$\mathcal{L}_2$-Polynomial Regression is Agnostic PAC Learner for k-Juntas

Mohsen Heidari MHEIDARI@PURDUE.EDU
Department of Computer Sciences, Purdue University
West Lafayette, IN, USA

Wojciech Szpankowski SZPAN@PURDUE.EDU
Department of Computer Sciences, Purdue University
West Lafayette, IN, USA

Abstract

Agnostic PAC learning is a theoretical framework to derive distribution-free guarantees. Learning algorithms often rely on loss functions other than the natural 0-1 loss to gain computational efficiency. Particularly, learning with least absolute regression ($\mathcal{L}_1$ distance) and least square regression ($\mathcal{L}_2$ distance) are of interest. The $\mathcal{L}_1$-based approaches are proved to be *agnostic* PAC learners for a variety of important concept classes such as *juntas*, and *half-spaces*. On the other hand, $\mathcal{L}_2$-based approaches are preferable because of computational efficiency — with complexity linear in the number of samples. However, for such approaches, PAC guarantees have been proved only under distributional assumptions such as *realizability* ($P_{opt} = 0$) and uniform input distribution. The question whether $\mathcal{L}_2$ approaches are agnostic PAC learners has been open up to now.

In this paper, we prove that agnostic PAC learning of k-juntas is possible using $\mathcal{L}_2$ polynomial regression; hence addressing this question for k-juntas. Moreover, we propose a generalization of Linial et al. (1993)'s low-degree algorithm for learning with arbitrary (non-product) distributions. We prove that this algorithm is also an agnostic PAC learner for k-juntas. We establish our results by developing a framework using Hilbert spaces as a proxy to analyze PAC learning problems for non-product distributions. Using a joint Hilbert space incorporating the stochastic relation between the true label and the predictor, we demonstrate a fundamental connection between agnostic PAC learning with 0-1 loss and the least square error in polynomial regressions.

Keywords: Agnostic PAC, L2 Polynomial Regression, Juntas, Boolean Fourier Expansion

1. Introduction

To gain computational efficiency or analytic tractability, many conventional learning methods such as support-vector machine (SVM) rely on intermediate loss functions other than the natural $0-1$ loss. Absolute difference ($\mathcal{L}_1$ distance) is an example. It is known that polynomial regression under $\mathcal{L}_1$ distance leads to *agnostic probably approximately correct* (PAC) learners (Kalai et al., 2008) for various hypothesis classes such as k-juntas, *polynomial-approximated* predictors, and *half-spaces*. However, computing $\mathcal{L}_1$ distance is an intensive task computationally.

Square loss ($\mathcal{L}_2$ distance) on the other hand is an alternative with computational complexity linear in the size of the data. This has been an incentive to use learning algorithms such as the *low-degree* algorithm (Linial et al., 1993) and LS-SVM (Suykens and Vandewalle, 1999). From the learning theoretic perspective, PAC learning using $\mathcal{L}_2$-based approaches has been studied for the aforementioned concept classes and with distributional assumptions (Linial et al., 1993; Kalai et al., 2008; Jackson, 2006). For instance, under the *realizability* assumption, where zero generalization 0-1 loss is possible ($P_{opt} = 0$), the $\mathcal{L}_2$-polynomial regression is a PAC learner. In addition to the realizability assumption, under the uniform input distribution, the low-degree algorithm is also a PAC learner (Mossel et al., 2004, 2003; Blais et al., 2010). Under the distribution-free (*agnostic)* setting, PAC bounds of the form $c(P_{opt})$ with $c \geq 2$, and as high as $c = 8$, have been proved so far for various concept classes (Kalai et al., 2005; Kearns et al., 1994; Jackson, 2006). Therefore, *agnostic* PAC learnability of $\mathcal{L}_2$-based approaches is still unknown.

This paper resolves this problem for learning k-juntas, i.e., functions over d inputs whose output depends on at most $k < d$ variables. Notably, we prove that agnostic PAC learning is achievable using $\mathcal{L}_2$-polynomial regression for k-juntas. Moreover, we present a Fourier-based algorithm that generalizes the well-known low-degree algorithm to arbitrary distributions on binary inputs (non-uniform and no product). We show that this algorithm is also an agnostic PAC learner w.r.t to k-juntas. Learning/testing juntas has been studied extensively in the literature (Kalai et al., 2005; Klivans et al., 2009; Birnbaum and Shwartz, 2012; Diakonikolas et al., 2019). Learning juntas are motivated by the *feature selection* problem in machine learning (Guyon and Elisseeff, 2003; Heidari et al., 2021b).

1.1 Summary of the Contributions

Learning k-juntas with least square regression: The focus of this paper is PAC learning of k-junta class using $\mathcal{L}_2$-regression and with the usual 0-1 loss. The training set in this model contains n samples $\{(\mathbf{x}(i), y(i))\}_{i=1}^{n}$ with binary label $y(i) \in \{-1, 1\}$. The objective of the $\mathcal{L}_2$-polynomial regression is to minimize the empirical square loss between the target label y and a polynomial $p(\mathbf{x})$ of degree up to k. Given such a polynomial, a predictor g is created by simply taking the sign of this polynomial as $g(\mathbf{x}) = \text{sign}[p(x)]$.

The first main result of this paper is that this approach agnostically PAC learns k-juntas. More precisely, with probability at least $(1-\delta)$, the generalization loss of the predictor g is within a small deviation of the optimal loss among all k-juntas, i.e., $\mathbb{P}\{Y \neq g(\mathbf{X})\} \leq P_{opt} + \epsilon$, with P_{opt} being the optimal loss in k-junta class. We further obtain bounds on the sample complexity of this approach. More formally, we present the following theorem.

Theorem 1 *The $\mathcal{L}_2$-polynomial regression with degree limit of $k \leq d$ agnostically PAC learns k-juntas with sample complexity less than $O(\frac{k2^k}{\epsilon^2} \log \frac{d}{\delta\epsilon^2})$ and running time linear in n and polynomial in d.*

Learning with Fourier algorithm: The low-degree algorithm is an efficient algorithm for binary samples with uniform distribution over $\{-1, 1\}^d$. It relies on the Fourier expansion for Boolean functions (Wolf, 2008). With this approach, one can perform $\mathcal{L}_2$ polynomial regression more efficiently by estimating an appropriate set of Fourier coefficients. This technique leverages a Parseval-type identity connecting the square loss to the estimation of the Fourier coefficients.

