Suppose θ is a random variable with the probability density function $f_\theta(t) = 1 - |t|$, for $t \in [-1, 1]$. Then, the following bound holds for any polynomial p*

$$\mathbb{E}_\theta\Big[\hat{\mathbb{P}}\Big\{Y \neq \mathsf{sign}[p(\mathbf{X}) - \theta]\Big\}\Big] \leqslant \frac{1}{2}\|Y - p\|_{2,\hat{D}}^2.$$

Proof. Note that $y \neq \mathsf{sign}(p(\mathbf{x}) - \theta)$, if θ is between y and $p(\mathbf{x})$. Hence, the expected empirical error of $\mathsf{sign}[p(\mathbf{X}) - \theta]$ with respect to the random θ equals to

$$\begin{aligned}
&\mathbb{E}_\theta\Big[\hat{\mathbb{P}}\Big\{Y \neq \mathsf{sign}[p(\mathbf{X}) - \theta]\Big\}\Big] \\
&\quad = \frac{1}{n}\sum_i \mathbb{E}_\theta\Big[\mathbb{1}\big\{y_i \neq \mathsf{sign}(p(\mathbf{x}_i) - \theta)\big\}\Big] \\
&\quad = \frac{1}{n}\sum_i \underbrace{\mathbb{P}\Big\{\theta \in [p(\mathbf{x}_i), y_i] \bigcup [y_i, p(\mathbf{x}_i)]\Big\}}_{\mathbb{P}_i}. \qquad (4)
\end{aligned}$$

Next, we show that $\mathbb{P}_i \leqslant \frac{1}{2}(y_i - p(\mathbf{x}_i))^2$ for all $(\mathbf{x}_i, y_i)$'s. Suppose $y_i = 1$. If $p(\mathbf{x}_i) > 1$, then $\mathbb{P}_i = 0$ as $\theta \leqslant 1$. If $p(\mathbf{x}_i) \in [0, 1]$, then

$$\begin{aligned}
\mathbb{P}_i &= \mathbb{P}\Big\{\theta \in [p(\mathbf{x}_i), 1]\Big\} = \int_{p(\mathbf{x}_i)}^{1} (1 - t)dt \\
&= \frac{1}{2}\big(1 - p(\mathbf{x}_i)\big)^2 = \frac{1}{2}\big(y_i - p(\mathbf{x}_i)\big)^2.
\end{aligned}$$

If $p(\mathbf{x}_i) \in [-1, 0]$, then

$$\begin{aligned}
\mathbb{P}_i &= \mathbb{P}\Big\{\theta \in [p(\mathbf{x}_i), 1]\Big\} = \int_{p(\mathbf{x}_i)}^{1} 1 - |t|dt \\
&= \frac{1}{2} + \int_{p(\mathbf{x}_i)}^{0} (1 + t)dt \\
&= \frac{1}{2} - p(\mathbf{x}_i) - \frac{1}{2}(p(\mathbf{x}_i))^2 \\
&\leqslant \frac{1}{2}(1 + |p(\mathbf{x}_i)|)^2 = \frac{1}{2}(y_i - p(\mathbf{x}_i))^2.
\end{aligned}$$

Lastly, if $p(\mathbf{x}_i) < -1$, then $\mathbb{P}_i = 1$ because $\theta \geqslant -1$. In this case also $\mathbb{P}_i \leqslant \frac{1}{2}(y_i - p(\mathbf{x}_i))^2$. The case for $y_i = -1$ follows by symmetricity. Hence, we obtain the following inequality

$$\mathbb{E}_\theta\Big[\hat{\mathbb{P}}\Big\{Y \neq \hat{g}(\mathbf{X})\Big\}\Big] \leqslant \frac{1}{n}\sum_i \frac{1}{2}\big(y_i - p(\mathbf{x}_i)\big)^2.$$

The proof is complete by noting that the right-hand side equals to $\frac{1}{2}\|Y - p\|^2_{2,\hat{D}}$. □

Consequently, from the lemma and due the fact that θ in the algorithm is selected to minimize the empirical error, we obtain that

$$\hat{\mathbb{P}}\Big\{Y \neq \hat{g}(\mathbf{X})\Big\} \leqslant \frac{1}{2}\|Y - \hat{p}\|^2_{2,\hat{D}}, \qquad (5)$$

where $\hat{p}$ is the output of $\mathcal{L}_2$-polynomial regression and $\hat{g} \equiv \mathsf{sign}[\hat{p} - \theta]$, as in Algorithm 1. Let c^* be the predictor with minimum generalization error in the (ϵ, k)-approximated concept class. Let p be a degree

k polynomial such that $\|c^* - p\|_2 \leqslant \epsilon$. Since $\hat{p}$ minimizes the empirical 2-norm, then the right-hand side of (5) satisfies

$$\frac{1}{2}\|Y - \hat{p}\|_{2,\hat{D}}^2 \leqslant \frac{1}{2}\|Y - p^*\|_{2,\hat{D}}^2. \tag{6}$$

We proceed by taking the expected error of the empirical error with respect to the random training samples. From (5) and (6) we obtain the following inequalities

$$\begin{aligned}
\mathbb{E}\Big[\hat{\mathbb{P}}\Big\{Y \neq \hat{g}(\mathbf{X})\Big\}\Big] &\leqslant \frac{1}{2}\mathbb{E}\Big[\|Y - p^*\|_{2,\hat{D}}^2\Big] = \frac{1}{2}\|Y - p^*\|_{2,D}^2 \\
&\overset{(a)}{\leqslant} \frac{1}{2}\Big(\|Y - c^*\|_{2,D} + \|p^* - c^*\|_{2,D}\Big)^2 \\
&\leqslant \frac{1}{2}\Big(\|Y - c^*\|_{2,D} + \epsilon\Big)^2 \\
&\overset{(b)}{\leqslant} \frac{1}{2}\Big(\|Y - c^*\|_{2,D}^2 + 4\epsilon + \epsilon^2\Big) \\
&\overset{(c)}{\leqslant} 2\mathsf{P}_{opt} + \frac{5}{2}\epsilon,
\end{aligned} \tag{7}$$

where (a) holds from Minkowski's inequality for 2-norm, (b) holds as $\|Y - c^*\|_{2,D} \leqslant 2$, and (c) holds because of the second equality in (1) and that $\mathsf{P}_{opt} = \mathbb{P}\{Y \neq c^*(\mathbf{X})\}$.

Next, we connect the empirical error of $\hat{g}$ to its generalization error. Note that the Vapnik–Chervonenkis (VC) dimension of all functions of the form $\mathsf{sign}[p]$ for some polynomial of degree upto k does not exceed d^{k+1}. Therefore, from VC theory (See Corollary 3.19 in [19]) for any δ, with probability at least $(1-\delta)$, the following inequality holds

$$\begin{aligned}
\mathbb{P}\Big\{Y \neq \hat{g}(\mathbf{X})\Big\} \leqslant \hat{\mathbb{P}}\Big\{Y \neq \hat{g}(\mathbf{X})\Big\} &+ \sqrt{\frac{2\ d^{k+1}}{n}\log\frac{en}{d^{k+1}}} \\
&+ \sqrt{\frac{\log\frac{1}{\delta}}{2n}}.
\end{aligned} \tag{8}$$

Set $\delta = \exp\{-\frac{1}{2}n\epsilon^2\}$. Therefore, the proof is complete by taking the expectation and combining it with the last bound in (7). □

PAC bounds for the Fourier algorithm: Next, we employ a low-degree (Fourier) algorithm (Algorithm 2) for PAC learning with respect to the polynomially approximated hypothesis class.

Theorem 2. *Let D be a joint probability distribution with marginal D_X that is a product probability distribution on $\{-1,1\}^d$. Then, for any $\delta \in [0,1]$, with probability at least $1-\delta$, the Fourier-based algorithm agnostically PAC learns any (ϵ, k)-approximated concept class with generalization error up to*

$$2\mathsf{P}_{opt} + 2\epsilon + O\Big(\sqrt{\frac{d^k c_k}{(k-1)!n}\log\frac{4d^k}{(k-1)!\delta}}\Big), \tag{9}$$

where $c_k \triangleq \max_{\mathcal{S}\subseteq[d],|\mathcal{S}|\leqslant k}\|\psi_{\mathcal{S}}\|_\infty^2$.

Proof. We prove the theorem by characterizing the effect of 2-norm estimation on the error probability. Let

$$p^* = \arg\min_{p\in\mathcal{P}_k}\|Y-p\|_{2,D}^2.$$

From the second equality in (1), the generalization error of $\hat{g}$ in Algorithm 2 satisfies

$$\begin{aligned}\mathbb{P}\Big\{Y\neq\hat{g}(\mathbf{X})\Big\} &= \frac{1}{4}\|Y-\hat{g}\|_{2,D}^2\\ &\leqslant \frac{1}{4}\Big(\|Y-p^*\|_{2,D}+\|p^*-\hat{g}\|_{2,D}\Big)^2\\ &\leqslant \frac{1}{2}\Big(\|Y-p^*\|_{2,D}^2+\|p^*-\hat{g}\|_{2,D}^2\Big),\end{aligned} \tag{10}$$

where the inequality follows from Minkowski's inequality for 2-norm. Observe that (10) is an upper bound on the generalization error in terms of 2-norm quantities. Since p^* minimizes the square loss, the first term in (10) equals

$$\|Y-p^*\|_{2,D}^2 = 1-\|p^*\|_2^2.$$

We proceed by bounding the second term in (10). From Minkowski's inequality for 2-norm and by adding and subtracting $\hat{\Pi}_Y$ as in Algorithm 2, we have that

$$\begin{aligned}\|p^*-\hat{g}\|_2^2 \leqslant \|p^*-\hat{\Pi}_Y\|_2^2+\|\hat{\Pi}_Y-\hat{g}\|_2^2\\ +2\|p^*-\hat{\Pi}_Y\|_2\|\hat{\Pi}_Y-\hat{g}\|_2.\end{aligned} \tag{11}$$

The first term in (11) is bounded from Lemma 1. As a result, $\|p^*-\hat{\Pi}_Y\|_2=\epsilon_n'$, where

$$\epsilon_n' = O\Big(\sqrt{\frac{d^k c_k}{(k-1)!n}\log\frac{4d^k}{(k-1)!\delta}}\Big),$$

with probability at least $(1-\delta)$. As for the second term in (11), we use the identity $|h-\mathsf{sign}[h]| = |1-|h||$ for any function h. Therefore, as $\hat{g}=\mathsf{sign}[\hat{\Pi}_Y]$, we obtain that

$$\begin{aligned}\|\hat{\Pi}_Y-\hat{g}\|_2^2 &= \mathbb{E}\Big[(1-|\hat{\Pi}_Y(\mathbf{X})|)^2\Big]\\ &= 1+\|\hat{\Pi}_Y\|_2^2-2\|\hat{\Pi}_Y\|_1.\end{aligned} \tag{12}$$

Next, we show that the third term in (11) is less than $4\epsilon_n'$. It suffices to show that $\|\hat{\Pi}_Y-\hat{g}\|_2\leqslant 2$. For that, we use the equality in (12). By removing the last term in (12) and taking the square root we have

$$\|\hat{\Pi}_Y-\hat{g}\|_2 \leqslant \sqrt{1+\|\hat{\Pi}_Y\|_2^2}.$$

From the Minkowski's inequality we have that

$$\begin{aligned}\|\hat{\Pi}_Y\|_2 &\leqslant \|p^*\|_2 + \|p^* - \hat{\Pi}_Y\|_2 \\ &\leqslant \|p^*\|_2 + \epsilon'_n \leqslant 1 + \epsilon'_n.\end{aligned}$$