In this paper, we present a more general version of the low-degree algorithm incorporating arbitrary probability distributions on the Boolean cube. We refer to this approach as the Fourier algorithm. For that, we employ a distribution-dependent Fourier expansion by applying a Gram-Schmidt orthogonalization to find an orthonormal basis for Boolean functions. We then show that this algorithm also agnostically PAC learns the k-junta class.

Theorem 2 *Given $k < d$, the Fourier algorithm agnostically PAC-learns k-juntas with sample complexity $O(\frac{c_k 2^k}{\epsilon^2} \log \frac{d^k}{\delta})$ and with and running time linear in n and polynomial in d.*

Table 1 compares our approach with well-known PAC learning algorithms. To the best of our knowledge, our algorithms improve both the sample and computational complexities when compared to previously known PAC-learning algorithms. In particular, we improve the loss of the low-degree algorithm (Linial et al., 1993) from $2P_{opt}$ to P_{opt}, which has a comparable computational complexity. As compared to Kalai et al. (2008)'s approach using $\mathcal{L}_1$-polynomial regression, we obtain a lower sample complexity and significantly lower computational complexity (especially for large data sets). These algorithms are explained in Section 1.2.

Our Approach for Deriving PAC Bounds: We develop a framework using Hilbert spaces as a proxy to derive our PAC learning bounds. We consider a joint Hilbert space for functions on the feature-label set $\mathcal{X} \times \mathcal{Y}$, incorporating the sample-label relation and the underlying joint distribution D. In this perspective, the true labeling is viewed as the function $(\mathbf{x}, y) \mapsto y$ and predictors are mappings of the form $(\mathbf{x}, y) \mapsto g(\mathbf{x})$. The inner product in this space is distribution-dependent and is expressed as $\langle f, h \rangle_D = \mathbb{E}_D[f(\mathbf{X}, Y) g(\mathbf{X}, Y)]$ for any pair of functions f, g. To analyze the empirical loss, we consider a similar Hilbert space but with the empirical distribution instead of D. This technique enables bridging the PAC learning model with elegant theories on vector (Hilbert) spaces. Particularly, we make a connection between the 0-1 loss and amenable quantities such as 1-norm and 2-norm (see Lemma 1 and 2 in Section 3). A notable feature of this approach is that the expressions are quite compact and insightful.

Learning other hypothesis classes: In addition to k-juntas, we study learning with another related hypothesis class called polynomial-approximated functions. Given a positive integer k and $\epsilon > 0$, it consists of Boolean-valued functions approximated by a degree k polynomial with square error up to ϵ^2. It is known that with $\mathcal{L}_2$ polynomial regression, one obtains error up to $8P_{opt}$. Further, this PAC bound can be reduced to $2P_{opt}$ using a nondeterministic rounding proposed in (Blum et al., 1994; Jackson, 2006), instead of taking the sign. As a by-product of our Hilbert-space analysis, we show in Section 4, that the bound $2P_{opt}$ is still achievable even with deterministic rounding with the sign function. We show that in a distribution-free setting the Fourier algorithm and the $\mathcal{L}_2$-polynomial regression learn this class with error $2P_{opt}$. Hence, we extend the result of (Jackson, 2006) to non-uniform and non-product distributions.

1.2 Related Works

Learning k-juntas: There have been a large body of works on learning and testing of juntas (Mossel et al., 2004; Bshouty and Costa, 2016; Liu et al., 2019; Arpe and Mossel, 2008; Fischer et al., 2004; Servedio et al., 2015; De et al., 2019; Vempala and Xiao, 2011; Chen et al., 2021; Iyer et al., 2021). Juntas has been of significant interest in learning theory as they are connected to other fundamental problems such as learning with feature selection (Guyon and Elisseeff, 2003), DNF formulas, and decision trees (Mossel et al., 2004). Particularly, learning with feature selection can be expressed as learning k-juntas (with k out of d features). Additionally, every k-junta is implemented by a decision tree or DNF formula of size 2^k and conversely, any size-k decision tree is also a k-junta, and any k-term DNF is ϵ-approximated by a $k \log(\frac{k}{\epsilon})$-junta. Hence, obtaining efficient algorithms for these problems is closely related to learning juntas (Mossel et al., 2004). PAC learning with respect to k-juntas has been studied using various approaches. We briefly review the approaches for learning these concept classes below and summarize in Table 1.

Table 1: Comparison of the approaches for PAC-learning of k-juntas.

Approach	Sample Cmpx.	Comput. Cmpx.	PAC bounds
ERM	$O(\frac{k2^k}{\epsilon^2}\log\frac{d}{\delta})$	$O(nd^k 2^{2^k})$	$P_{opt}+\epsilon$
$\mathcal{L}_1$-Poly. Reg. (Kalai et al., 2008)	$O(d^{\Theta(k)/\epsilon})$	$O(n^2 d^{(3+\omega)3k})$	$P_{opt}+\epsilon$
Low-degree (*uniform dist.*) (Linial et al., 1993)	$O(2^k\log\frac{d^k}{\delta})$	$O(nkd^k)$	• $8P_{opt}+\epsilon$ (Linial et al., 1993) • $2P_{opt}+\epsilon$ (randomized rounding) (Jackson, 2006) • $\frac{1}{4}+P_{opt}(1-P_{opt})+\epsilon$ (Kearns et al., 1994) • $P_{opt}+\epsilon$, [**This work**]
Generalized Fourier [**This work**]	$O(\frac{c_k 2^k}{\epsilon^2}\log\frac{d^k}{\delta})$	$O(nk(2d)^k)$	$P_{opt}+\epsilon$
$\mathcal{L}_2$-Poly. Reg.	$O(\frac{k2^k}{\epsilon^2}\log\frac{d}{\epsilon^2\delta})$	$O(nd^{\Theta(k)})$	• $2P_{opt}+\epsilon$ (randomized rounding) (Jackson, 2006) • $P_{opt}+\epsilon$, [**This work**]

Naive Empirical Risk Minimization (ERM): This is the usual exhaustive search over all predictors to minimize the empirical loss. For k-juntas, ERM is an agnostic PAC learning algorithm with sample complexity $O(\frac{k2^k}{\epsilon^2}\log\frac{d}{\delta})$ and computational complexity $O(nd^k2^{2^k})$ (Shalev-Shwartz and Ben-David, 2014). With the computational complexity of doubly exponential with respect to k, ERM is prohibitive even for small values of k.