Hence, we get the desired bound $\|\hat{\Pi}_Y - \hat{g}\|_2 \leqslant \sqrt{1 + (1+\epsilon'_n)^2} \leqslant 2$, assuming that $\epsilon'_n \leqslant 1/3$. Combining the bounds for each term in (11), we get

$$\begin{aligned}\|p^* - \hat{g}\|_2^2 &\leqslant {\epsilon'_n}^2 + 1 + \|\hat{\Pi}_Y\|_2^2 - 2\|\hat{\Pi}_Y\|_1 + 4\epsilon'_n \\ &\leqslant 1 + \|\hat{\Pi}_Y\|_2^2 - 2\|\hat{\Pi}_Y\|_1 + 5\epsilon'_n.\end{aligned}$$

We plug this inequality in (10). After rearranging the terms by adding and subtracting $\|p^*\|_1$, we obtain the following inequality

$$\begin{aligned}\mathbb{P}\Big\{Y \neq \hat{g}(\mathbf{X})\Big\} &\leqslant \frac{1}{2}\Big(2 - 2\|p^*\|_1 + 5\epsilon'_n \\ &\quad + 2\big(\|p^*\|_1 - \|\hat{\Pi}_Y\|_1\big) + \big(\|\hat{\Pi}_Y\|_2^2 - \|p^*\|_2^2\big)\Big) \\ &\leqslant 2\mathsf{P}_{opt} + 2\epsilon + 5\epsilon'_n,\end{aligned}$$

where the last inequality follows from Lemma 2 and the following argument for bounding the last two terms in the first inequality:

For the 1-norm difference, the Minkowski's inequality for 1-norm gives

$$\|p^*\|_1 - \|\hat{\Pi}_Y\|_1 \leqslant \|p^* - \hat{\Pi}_Y\|_1 \leqslant \|p^* - \hat{\Pi}_Y\|_2 = \epsilon'_n,$$

where the last inequality follows from the Jensen's inequality implying that $\|\cdot\|_1 \leqslant \|\cdot\|_2$.

For the difference of square of 2-norms, we apply the Minkowski's inequality for 2-norm and obtain

$$\begin{aligned}\|\hat{\Pi}_Y\|_2^2 &\leqslant \|p^*\|_2^2 + \|p^* - \hat{\Pi}_Y\|_2^2 + 2\|p^*\|_2\|p^* - \hat{\Pi}_Y\|_2 \\ &\leqslant \|p^*\|_2^2 + 3\epsilon'_n.\end{aligned}$$

where the last inequality holds as $\|p^*\|_2 \leqslant 1$. □

We end this section by presenting a simplified result of Theorem 2.

Corollary 1. *If the expected value of each X_j satisfies $|\mu_j| \leqslant 1 - \frac{1}{k}$, then the generalization error of the Fourier algorithm is upper bounded by*

$$2\mathsf{P}_{opt} + 2\epsilon + O\Big(\sqrt{\frac{\sqrt{k}(ed)^k}{n}\big(k\log\frac{ed}{k} + \log\frac{2\sqrt{k}}{\delta}\big)}\Big).$$

Proof. From the definition of c_k, we can write

$$c_k = \max_{\mathcal{S}:|\mathcal{S}|\leqslant k} \max_{\mathbf{x}\in\{-1,1\}^d} |\psi_{\mathcal{S}}(\mathbf{x})|^2 \leqslant \max_{\mathcal{S}:|\mathcal{S}|\leqslant k} \prod_{j\in\mathcal{S}} \frac{(1+|\mu_j|)^2}{\sigma_j^2}$$
$$= \max_{\mathcal{S}:|\mathcal{S}|\leqslant k} \prod_{j\in\mathcal{S}} \frac{1}{1-|\mu_j|},$$

where the first inequality is from the definition of $\psi_{\mathcal{S}}$. The laste quality holds as $\sigma_j^2 = 1-\mu_j^2$. Therefore, under the assumption that $|\mu_j| \leqslant 1 - \frac{1}{k}$, the following inequality holds

$$c_k \leqslant \max_{\mathcal{S}:|\mathcal{S}|\leqslant k} \prod_{j\in\mathcal{S}} k \leqslant k^k$$

Hence,

$$\frac{c_k}{(k-1)!} \leqslant \frac{kk^k}{k!} = 2^{(k+1)\log_2 k - \log_2 k!}$$

From Stirling approximation $\log_2 k! = k\log_2 k - k\log_2 e + O(\log_2 k)$. Hence,

$$\frac{c_k}{(k-1)!} \leqslant 2^{k\log_2 e + O(\log_2 k)} = e^k + O(k).$$

Using the above inequality and Theorem 2 in the main text, we obtain the corollary.

□

V. Learning Other Hypothesis Classes

In this section, we extend our results to two other type of concept classes. The first one is called *half-spaces* and the other one is a generalized version of the concentrated hypothesis classes.

A. Half-spaces

In this section, we consider learning another class of functions called half-spaces. More precisely, a half-space a Boolean-valued function of the form

$$c(\mathbf{x}) = \mathsf{sign}[a_0 + \sum_{j=1}^{d} a_j x_j)], \quad \forall \mathbf{x} \in \mathbb{R}^d$$

where $a_j \in \mathbb{R}$. We start with a lower-bound on the optimal classification error of the class.