Learning with $\mathcal{L}_1$ Regression. Kalai et al. (2008) introduced polynomial regression as an approach for PAC learning with the $0-1$ loss function. They showed that $\mathcal{L}_1$-Polynomial regression agnostically PAC learns with respect to (k,ϵ)-concentrated hypothesis class which includes k-juntas. The main idea behind this approach is first to find a degree-k polynomial minimizing the absolute loss. Then a predictor is obtained by taking the sign of this polynomial. Adopting this algorithm to k-juntas requires a sample complexity $O(d^{\Theta(k)/\epsilon})$. With a *linear programming* implementation, the computational complexity of this algorithm is $O(n^2d^{(3+\omega)3k})$, where $\omega<2.4$ is the matrix-multiplication exponent. A more efficient implementation is SVM with degree-k polynomial kernel and without any regularization (Kalai et al., 2008). However, this implementation PAC learns in the non-agnostic setting, that is when the target labeling function itself belongs to the hypothesis class (Blais et al., 2010).

Learning with $\mathcal{L}_2$ Polynomial Regression. This approach is similar to its $\mathcal{L}_1$ counterpart with absolute error replaced by the square loss. PAC learning using this approach has been studied in (Kalai et al., 2008; Jackson, 2006). In the agnostic setting, It is shown that this approach is a *weak learner* with error $8P_{opt}$. One can show that, instead of taking the sign, with the use of a nondeterministic rounding proposed in (Blum et al., 1994; Jackson, 2006), the PAC bound can be reduced to $2P_{opt}$. Using our Hilbert-space analysis, we show that for k-juntas P_{opt} is obtained. For other classes, in Section 4, we show that the bound $2P_{opt}$ is achievable even with deterministic rounding. Fast implementations of $\mathcal{L}_2$ regression with complexity linear in the number of samples has been studied (Drineas et al., 2006, 2010).

Low-degree Algorithms Linial et al. (1993) investigated PAC learning from an alternative perspective and introduced the well-known "Low-Degree Algorithm". They provide theoretical guarantees under the *uniform* and *known* distribution on $\{-1,1\}^d$ of the examples. As Kalai et al. (2008) showed, under the uniform distribution, the low-degree algorithm agnostically learns the (k,ϵ)-concentrated hypothesis classes as well with an error up to $8P_{opt}+\epsilon$. Again the factor 8 can be reduced to 2 if nondeterministic rounding is allowed. Moreover, Kearns et al. (1994) proved the PAC bound of $(\frac{1}{4}+P_{opt}(1-P_{opt}))$ under the uniform distribution. The low-degree is based on the Fourier expansion on the Boolean cube. Although computationally efficient, this algorithm has limited practical applications due to its distributional restrictions — uniform (and known) distribution is unrealistic in many applications. Furst et al. (1991) relaxed such a distributional restriction by adopting a low-degree algorithm for learning AC^0 functions under the product probability distributions.

Fourier Expansion: The Fourier expansion has been used to analyze Boolean functions (Wolf, 2008; O'Donnell, 2014) with a wide range of applications, namely computational learning (Linial et al., 1993; Mossel et al., 2004), noise sensitivity (O'Donnell, 2014; Kalai, 2005; Li and Médard, 2018; Heidari et al., 2019), approximation (Blais et al., 2010), and other information-theoretic scenarios (Courtade and Kumar, 2014; Weinberger and Shayevitz, 2017, 2018). Such works mainly consider Boolean functions with the uniform distribution on the inputs. In this paper, we extend the applications of Fourier estimation to stochastic mappings on the Boolean cube with arbitrary distributions.

2. Problem Formulation and Main Results

We start with the usual formulation of *agnostic* PAC learning model (Valiant, 1984; Kearns et al., 1994). The focus of this paper is on binary classification with the 0-1 loss function. Therefore, each samples consists of a feature vector $\mathbf{x}\in\mathcal{X}$ of length d with a binary label $y\in\{-1,1\}$. Typically, we consider binary features $\mathcal{X}=\{-1,1\}^d$. Available is a set of n labeled samples $\mathcal{S}_n=\{(\mathbf{x}(i),y(i))\}_{i=1}^n$ generated independent and identically distributed (i.i.d.) from an unknown but fixed probability distribution D on $\mathcal{X}\times\{-1,1\}$. The expected (generalization) loss of a predictor $g:\mathcal{X}\rightarrow\{-1,1\}$, is given by $\mathbb{P}_D\big\{Y\neq g(\mathbf{X})\big\}$.

Agnostic PAC: As a benchmark, a hypothesis class, denoted by $\mathcal{H}$, is a set of potential predictors $h:\mathcal{X}\rightarrow\{-1,1\}$. In agnostic PAC, we seek a learning algorithm that given the training set $\mathcal{S}_n$, outputs a predictor $g:\mathcal{X}\rightarrow\{-1,1\}$ whose expected loss is close to that of the best predictor in $\mathcal{H}$. More precisely, an algorithm agnostically PAC learns $\mathcal{H}$, if, for any $\epsilon,\delta\in(0,1)$, and given $n>n(\epsilon,\delta)$ training samples drawn from any distribution D, it outputs with probability $(1-\delta)$ a predictor g whose expected loss is bounded by

$$\mathbb{P}_D\Big\{Y\neq g(\mathbf{X})\Big\}\leq P_{opt}+\epsilon,$$

where P_{opt} is the minimum expected loss among all predictors in $\mathcal{H}$.

k-junta class: The main hypothesis classes considered in this paper are called k-juntas. A k-junta is a function $h:\mathcal{X}\rightarrow\{-1,1\}$ with d input variables whose output depends on at most k out of d inputs.

Notation: For compactness, we present a few notations. For any natural number d, the set $\{1, 2, \cdots, d\}$ is denoted by $[d]$. For a pair of functions f, g on $\mathcal{X}$, the notation $f \equiv g$ means that $f(x) = g(x)$ for all $x \in \mathcal{X}$.

2.1 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: Training samples $\mathcal{S}_n = \{(\mathbf{x}(i), y(i))\}_{i=1}^{n}$, depth parameter $k$.
Find a polynomial $\hat{p}$ of degree up to $k$ that minimizes $\frac{1}{n}\sum_i \big(y(i) - p(\mathbf{x}(i))\big)^2$.
Find $\theta \in [-1, 1]$ such that the empirical error of $\mathrm{sign}[\hat{p}(\mathbf{x}) - \theta]$ is minimized.
return $\hat{g} \equiv \mathrm{sign}[\hat{p} - \theta]$.
```

Next, we use the $\mathcal{L}_2$-polynomial regression algorithm for learning with respect to k juntas. The idea is to fix $\mathcal{J}$ and run Algorithm 1 on the training set with inputs only from $\mathcal{J}$. Then we repeat this experiment for all k-element subsets $\mathcal{J}$ and find the best one. With this approach, we establish the following theorem.

Theorem 1 *The $\mathcal{L}_2$-polynomial regression as in Algorithm 1, with a degree limit of $k \leq d$, agnostically PAC learns k-juntas . More precisely, given $\delta \in [0, 1]$, with probability $(1 - \delta)$ the generalization loss of this algorithm does not exceed the following*

$$\mathrm{P}_{opt} + O\Bigg(\sqrt{\frac{(2^k + k\log d)}{n}\log\frac{n}{(2^k + k\log d)}}\Bigg) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}},$$

where n is the number of samples.

By simplifying the above expression, we get a sample complexity bound of $n(\delta, \epsilon) = O(\frac{k2^k}{\epsilon^2}\log\frac{d}{\epsilon^2\delta})$.

2.2 Fourier-Based Learning Algorithm

We present another variant of $\mathcal{L}_2$ polynomial regression, known as the low-degree (Fourier) algorithm (Linial et al., 1993). The low-degree algorithm was originally designed for uniform distribution on the Boolean cube. Later it was generalized to non-uniform but product distributions on $\{-1, 1\}^d$ (Furst et al., 1991). In this paper, we further generalize this algorithm for non-product distributions.

The main idea behind the low-degree algorithm is to estimate the polynomial p^* that minimizes the square loss under the true distribution. This method is based on the Fourier expansion on the Boolean cube (O'Donnell, 2014; Wolf, 2008) and is summarized in the following.

```
Algorithm 2: Generalized Fourier
  Input: Training samples $\{(\mathbf{x}(i), y(i))\}_{i=1}^{n}$, depth parameter $k$.
  Output: Predictor $\hat{g}$
  Run FourierOrth, as in Procedure 1, to obtain the empirical basis $\widehat{\psi}_{\mathcal{S}}, \mathcal{S} \subseteq [d]$ with
    $|\mathcal{S}| \leq k$.
  Function FourierPredict($\{\widehat{\psi}_{\mathcal{S}}\}, k$):
      Compute the empirical Fourier coefficients as $a_{\mathcal{S}} = \frac{1}{n}\sum_{i=1}^{n} y(i)\widehat{\psi}_{\mathcal{S}}(\mathbf{x}(i))$.
      Construct the function $\hat{\Pi}_Y$ as $\hat{\Pi}_Y(\mathbf{x}) \triangleq \sum_{\mathcal{S}\subseteq\mathcal{J}} a_{\mathcal{S}}\widehat{\psi}_{\mathcal{S}}(\mathbf{x})$.
  return $\hat{g} \equiv \mathrm{sign}[\hat{\Pi}_Y^{\hat{\mathcal{J}}}]$.
```

Fourier expansion on the Boolean cube: Consider the vector space of all functions $f : \{-1, 1\}^d \to \mathbb{R}$ with the inner product $\langle f, g\rangle = \mathbb{E}_D[f(\mathbf{X})g(\mathbf{X})]$, where D is the underlying distribution. It is known that when D is uniform, the *monomials* form an orthonormal basis. Such monomials are defined as $\chi_{\mathcal{S}}(\mathbf{x}) = \prod_{j\in\mathcal{S}} x_j, \mathbf{x} \in \{-1, 1\}^d$, for all input subsets $\mathcal{S} \subseteq [d]$. For a general D, the monomials are not necessarily orthogonal. In that case, one needs to find a different basis denoted as $\psi_{\mathcal{S}}, \mathcal{S} \subseteq [d]$. Having this basis, any bounded function in this spaces is decomposed as

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

where $f_{\mathcal{S}}$'s are called the Fourier coefficients and calculated as $f_{\mathcal{S}} \triangleq \langle f, \psi_{\mathcal{S}}\rangle$.

The main challenge in this approach is to find the orthonormal basis for an arbitrary distribution D. For that, we adopt a Gram-Schmidt process to make the monomials $\chi_{\mathcal{S}}$ orthogonal (Heidari et al., 2021a). The orthognolaziation process is described as in the following.

Orthogonalization: Fix the following ordering for subsets of $[d]$:

$$\emptyset, \{1\}, \{2\}, \{1, 2\}, \{3\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}, \cdots, \{1, 2, ..., d\}. \tag{1}$$

We apply the Gram-Schmidt process on $\chi_{\mathcal{S}_j}$ with the above ordering. The first element of the basis is trivially given by $\psi_{\emptyset} \equiv 1$. The jth basis function corresponding to $\mathcal{S}_j$ is obtained from the following operation:

$$\tilde{\psi}_{\mathcal{S}_j} \equiv \chi_{\mathcal{S}_j} - \sum_{\ell=1}^{j-1} \langle \psi_{\mathcal{S}_\ell}, \chi_{\mathcal{S}_j}\rangle_D\ \psi_{\mathcal{S}_\ell}, \qquad \psi_{\mathcal{S}_j} \equiv \begin{cases} \frac{\tilde{\psi}_{\mathcal{S}_j}}{\|\tilde{\psi}_{\mathcal{S}_j}\|_{2,D}} & \text{if } \|\tilde{\psi}_{\mathcal{S}_j}\|_{2,D} > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

where $\|\tilde{\psi}_{\mathcal{S}_j}\|_{2,D} = \sqrt{\langle \tilde{\psi}_{\mathcal{S}_j}, \tilde{\psi}_{\mathcal{S}_j}\rangle_D}$. As a result of this process, we obtain an orthonormal basis. Therefore, we can write the Fourier decomposition of the optimal polynomial p^*. For that, we have the following statement which is a consequence of the Parseval's identity connecting the square loss to the Fourier coefficients. The proof is given in Appendix A.

Proposition 3 *Let $\psi_{\mathcal{S}}, \mathcal{S} \subseteq [d]$ be a orthonormal basis form from the process in (2). Then, the optimal degree k polynomial p^* achieving the least-square error admits the following Fourier decomposition*

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

We note that the ordering of the subsets in (1) is important to obtain the above statement. As a result, the main idea behind the Fourier algorithm is to estimate p^* by estimating the Fourier coefficients associated with the true labeling, that is $\langle Y, \psi_{\mathcal{S}} \rangle$ for all $\mathcal{S}$ with $|\mathcal{S}| \leq k$. This algorithm is presented in Algorithm 2.

In practice when the distribution D is unknown, we apply the orthogonalization process on the empirical distribution $\hat{D}$ derived using the training samples. This process can be implemented using recursive formulas as described in Procedure 1. Once the orthogonal basis is obtained, the Fourier algorithm computes the empirical estimate of $\langle Y, \psi_{\mathcal{S}} \rangle_D$. In the following theorem, we provide PAC bound when using this approach for learning k-juntas.

Theorem 2 *The Fourier algorithm algorithm agnostically learns k juntas with generalization error less than*

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

with probability at least $(1-\delta)$*, where* n *is the number of samples and* $c_k \triangleq \max_{\mathcal{S}\subseteq[d], |\mathcal{S}|\leq k} \|\psi_{\mathcal{S}}\|_\infty^2$.

Procedure 1: FourierOrth