Lemma 4. *Let D be any joint probability distribution on $\mathbb{R}^d \times \{-1,1\}$ with marginal $D_{\mathbf{x}}$ that is the uniform distribution on $\mathbb{S}^{d-1}$ or jointly Gaussian on $\mathbb{R}^d$. Then, for any $\epsilon > 0$, the minimum generalization error of learning with respect to half-spaces satisfy the following lower bound*

$$\mathsf{P}_{opt} \geqslant \frac{1}{2} - \frac{1}{2}\|p_\epsilon^*\|_{1,D_X} - \epsilon,$$

where p_ϵ^* is a polynomial of degree up to $O(\frac{1}{\epsilon^4})$ minimizing $\|Y - p\|_{2,D}$ among all such polynomials.

The proof of the lemma follows from Lemma 2 and [5]'s result (Theorem 6) on the sign function. This result is stated as

Lemma 5 ([5]). *Let X be a random variable with uniform distribution on $\mathbb{S}^{d-1}$ or jointly Gaussian on $\mathbb{R}^d$. Then, for any $\epsilon > 0$, there exists a polynomial p of degree $O(\frac{1}{\epsilon^4})$ such that $\mathbb{E}\Big[\big(p(\mathbf{X}) - \mathsf{sign}(\mathbf{X})\big)^2\Big] \leqslant \epsilon^2$.*

This lemma makes a connection between half-spaces and the polynomial-approximated class. That said, in the following theorems we show our results for PAC learning using Algorithm 1.

Theorem 3. *Let D be any joint probability distribution on $\mathbb{R}^d \times \{-1, 1\}$, with marginal D_X that is uniform on the unit sphere or jointly Gaussian. Then, $\mathcal{L}_2$-polynomial regression PAC learns half-spaces with expected generalization error up to*

$$2\mathsf{P}_{opt} + 3\epsilon + \sqrt{\frac{d^{O(\frac{1}{\epsilon^4})}}{n} \log \frac{n}{d^{O(\frac{1}{\epsilon^4})}}}.$$

B. Generalized approximated class

Lastly, we finish this paper by extending our results to a more general hypothesis class. Fix a set of functions $e_1(\mathbf{x}), e_2(\mathbf{x}), ..., e_m(\mathbf{x})$ and let $\mathcal{H}$ be a Hilbert space spanned by a these functions. Let $\mathcal{C}$ be a class of functions each of which approximated by elements of $\mathcal{H}$ with square error up to ϵ, that is,

$$\inf_{h \in \mathcal{H}} \|c - h\|_{2,D} \leqslant \epsilon,$$

for any $c \in \mathcal{C}$. As a special case, suppose e_i's are all the functions of the form $e(\mathbf{x}) = \prod_{j \in [d]} x_j^{\alpha_j}$ where α_j's are non-negative integers adding up to k. Then $\mathcal{C}$ is a (k, ϵ)-approximated class as in Section IV.

Theorem 4. *Suppose A is any algorithm that given n training instances finds a function $\hat{h} \in \mathcal{H}$ so that the empirical loss $\|Y - h\|_{2,\hat{D}}$ is minimized. Then, the predictor $\mathsf{sign}[\hat{h}]$ learns $\mathcal{C}$ with expected generalization error up to*

$$2\mathsf{P}_{opt} + 3\epsilon + O\Big(\sqrt{\frac{\mathrm{VC}(\mathcal{C})}{n} \log \frac{n}{\mathrm{VC}(\mathcal{C})}}\Big),$$

where $\mathrm{VC}(\mathcal{C})$ is the VC dimension of $\mathcal{C}$.

ACKNOWLEDGEMENT

This work was supported in part by NSF Center on Science of Information Grants CCF-0939370 and NSF Grants CCF-1524312, CCF-2006440, CCF-2007238, and Google Research Award.

APPENDIX A
PROOF OF LEMMA 1

a) Mean and variance estimations:: We first take into account the effect of the imperfections in mean and variance estimation. For tractability of our analysis, we use a fraction of the training samples just for the mean and variance estimations. As a measure of accuracy of the estimations, we require the differences $|\hat{\mu}_j - \mu_j|$ and $|1 - \frac{\sigma}{\hat{\sigma}}|$ to be sufficiently small with probability close to one. This is a deviation from standard measures of estimations in which the variance of the differences are required to be small. In the following lemma, we bound the estimation errors in terms of the number of the samples.

Lemma 6. *Given $\epsilon_0, \delta_0 \in (0,1)$ the following inequalities hold with probability at least $(1-\delta_0)$*

$$|\hat{\mu}_j - \mu_j| \leqslant \epsilon_0, \qquad |1 - \frac{\sigma_j}{\hat{\sigma}_j}| \leqslant \frac{2\epsilon_0}{\sigma_j^2}, \tag{13}$$

for all $j \in [d]$, provided that atleast $n_0(\epsilon_0, \delta_0) = \frac{2}{\epsilon_0^2} \log \frac{2d}{\delta_0}$ samples are available.

Proof. Form McDiarmid's inequality, for each $j \in [d]$ we have

$$\mathbb{P}\{|\hat{\mu}_j - \mu_j| \geqslant \epsilon_0\} \leqslant 2\exp\{-\frac{n\epsilon_0^2}{2}\}.$$

Therefore, applying the union bound gives

$$\mathbb{P}\Big\{\bigcup_{j=1}^{d} \{|\hat{\mu}_j - \mu_j| \geqslant \epsilon_0\}\Big\} \leqslant 2d\exp\{-\frac{n\epsilon_0^2}{2}\}.$$

Thus, the right-hand side of the above inequality is less than δ_0, if $n \geqslant \frac{2}{\epsilon_0^2}\log(\frac{2d}{\delta_0})$. As a result we obtain the inequalities for the estimation of μ_j's. Next, we prove the inequalities for the estimation of σ_j's. For any fixed $\hat{\mu} \in (-1,1)$, define the function $h_{\hat{\mu}}(x) = \frac{\sqrt{1-x^2}}{\sqrt{1-\hat{\mu}^2}}$. From Taylor's theorem, there exists $\zeta \in (-1,1)$ which is between x and $\hat{\mu}$ such that

$$h_{\hat{\mu}}(x) = 1 - \frac{\zeta(x-\hat{\mu})}{\sqrt{(1-\zeta^2)(1-\hat{\mu}^2)}}.$$

As a result,

$$|h_{\hat{\mu}}(x) - 1| = \frac{|\zeta||x-\hat{\mu}|}{\sqrt{(1-\zeta^2)(1-\hat{\mu}^2)}} \leqslant \frac{|x-\hat{\mu}|}{\sqrt{(1-(\max\{x,\hat{\mu}\})^2)(1-\hat{\mu}^2)}}.$$

Now by setting $x = \mu_j$ and that $|\hat{\mu}_j - \mu_j| \leqslant \epsilon_0$, we have

$$|\frac{\sigma_j}{\hat{\sigma}_j} - 1| = |h_{\hat{\mu}}(\mu) - 1| \leqslant \frac{\epsilon_0}{\hat{\sigma}\min\{\hat{\sigma}, \sigma\}}.$$

Note that, $|\hat{\mu}_j| \leqslant |\mu_j| + \epsilon_0$. Therefore,

$$\hat{\sigma}_j^2 \geqslant 1 - (|\mu_j| + \epsilon_0)^2 \geqslant \sigma_j^2 - 2\epsilon_0|\mu_j| - \epsilon_0^2 \geqslant \sigma_j^2 - 3\epsilon_0.$$

As a result,

$$|\frac{\sigma_j}{\hat{\sigma}_j} - 1| \leqslant \frac{\epsilon_0}{\sigma_j^2 - 3\epsilon_0} \leqslant \frac{2\epsilon_0}{\sigma_j^2},$$

which completes the proof of the lemma. □

Now we proceed with the proof of the lemma. Let $\bar{\Pi}_Y$ denote the version of $\hat{\Pi}_Y$ under the assumption that $\hat{\mu}_j = \mu_j$ and $\hat{\sigma}_j = \sigma_j$ for all $j \in [d]$. Also, let B be the even that the inequalities in (13) hold. From Minkowsky's inequality, by adding and subtracting $\bar{\Pi}_Y$ we have

$$\|\Pi_Y - \hat{\Pi}_Y\|_2 \leqslant \underbrace{\|\Pi_Y - \bar{\Pi}_Y\|_2}_{V} + \underbrace{\|\bar{\Pi}_Y - \hat{\Pi}_Y\|_2}_{W}.$$

Let V and W denote the first and the second term above, respectively. We proceed by the following lemmas.

Lemma 7. *Given any $\delta > 0$, the inequality $\|\Pi_Y - \bar{\Pi}_Y\|_2 \leqslant \sqrt{\frac{2d^k c_k}{(k-1)!n} \log \frac{2d^k}{(k-1)!\delta}}$ holds with probability $(1-\delta)$.*

Proof. Recall that $\bar{\Pi}_Y$ is defined as

$$\bar{\Pi}_Y(x^d) \triangleq \sum_{\mathcal{S}:|\mathcal{S}|\leqslant k} \bar{f}_S \psi_{\mathcal{S}}(x^d),$$

where the Fourier-estimates $\bar{f}_S$ are defined as $\bar{f}_S \triangleq \frac{1}{n}\sum_i Y(i)\psi_{\mathcal{S}}(X(i))$. In addition, by definition of the projection function Π_Y, we have

$$\Pi_Y(\mathbf{x}) = \sum_{\mathcal{S}:|\mathcal{S}|\leqslant k} f_{\mathcal{S}}\ \psi_{\mathcal{S}}(\mathbf{x}), \qquad \forall \mathbf{x} \in \mathcal{X}^d.$$

Therefore, from Parseval's identity, the 2-norm factors as

$$\|\Pi_Y - \bar{\Pi}_Y\|_2^2 = \sum_{\mathcal{S}:|\mathcal{S}|\leqslant k} |f_{\mathcal{S}} - \bar{f}_S|^2.$$

In what follows, we show that $|f_{\mathcal{S}} - \bar{f}_S| \leqslant \epsilon$ for all subsets $\mathcal{S} \subseteq [d]$ with $|\mathcal{S}| \leqslant k$. Note that $\bar{f}_S$ is a function of the training random samples $(X(i), Y(i)), i = 1, 2, ..., n$. Observe that $\mathbb{E}[\bar{f}_S] = f_{\mathcal{S}}$ which implies that $\bar{f}_S$ is an unbiased estimation of $f_{\mathcal{S}}$. Since the samples are drawn independent and identically distributed (i.i.d.), we apply McDiarmid's inequality to bound the probability of the event $|f_{\mathcal{S}} - \bar{f}_S| \geqslant \epsilon'$.

For that, fix $i \in [d]$ and suppose $(\mathbf{X}(i), Y(i))$ in the training set is replaced with an i.i.d. copy $(\tilde{\mathbf{X}}(i), \tilde{Y}(i))$. With this replacement $\bar{f}_S$ is changed to another random variable denoted by $\tilde{f}_S$. Then

$$\begin{aligned}
|\bar{f}_S - \tilde{f}_S| &= \frac{1}{n}|Y(i)\psi_{\mathcal{S}}(\mathbf{X}(i)) - \tilde{Y}(i)\psi_{\mathcal{S}}(\tilde{\mathbf{X}}(i))| \\
&\leqslant \frac{1}{n}|Y(i)\psi_{\mathcal{S}}(\mathbf{X}(i))| + |\tilde{Y}(i)\psi_{\mathcal{S}}(\tilde{\mathbf{X}}(i))| \\
&\leqslant \frac{1}{n}|\psi_{\mathcal{S}}(\mathbf{X}(i))| + |\psi_{\mathcal{S}}(\tilde{\mathbf{X}}(i))| \\
&\leqslant \frac{2}{n}\|\psi_{\mathcal{S}}\|_\infty,
\end{aligned}$$