```
Input: Samples $\{\mathbf{x}(i)\}_{i=1}^n$, depth parameter $k$, threshold $\varepsilon \in (0,1)$
Procedure FourierOrth:
    Generate all subsets $\mathcal{S}_j \subseteq [d]$ with size at most $k$ and with the standard ordering as in (1).
    Compute the covariance matrix $\hat{\mathbf{B}}$ with elements: $\hat{b}_{j,\ell} \leftarrow \frac{1}{n}\sum_{l=1}^n \mathbf{x}^{\mathcal{S}_j}\mathbf{x}^{\mathcal{S}_\ell}$.
    Set $\hat{\mathbf{A}} \leftarrow \hat{\mathbf{B}}$.
    for row $j$ of $\hat{\mathbf{A}}$ do
        Update the $j$th row: $\hat{\mathbf{A}}_{j,*} \leftarrow \hat{\mathbf{A}}_{j,*} - \sum_{\ell<j} \hat{a}_{\ell,j} \hat{\mathbf{A}}_{\ell,*}$
        Compute $\nu_j \leftarrow \hat{b}_{j,j} - \sum_{\ell<j} \hat{a}_{\ell,j}^2$
        if $\nu_j \geq \varepsilon$ then
            Normalize the $j$th row: $\hat{\mathbf{A}}_{j,*} \leftarrow \frac{\hat{\mathbf{A}}_{j,*}}{\sqrt{\nu_j}}$
            Set $\psi_{\mathcal{S}_j} \equiv \frac{1}{\sqrt{\nu_j}}(\chi_{\mathcal{S}_j} - \sum_{\ell<j} \hat{a}_{\ell,j}\, \psi_{\mathcal{S}_\ell})$
        else
            Set the $j$th row of $\hat{\mathbf{A}}$ zero: $\hat{\mathbf{A}}_{j,*} \leftarrow \mathbf{0}$
            Set $\psi_{\mathcal{S}_j} \equiv 0$
    return all $\psi_{\mathcal{S}_j}$'s.
```

3. Theoretical Analysis

The main results of this paper rely on a fundamental connection between square loss and the 0-1 loss presented as Lemma 1 and 2 in Section 3.2. In this section, we present this connection and describe the steps in proving Theorem 1 and 2. We start with presenting the Hilbert space framework.

3.1 A Hilbert Space Representation

We define a joint Hilbert space for functions on $\mathcal{X} \times \{-1, 1\}$ incorporating the joint sample-label relation and the underlying joint distribution. Let D be a joint probability distribution on the input-output set $\mathcal{X} \times \{-1, 1\}$. For analytic tractability assume that the marginal $D_{\mathbf{x}}$ of D has finite moments. Consider a Hilbert space of all real-valued functions $f : \mathcal{X} \times \{-1, 1\} \mapsto \mathbb{R}$ for which $\mathbb{E}_D[f(\mathbf{X}, Y)^2]$ is finite. The inner product between two members f, g is defined as $\langle f, g \rangle_D \triangleq \mathbb{E}_D[f(\mathbf{X}, Y)g(\mathbf{X}, Y)]$[1]. The usual 1-norm and 2-norm of a function f are defined as $\|f\|_{1,D} \triangleq \mathbb{E}_D[|f(\mathbf{X}, Y)|]$ and $\|f\|_{2,D} \triangleq \sqrt{\mathbb{E}_D[f(\mathbf{X}, Y)^2]}$, respectively.

We use this formulation to study our PAC learning problem. In this view, the true labeling is the mapping $(\mathbf{x}, y) \mapsto y$ for all $\mathbf{x} \in \mathcal{X}$ and $y \in \{-1, 1\}$. Therefore, it is not difficult to check that the generalization loss of a predictor g can be written in terms of the inner products as

$$\mathbb{P}_D\Big\{Y \neq g(\mathbf{X})\Big\} = \frac{1}{2} - \frac{1}{2}\langle Y, g \rangle_D = \frac{1}{4}\|Y - g\|_{2,D}^2, \tag{3}$$

where, with slight abuse of notation, Y and g are understood as the mappings $(\mathbf{x}, y) \mapsto y$ and $(\mathbf{x}, y) \mapsto g(\mathbf{x})$, respectively.

The empirical variant of the above formulation is easily derived by replacing D with the empirical distribution $\hat{D}$ that is uniform on the training set and zero outside of it. Therefore, the empirical loss of g immediately equals to

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

In what follows, we use this formulation to derive bounds on the expected and empirical loss and prove the main theorems.

3.2 Main Technical Results

We start with characterizing the minimum loss in the class of k-juntas.

Characterizing P_{opt}: Let $\mathcal{J}$ be a subset of $[d]$ with k elements. Consider all polynomials of degree up to k on the coordinates of $\mathcal{J}$ as the input variables. The polynomial that minimizes the square loss is defined as the projection of Y onto the subset $\mathcal{J}$. This polynomial is formally defined as

$$\Pi_Y^{\mathcal{J}} \triangleq \arg\min_{p \in \mathcal{P}_k} \|Y - p(X^{\mathcal{J}})\|_{2,D} \tag{4}$$

where $\mathcal{P}_k$ is the set of polynomials of degree at most k. With this definition, in the following lemma, we characterize the minimum generalization loss obtained by the k-juntas.

Lemma 1 *Let* P_{opt} *be the minimum expected 0-1 loss under any distribution* D *and among all the* k*-juntas for a fixed* $k \leq d$*. Then, the following equality holds*

$$\mathrm{P}_{opt} = \frac{1}{2} - \frac{1}{2}\max_{\mathcal{J} \subseteq [d],\, |\mathcal{J}| = k} \|\Pi_Y^{\mathcal{J}}\|_{1,D}. \tag{5}$$

[1] The zero function in this space is a function that maps all $(\mathbf{x}, y) \in \mathcal{X} \times \{-1, 1\}$ to 0 except a zero-probability subset. Therefore, a pair of functions f, g in this space are equal if $f - g$ is the zero function.

Proof Fix a k-element subset $\mathcal{J} \subseteq [d]$ and let $g : \mathcal{X} \mapsto \{-1, 1\}$ be a k-junta that depends on the coordinates of $\mathcal{J}$. From (3) its generalization loss can be written as $\mathbb{P}\Big\{Y \neq g(\mathbf{X})\Big\} = \frac{1}{2} - \frac{1}{2}\langle Y, g\rangle_D$. Next, consider the collection of all bounded real-valued functions whose output depends only on the coordinate $\mathcal{J}$. It is not difficult to check that this collection forms a subspace inside the original Hilbert space. Therefore, g belongs to this subspace. Further, as $\Pi_Y^{\mathcal{J}}$ minimizes the $\mathcal{L}_2$-distance with Y, it is the *projection* of Y onto this space. This implies that inner product $\langle Y, g\rangle_D$ equals to $\langle \Pi_Y^{\mathcal{J}}, g\rangle_D$. Hence, the generalization loss of g satisfies the following

$$\begin{aligned}\mathbb{P}\Big\{Y \neq g(\mathbf{X})\Big\} &= \frac{1}{2} - \frac{1}{2}\langle \Pi_Y^{\mathcal{J}}, g\rangle_D \geq \frac{1}{2} - \frac{1}{2}\langle |\Pi_Y^{\mathcal{J}}|, |g|\rangle_D \\ &\geq \frac{1}{2} - \frac{1}{2}\|\Pi_Y^{\mathcal{J}}\|_{1,D},\end{aligned}$$

where the last inequality follows as $|g(\mathbf{x})| = 1$. Therefore, maximizing over all k-element subsets $\mathcal{J}$ gives the bound

$$P_{opt} \geq \frac{1}{2} - \frac{1}{2}\max_{\mathcal{J}:|\mathcal{J}|=k} \|\Pi_Y^{\mathcal{J}}\|_{1,D}.$$

Hence, we established a lower-bound on P_{opt}. Next, we show that this bound is achievable. For that construct a k-junta as $g = \text{sign}[\Pi_Y^{\mathcal{J}}]$ where $\mathcal{J}$ is a k-element subset of $[d]$. Then, from the above argument, the generalization error of such g equals

$$\begin{aligned}\mathbb{P}\Big\{Y \neq \text{sign}[\Pi_Y^{\mathcal{J}}]\Big\} &= \frac{1}{2} - \frac{1}{2}\langle \Pi_Y^{\mathcal{J}}, \text{sign}[\Pi_Y^{\mathcal{J}}]\rangle_D \\ &= \frac{1}{2} - \frac{1}{2}\|\Pi_Y^{\mathcal{J}}\|_{1,D},\end{aligned}$$

where the last equality follows due to the identity $\langle h, \text{sign}[h]\rangle = \|h\|_1$ for any function h. Therefore, optimizing over $\mathcal{J}$ gives

$$P_{opt} \leq \frac{1}{2} - \frac{1}{2}\max_{\mathcal{J}:|\mathcal{J}|\leq k} \|\Pi_Y^{\mathcal{J}}\|_{1,D}.$$

As a result,we established an upper-bound that matches with the lower-bound and the completes the proof. ■

Connection to Square Loss : The main ingredient in the proof of the main results (Theorem 1 and 2) is the following lemma connecting the square error to the $0 - 1$ loss in learning k-juntas.

Lemma 2 *Let D be any joint probability distribution on $\mathcal{X} \times \{-1, 1\}$. Given a subset $\mathcal{J} \subseteq [d]$, let $h_{\mathcal{J}}$ denote an arbitrary bounded real-valued function on $\mathcal{X}$ that depends only on the coordinates of $\mathcal{J}$. Then,*

$$\mathbb{P}\Big\{Y \neq \text{sign}[h_{\mathcal{J}}(\mathbf{X})]\Big\} \leq \frac{1}{2}\Big(1 - \|\Pi_Y^{\mathcal{J}}\|_{1,D} + U\big(\|\Pi_Y^{\mathcal{J}} - h_{\mathcal{J}}\|_{2,D}\big)\Big),$$

where $\Pi_Y^{\mathcal{J}}$ is as in (4) and U is a polynomial defined as $U(x) = x^3 + \frac{3}{2}x^2 + \frac{5}{4}x$.

The proof of the lemma is given in Appendix B.

3.3 PAC learnability of $\mathcal{L}_2$ Polynomial Regression

Having the above lemmas, we are ready to prove Theorem 1 on agnostic PAC learnability of k-juntas using $\mathcal{L}_2$ polynomial regression.

Proof of Theorem 1. For any $\mathcal{J}$, let $\hat{p}_{\mathcal{J}}$ be the output of the polynomial regression, that is $\hat{p}_{\mathcal{J}} = \arg\min_{p\in\mathcal{P}_k} \|Y - p_{\mathcal{J}}\|_{2,\hat{D}}$, where $\hat{D}$ is the empirical distribution. Note that the selected predictor is of the form $\text{sign}[\hat{p}_{\mathcal{J}}]$. As a result, from Lemma 2 with D replaced with $\hat{D}$, the empirical loss of $\text{sign}[\hat{p}_{\mathcal{J}}]$ is bounded as

$$\mathbb{P}_{\hat{D}}\Big\{Y \neq \text{sign}[\hat{p}_{\mathcal{J}}(\mathbf{X})]\Big\} \leq \frac{1}{2} - \frac{1}{2}\|\hat{p}_{\mathcal{J}}\|_{1,\hat{D}},$$

where the $U(\cdot)$ term in Lemma 2 is zero, as $\hat{p}_{\mathcal{J}} = \Pi_Y^{\mathcal{J}}$ when D is replaced with $\hat{D}$. Next, we minimize both sides over all k-element subsets $\mathcal{J}$. From Lemma 1, with D replaced by $\hat{D}$, the right-hand side of the above inequality minimized over $\mathcal{J}$ is the minimum empirical loss $\widehat{P}_{opt}$. This implies that

$$\min_{\mathcal{J}:|\mathcal{J}|=k} \mathbb{P}_{\hat{D}}\Big\{Y \neq \text{sign}[\hat{p}_{\mathcal{J}}(\mathbf{X})]\Big\} = \widehat{P}_{opt}.$$

Hence, we proved that the minimum empirical loss is achieved using the $\mathcal{L}_2$ polynomial regression. Naturally, the next step is to extend this result to the generalization loss. This part follows from the standard arguments in VC theory (See Corollary 3.19 in (Mohri et al., 2018)) and the fact that the VC dimension of the k-junta class is less than $2^k + O(k\log d)$. Particularly, given $\delta \in (0,1)$, with probability $(1-\delta)$, the generalization loss is less than

$$P_{opt} + O\Big(\sqrt{\frac{(2^k + k\log d)}{n}\log\frac{n}{(2^k + k\log d)}}\Big) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}},$$

where n is the number of the samples. With this inequality, the theorem is proved. ■

3.4 PAC learning using the Fourier algorithm

In this part, we use the lemmas in the previous section to prove Theorem 2. Recall that $\hat{\Pi}_{\mathcal{J}}$ is the Fourier polynomial constructed at Algorithm 2 for any coordinate subset $\mathcal{J}$.

Proof of Theorem 2. From Lemma 2, we have that

$$\mathbb{P}\Big\{Y \neq \text{sign}[\hat{\Pi}_{\mathcal{J}}]\Big\} \leq \frac{1}{2}(1 - \|\Pi_Y^{\mathcal{J}}\|_{1,D}) + \frac{1}{2}U(\|\Pi_Y^{\mathcal{J}} - \hat{\Pi}_{\mathcal{J}}\|_{2,D}).$$