where $\|\psi_S\|_\infty = \max_{\mathbf{x}} |\psi_S(\mathbf{x})|$. Let $c_k = \max_{\mathcal{S}\subseteq[d],|\mathcal{S}|\leqslant k}\|\psi_{\mathcal{S}}\|_\infty^2$. Then, from McDiarmid's inequality, for any $\epsilon' \in (0,1)$

$$\mathbb{P}\Big\{\max_{\mathcal{S}:|\mathcal{S}|\leqslant k}|\bar{f}_S - f_{\mathcal{S}}| \geqslant \epsilon'\Big\} \leqslant 2\Big[\sum_{m=0}^{k}\binom{d}{m}\Big]\exp\big\{-\frac{n\epsilon'^2}{2c_k}\big\}, \tag{14}$$

where we also used the union bound. For $k \leqslant d/2$, we obtain that

$$\sum_{m=0}^{k}\binom{d}{m} \leqslant k\frac{d^k}{k!}.$$

As a result, with probability at least $(1-\delta)$, $\max_{\mathcal{S}:|\mathcal{S}|\leqslant k}|\bar{f}_S - f_{\mathcal{S}}| \leqslant \sqrt{\frac{2c_k}{n}\log\frac{2d^k}{(k-1)!\delta}}$. Hence, we with probability at least $(1-\delta)$

$$\|\Pi_Y - \bar{\Pi}_Y\|_2^2 \leqslant \frac{2d^k c_k}{(k-1)!n}\log\frac{2d^k}{(k-1)!\delta},$$

and the proof is complete by taking the square root of both sides. □

Lemma 8. *Conditioned on B, the inequalities $\|\bar{\Pi}_Y - \hat{\Pi}_Y\|_\infty \leqslant \lambda(\epsilon)$ hold, almost surely, for all k-element subsets $\mathcal{J} \subset [d]$, where λ is a function satisfying $\lambda(\epsilon_0) = O(\frac{kd^k c_k}{(k-1)!}\epsilon_0)$ as $\epsilon_0 \to 0$.*

Recall that the function $\bar{\Pi}_Y$ is defined as

$$\bar{\Pi}_Y(x^d) \triangleq \sum_{\mathcal{S}:|\mathcal{S}|\leqslant k}\bar{f}_S\psi_{\mathcal{S}}(x^d),$$

where the Fourier-estimates $\bar{f}_S$ are defined as

$$\bar{f}_S \triangleq \frac{1}{n}\sum_i Y(i)\psi_{\mathcal{S}}(X(i)).$$

From triangle inequality for ∞-norm and the definition of $\hat{\Pi}_Y$ and $\bar{\Pi}_Y$ we obtain

$$\|\hat{\Pi}_Y - \bar{\Pi}_Y\|_\infty \leqslant \sum_{\mathcal{S}:|\mathcal{S}|\leqslant k}\|\hat{f}_S\ \widehat{\psi}_{\mathcal{S}} - \bar{f}_S\ \psi_S\|_\infty. \tag{15}$$

Again by triangle inequality and by adding and subtracting $\bar{f}_S \widehat{\psi}_{\mathcal{S}}$, we obtain that

$$\begin{aligned}\|\hat{f}_S\ \widehat{\psi}_{\mathcal{S}} - \bar{f}_S\ \psi_S\|_\infty &\leqslant \|\hat{f}_S\ \widehat{\psi}_{\mathcal{S}} - \bar{f}_S\ \widehat{\psi}_{\mathcal{S}}\|_\infty + \|\bar{f}_S\ \widehat{\psi}_{\mathcal{S}} - \bar{f}_S\ \psi_S\|_\infty \\ &= |\hat{f}_S - \bar{f}_S|\ \|\widehat{\psi}_{\mathcal{S}}\|_\infty + |\bar{f}_S|\ \|\widehat{\psi}_{\mathcal{S}} - \psi_S\|_\infty.\end{aligned}$$

Next, note that from triangle inequality

$$|\hat{f}_S - \bar{f}_S| \leqslant \frac{1}{n}\sum_i |\widehat{\psi}_{\mathcal{S}}(\mathbf{x}(i)) - \psi_{\mathcal{S}}(\mathbf{x}(i))| \leqslant \|\psi_{\mathcal{S}} - \widehat{\psi}_{\mathcal{S}}\|_\infty.$$

Therefore,

$$\|\hat{f}_S\ \widehat{\psi}_{\mathcal{S}} - \bar{f}_S\ \psi_S\|_\infty \leqslant \big(\|\widehat{\psi}_{\mathcal{S}}\|_\infty + |\bar{f}_S|\big)\|\widehat{\psi}_{\mathcal{S}} - \psi_S\|_\infty. \tag{16}$$

We proceed by bounding each term above. As for the first term we have, that $\|\widehat{\psi}_{\mathcal{S}}\|_\infty \leqslant \|\psi_{\mathcal{S}}\|_\infty + \|\widehat{\psi}_{\mathcal{S}} - \psi_{\mathcal{S}}\|_\infty$. As for the second term, we have

$$\bar{f}_S = \frac{1}{n}\sum_i Y(i)\psi_{\mathcal{S}}(\mathbf{X}(i)) \leqslant \|\psi_{\mathcal{S}}\|_\infty.$$

Lastly, the third term is bounded using the following lemma.

Lemma 9. *Conditioned on B, the inequality $\|\psi_{\mathcal{S}} - \widehat{\psi}_{\mathcal{S}}\|_\infty \leqslant \gamma(\epsilon_0)$ holds, almost surely, where γ is a function satisfying $\gamma(\epsilon_0) = O(k\epsilon_0\sqrt{c_k})$ as $\epsilon_0 \to 0$.*

Before proving this lemma, we complete our argument. As a result of this lemma and using the triangle inequality, we obtain from (16) that

$$\begin{aligned}\|\hat{f}_S\ \widehat{\psi}_{\mathcal{S}} - \bar{f}_S\ \psi_S\|_\infty &\leqslant \big(2\|\psi_{\mathcal{S}}\|_\infty + \|\widehat{\psi}_{\mathcal{S}} - \psi_{\mathcal{S}}\|_\infty\big)\|\widehat{\psi}_{\mathcal{S}} - \psi_S\|_\infty \\ &\leqslant \big(2\sqrt{c_k} + \gamma(\epsilon_0)\big)\gamma(\epsilon_0).\end{aligned}$$

Lastly, from (15) we get the following bound

$$\|\hat{\Pi}_Y - \bar{\Pi}_Y\|_\infty \leqslant \lambda(\epsilon_0) \triangleq \frac{d^k}{(k-1)!}\big(2\sqrt{c_k}\gamma(\epsilon_0) + \gamma^2(\epsilon_0)\big).$$

It is not difficult to check that $\lambda(\epsilon_0) = O(\frac{kd^k c_k}{(k-1)!}\epsilon_0)$ as $\epsilon_0 \to 0$. Now it remains to prove Lemma 9 which is given below:

Proof of Lemma 9: We start with the triangle inequality for ∞-norm by adding and subtracting $b_{\mathcal{S}}\psi_{\mathcal{S}}$:

$$\|\psi_{\mathcal{S}} - \widehat{\psi}_{\mathcal{S}}\|_\infty \leqslant \|\psi_{\mathcal{S}} - b_{\mathcal{S}}\psi_{\mathcal{S}}\|_\infty + \|b_{\mathcal{S}}\psi_{\mathcal{S}} - \widehat{\psi}_{\mathcal{S}}\|_\infty.$$

Note that $b_{\mathcal{S}}\psi_{\mathcal{S}} \equiv \prod_{j\in\mathcal{S}} \frac{x_j-\mu_j}{\hat{\sigma}_i}$. Now, using the triangle inequality on the second term above, we have

$$\begin{aligned}
\|b_{\mathcal{S}}\psi_{\mathcal{S}} - \widehat{\psi}_{\mathcal{S}}\|_\infty &= \|b_{\mathcal{S}}\psi_{\mathcal{S}} \pm \big(\sum_{l\in\mathcal{S}}\prod_{j\leqslant l}\frac{x_j-\hat{\mu}_j}{\hat{\sigma}_i}\prod_{r>l}\frac{x_r-\mu_r}{\hat{\sigma}_r}\big) - \widehat{\psi}_{\mathcal{S}}\|_\infty \\
&\leqslant \sum_{l\in\mathcal{S}}\frac{|\mu_l-\hat{\mu}_l|}{\hat{\sigma}_l}\,\|\prod_{j<l}\frac{(x_j-\hat{\mu}_j)}{\hat{\sigma}_j}\prod_{r>l}\frac{(x_r-\mu_r)}{\hat{\sigma}_r}\|_\infty \\
&\leqslant \frac{\epsilon}{\sigma_{\min}}\sum_{l\in\mathcal{S}}\,\|\prod_{j<l}\frac{(x_j-\hat{\mu}_j)}{\hat{\sigma}_j}\prod_{r>l}\frac{(x_r-\mu_r)}{\hat{\sigma}_r}\|_\infty \\
&\leqslant \frac{\epsilon}{\sigma_{\min}}\sum_{l\in\mathcal{S}}\,\prod_{j<l}\frac{(1+|\hat{\mu}_j|)}{\hat{\sigma}_j}\prod_{r>l}\frac{(1+|\mu_r|)}{\hat{\sigma}_r} \\
&\overset{(a)}{\leqslant} \frac{\epsilon}{\sigma_{\min}}\sum_{l\in\mathcal{S}}\,\prod_{j<l}\frac{(1+|\mu_j|)(1+\epsilon)}{\hat{\sigma}_j}\prod_{r>l}\frac{(1+|\mu_r|)}{\hat{\sigma}_r} \\
&\overset{(b)}{\leqslant} \frac{\epsilon}{\sigma_{\min}}b_{\mathcal{S}}\sum_{l\in\mathcal{S}}\,\prod_{j\in\mathcal{S}}\frac{(1+|\mu_j|)(1+\epsilon)}{\sigma_j} \\
&\overset{(c)}{\leqslant} \frac{k\epsilon}{\sigma_{\min}}b_{\mathcal{S}}(1+\epsilon)^k\|\psi_{\mathcal{S}}\|_\infty,
\end{aligned}$$

where (a) follows from the inequality $(1+|\hat{\mu}_j|) \leqslant (1+|\mu_j|)(1+\epsilon)$, and (b) follows from $(1+|\mu_j|) \leqslant (1+|\mu_j|)(1+\epsilon)$. Lastly, (c) holds as $|\mathcal{S}| \leqslant k$ and because $\|\psi_{\mathcal{S}}\|_\infty = \prod_{j\in\mathcal{S}}\frac{1+|\mu_j|}{\sigma_j}$.

$$\|\psi_{\mathcal{S}} - \widehat{\psi}_{\mathcal{S}}\|_\infty \leqslant |1-b_{\mathcal{S}}|\|\psi_{\mathcal{S}}\|_\infty + \frac{k\epsilon}{\sigma_{\min}}b_{\mathcal{S}}(1+\epsilon)^k\|\psi_{\mathcal{S}}\|_\infty. \tag{17}$$