Next, we bound the second term in the right-hand side. For that, we present the following lemma which is proved in Appendix C.

Lemma 3 (Accuracy of the Fourier Estimations) *Given $\delta \in (0,1)$, with probability at least $(1-\delta)$, the following inequalities hold*

$$\|\Pi_Y^{\mathcal{J}} - \hat{\Pi}_Y^{\mathcal{J}}\|_{2,D} \leq O\Big(\sqrt{\frac{2^k c_k}{(k-1)!n}\log\frac{d^k}{(k-1)!\delta}}\Big), \tag{6}$$

for all k-element subsets $\mathcal{J}$, where c_k is as in Theorem 2 and n is the number of samples.

Therefore, with ϵ'_n being the right-hand side of (6) we have that

$$\begin{aligned}
\mathbb{P}\Big\{Y \neq \text{sign}[\hat{\Pi}_Y^{\mathcal{J}}]\Big\} &\le \frac{1}{2}(1 - \|\hat{\Pi}_Y^{\mathcal{J}}\|_{1,D}) + \frac{1}{2}U(\epsilon'_n) \\
&\overset{(a)}{=} \frac{1}{2}(1 - \|\Pi_Y^{\mathcal{J}}\|_{1,D}) + \frac{1}{2}\big(\|\Pi_Y^{\mathcal{J}}\|_{1,D} - \|\hat{\Pi}_Y^{\mathcal{J}}\|_{1,D}\big) + \frac{1}{2}U(\epsilon'_n) \\
&\overset{(b)}{\le} \frac{1}{2}(1 - \|\Pi_Y^{\mathcal{J}}\|_{1,D}) + \frac{\epsilon'_n}{2} + \frac{1}{2}U(\epsilon'_n),
\end{aligned} \tag{7}$$

where in (a) we add and subtract $\|\Pi_Y^{\mathcal{J}}\|_{1,D}$ and in (b) we use the following chain of inequalities

$$|\ \|\Pi_Y^{\mathcal{J}}\|_{1,D} - \|\hat{\Pi}_Y^{\mathcal{J}}\|_{1,D}| \le \|\Pi_Y^{\mathcal{J}} - \hat{\Pi}_Y^{\mathcal{J}}\|_{1,D} \le \|\Pi_Y^{\mathcal{J}} - \hat{\Pi}_Y^{\mathcal{J}}\|_{2,D} \le \epsilon'_n,$$

where the first inequality follows from the triangle inequality; the second is due to Jensen Inequality implying that $\mathbb{E}[|Z|] \le \sqrt{\mathbb{E}[Z^2]}$; and the third follows from the definition of ϵ'_n.

Next, we minimize both sides of (7) over all k-elements coordinates $\mathcal{J}$. Let $\mathcal{J}^*$ be the minimizer of the left-hand side. Hence, from Lemma 1, we obtain that

$$\mathbb{P}\Big\{Y \neq \text{sign}[\hat{\Pi}_Y^{\mathcal{J}^*}]\Big\} \le P_{opt} + \frac{\epsilon'_n}{2} + \frac{1}{2}U(\epsilon'_n) \le P_{opt} + \frac{19}{8}\epsilon'_n,$$

where the last inequality is because $U(x) \le \frac{15}{4}x$ for $x \in [0,1]$.

Recall that $\hat{\Pi}_Y^{\hat{\mathcal{J}}}$ is the polynomial on $\hat{\mathcal{J}}$ coordinates that is selected at Algorithm 2. Therefore, it remains to show that the loss on the left hand-side does not change significantly when $\mathcal{J}^*$ is replaced with $\hat{\mathcal{J}}$. As $\hat{\mathcal{J}}$ is the k-element coordinate that minimizes the empirical loss, we have that

$$\begin{aligned}
\mathbb{P}_{\hat{D}}\Big\{Y \neq \text{sign}[\hat{\Pi}_Y^{\hat{\mathcal{J}}}]\Big\} &\le \mathbb{P}_{\hat{D}}\Big\{Y \neq \text{sign}[\hat{\Pi}_Y^{\mathcal{J}^*}]\Big\} \\
&\le \mathbb{P}_{D}\Big\{Y \neq \text{sign}[\hat{\Pi}_Y^{\mathcal{J}^*}]\Big\} + O\Big(\sqrt{\frac{k}{n}\log\frac{2}{\delta}}\Big),
\end{aligned}$$

where the last inequality holds with probability $(1-\delta)$ that follows from McDiarmid's inequality and the fact that there are at most 2^k Boolean functions on coordinate $\mathcal{J}^*$. Hence, we proved that

$$\mathbb{P}_{\hat{D}}\Big\{Y \neq \text{sign}[\hat{\Pi}_Y^{\hat{\mathcal{J}}}]\Big\} \le P_{opt} + \frac{19}{8}\epsilon'_n + O\Big(\sqrt{\frac{k}{n}\log\frac{2}{\delta}}\Big).$$

The rest of the argument follows from VC theory for replacing $\hat{D}$ with D in the left-hand side . Hence, the proof is complete by replacing the ϵ'_n with the right-hand side of (6). ■

4. Learning Other Hypothesis Classes

In this section, we study learning of more general concept classes using the $\mathcal{L}_2$-based approaches.

4.1 Polynomially Approximated Class

An important concept class is the set of predictors that are approximated by fixed-degree polynomials as studied in (Kalai et al., 2008). We adopt Algorithm 1 for learning with respect to these classes. We start with the following formulation:

(ϵ, k)**-approximated concept class:** 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

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

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

As a byproduct of our analysis, we prove in Appendix D that the $\mathcal{L}_2$ based algorithms (without the randomized rounding of (Jackson, 2006)) learn the approximated concept class with error up to $2P_{opt} + \epsilon$.

Theorem 4 *Given $\epsilon > 0$ and $k \in \mathbb{N}$, the degree k $\mathcal{L}_2$ polynomial regression learns any (ϵ, k)-approximated concept class, with probability greater than $(1-\delta)$, and generalization loss up to*

$$2\text{P}_{opt} + 3\epsilon + O\Big(\sqrt{\frac{2\ d^{k+1}}{n}\log\frac{en}{d^{k+1}}}\Big) + \sqrt{\frac{1}{2n}\log\frac{1}{\delta}}$$

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

This result is derived using the following lemma proved in Appendix E that eliminates the need for randomized rounding.

Lemma 4 *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*

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

4.2 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, i.e., $\inf_{h\in\mathcal{H}}\|c - h\|_{2,D} \leq \epsilon$ 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 4.1.

Theorem 5 *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* $\text{sign}[\hat{h}]$ *learns $\mathcal{C}$ with expected generalization error up to*