From the assumption of the lemma and the definition of $b_{\mathcal{S}}$ we obtain that

$$1-(1+\epsilon)^{|S|} \leqslant 1-b_{\mathcal{S}} \leqslant 1-(1-\epsilon)^{|S|}.$$

Since $\epsilon \in (0,1)$ and $|S| \leqslant k$, then $(1-\epsilon)^{|S|} \geqslant 1-k\epsilon$. Also, from the fact that $(1+x) \leqslant e^x$ for all $x \in \mathbb{R}$, we obtain

$$1-e^{k\epsilon} \leqslant 1-b_{\mathcal{S}} \leqslant k\epsilon \leqslant e^{k\epsilon}-1. \tag{18}$$

Lastly, combining (17) and (18) gives the following inequality

$$\|\psi_{\mathcal{S}} - \widehat{\psi}_{\mathcal{S}}\|_\infty \leqslant (e^{k\epsilon}-1)\|\psi_{\mathcal{S}}\|_\infty + \frac{k\epsilon}{\sigma_{\min}}(1+\epsilon)^{2k}\|\psi_{\mathcal{S}}\|_\infty.$$

The proof is complete by noting that $\|\psi_{\mathcal{S}}\|_\infty \leqslant \sqrt{c_k}$. ■

From Lemma 8, we know that W is measurable with respect to B. In particular, conditioned on B, $W \leqslant \lambda(\epsilon_0)$. Therefore, from the above lemmas and using the inequality $\|\cdot\|_2 \leqslant \|\cdot\|_\infty$, we have, with probability $(1-\delta_0)(1-\delta)$ that

$$\|\Pi_Y - \bar{\Pi}_Y\|_2 \leqslant \sqrt{\frac{2d^k c_k}{(k-1)!n}\log\frac{2d^k}{(k-1)!\delta}} + \lambda(\epsilon_0).$$

Now set $\epsilon_0 = \sqrt{\frac{2}{n_0}\log\frac{2d}{\delta}}$ with $\delta_0 = \delta$. Then, with $n_0 = O(n)$, we get with probability $(1-\delta)^2$ that $\|\Pi_Y - \bar{\Pi}_Y\|_2 = O\Big(\sqrt{\frac{2d^k c_k}{(k-1)!n}\log\frac{2d^k}{(k-1)!\delta}}\Big)$. Now the proof is complete by changing δ to $\delta/2$ and noting that $(1-\delta/2)^2 \geqslant 1-\delta$. □

REFERENCES

[1] L. G. Valiant, "A theory of the learnable," *Communications of the ACM*, vol. 27, no. 11, pp. 1134–1142, nov 1984.

[2] M. J. Kearns, R. E. Schapire, and L. M. Sellie, "Toward efficient agnostic learning," *Machine Learning*, vol. 17, no. 2-3, pp. 115–141, 1994.

[3] J. Suykens and J. Vandewalle, *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999.

[4] N. Linial, Y. Mansour, and N. Nisan, "Constant depth circuits, Fourier transform, and learnability," *J. ACM*, vol. 40, no. 3, pp. 607–620, 1993.

[5] A. T. Kalai, A. R. Klivans, Yishay Mansour, and R. A. Servedio, "Agnostically learning halfspaces," in *Proc. 46th Annual IEEE Symp. Foundations of Computer Science (FOCS'05)*, Oct. 2005, pp. 11–20.

[6] E. Mossel, R. O'Donnell, and R. A. Servedio, "Learning functions of k relevant variables," *J. Comput. Syst. Sci*, vol. 69, no. 3, pp. 421–434, 2004.

[7] E. Mossel, R. O'Donnell, and R. P. Servedio, "Learning juntas," in *Proc. ACM Symp. on Theory of Computing*, 2003, pp. 206–212.

[8] E. Blais, R. O'Donnell, and K. Wimmer, "Polynomial regression under arbitrary product distributions," *Machine learning*, vol. 80, no. 2-3, pp. 273–294, 2010.

[9] M. Heidari, G. I. Shamir, and W. Szpankowski, "Fourier-based universal learning," *Journal of Machine Learning Research (JMLR)*, 2020.

[10] A. R. Klivans, P. M. Long, and R. A. Servedio, "Learning halfspaces with malicious noise." *Journal of Machine Learning Research*, vol. 10, no. 12, 2009.

[11] A. Birnbaum and S. S. Shwartz, "Learning halfspaces with the zero-one loss: time-accuracy tradeoffs," in *Advances in Neural Information Processing Systems*, 2012, pp. 926–934.

[12] I. Diakonikolas, T. Gouleakis, and C. Tzamos, "Distribution-independent pac learning of halfspaces with massart noise," in *Advances in Neural Information Processing Systems*, 2019, pp. 4749–4760.

[13] A. Klivans and P. Kothari, "Embedding hard learning problems into gaussian space," in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2014)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2014.

[14] V. Guruswami and P. Raghavendra, "Hardness of learning halfspaces with noise," *SIAM Journal on Computing*, vol. 39, no. 2, pp. 742–765, 2009.

[15] P. Awasthi, M. F. Balcan, and P. M. Long, "The power of localization for efficiently learning linear separators with noise," *Journal of the ACM*, vol. 63, no. 6, pp. 1–27, feb 2017.

[16] A. Daniely, "A ptas for agnostically learning halfspaces," in *Conference on Learning Theory*, 2015, pp. 484–502.

[17] M. L. Furst, J. C. Jackson, and S. W. Smith, "Improved learning of AC^0 functions," in *COLT*, vol. 91, 1991, pp. 317–325.

[18] R. O'Donnell, *Analysis of boolean functions*. Cambridge University Press, 2014.

[19] M. N. Y. U. Mohri, A. (Google, I. Rostamizadeh, A. U. of California, and B. Talwalkar, *Foundations of Machine Learning*. MIT Press Ltd, 2018.