$$2\min_{c\in\mathcal{C}} \mathbb{P}\Big\{Y \neq c(\mathbf{X})\Big\} + 3\epsilon + O\Big(\sqrt{\frac{\text{VC}(\mathcal{C})}{n}\log\frac{n}{\text{VC}(\mathcal{C})}}\Big),$$

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

References

Jan Arpe and Elchanan Mossel. Agnostically learning juntas from random walks. June 2008.

Aharon Birnbaum and Shai S Shwartz. Learning halfspaces with the zero-one loss: time-accuracy tradeoffs. In *Advances in Neural Information Processing Systems*, pages 926–934, 2012.

Eric Blais, Ryan O'Donnell, and Karl Wimmer. Polynomial regression under arbitrary product distributions. *Machine learning*, 80(2-3):273–294, 2010.

Avrim Blum, Merrick Furst, Jeffrey Jackson, Michael Kearns, Yishay Mansour, and Steven Rudich. Weakly learning DNF and characterizing statistical query learning using fourier analysis. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing - STOC 94*. ACM Press, 1994. doi: 10.1145/195058.195147.

Nader H. Bshouty and Areej Costa. Exact learning of juntas from membership queries. In *Algorithmic Learning Theory (ALT)*, pages 115–129. Springer International Publishing, 2016. doi: 10.1007/978-3-319-46379-7_8.

Xi Chen, Rajesh Jayaram, Amit Levi, and Erik Waingarten. Learning and testing junta distributions with sub cube conditioning. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1060–1113. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/chen21b.html.

Thomas A Courtade and Gowtham R Kumar. Which Boolean functions maximize mutual information on noisy inputs? *IEEE Trans. Inf. Theory*, 60(8):4515–4525, 2014.

Anindya De, Elchanan Mossel, and Joe Neeman. Junta correlation is testable. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, November 2019.

Ilias Diakonikolas, Themis Gouleakis, and Christos Tzamos. Distribution-independent pac learning of halfspaces with massart noise. In *Advances in Neural Information Processing Systems*, pages 4749–4760, 2019.

Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Sampling algorithms for ¡i¿l¡/i¿¡sub¿2¡/sub¿ regression and applications. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm*, SODA '06, page 1127–1136, USA, 2006. Society for Industrial and Applied Mathematics. ISBN 0898716055.

Petros Drineas, Michael W. Mahoney, S. Muthukrishnan, and Tamás Sarlós. Faster least squares approximation. *Numerische Mathematik*, 117(2):219–249, oct 2010. doi: 10.1007/s00211-010-0331-6.

E. Fischer, G. Kindler, D. Ron, S. Safra, and A. Samorodnitsky. Testing juntas. *Journal of Computer and System Sciences*, 68:753–787, 2004.

Merrick L Furst, Jeffrey C Jackson, and Sean W Smith. Improved learning of AC^0 functions. In *COLT*, volume 91, pages 317–325, 1991.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

Mohsen Heidari, S. Sandeep Pradhan, and Ramji Venkataramanan. Boolean functions with biased inputs: Approximation and noise sensitivity. In *Proc. IEEE Int. Symp. Information Theory (ISIT)*, pages 1192–1196, July 2019. doi: 10.1109/ISIT.2019.8849233.

Mohsen Heidari, Jithin Sreedharan, Gil I Shamir, and Wojciech Szpankowski. Information sufficiency via fourier expansion. In *Proc. IEEE Int. Symp. Information Theory (ISIT)*, July 2021a.

Mohsen Heidari, Jithin Sreedharan, Gil I Shamir, and Wojciech Szpankowski. Finding relevant information via a discrete fourier expansion. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4181–4191. PMLR, 18–24 Jul 2021b. URL https://proceedings.mlr.press/v139/heidari21a.html.

Vishnu Iyer, Avishay Tal, and Michael Whitmeyer. Junta distance approximation with subexponential queries. June 2021.

Jeffrey C. Jackson. Uniform-distribution learnability of noisy linear threshold functions with restricted focus of attention. In Gábor Lugosi and Hans Ulrich Simon, editors, *Learning Theory*, pages 304–318, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-35296-9.

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)*, pages 11–20, October 2005. doi: 10.1109/SFCS.2005.13.

Adam Tauman Kalai, Adam R. Klivans, Yishay Mansour, and Rocco A. Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, jan 2008. doi: 10.1137/060649057.

Gil Kalai. Noise sensitivity and chaos in social choice theory. Technical report, Hebrew University, 2005.

Michael J. Kearns, Robert E. Schapire, and Linda M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994. doi: 10.1007/bf00993468.

Adam R Klivans, Philip M Long, and Rocco A Servedio. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10(12), 2009.

Jiange Li and Muriel Médard. Boolean functions: Noise stability, non-interactive correlation, and mutual information. In *Proc. IEEE ISIT*, 2018.

Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, Fourier transform, and learnability. *J. ACM*, 40(3):607–620, 1993.

Zhengyang Liu, Xi Chen, Rocco A. Servedio, Ying Sheng, and Jinyu Xie. Distribution-free junta testing. *ACM Transactions on Algorithms*, 15(1):1–23, jan 2019. doi: 10.1145/3264434.

Mehryar (New York University) Mohri, Afshin (Google, Inc.) Rostamizadeh, Ameet (University of California, and Berkeley) Talwalkar. *Foundations of Machine Learning*. MIT Press Ltd, 2018. ISBN 0262039400.

Elchanan Mossel, Ryan O'Donnell, and Rocco P Servedio. Learning juntas. In *Proc. ACM Symp. on Theory of Computing*, pages 206–212, 2003.

Elchanan Mossel, Ryan O'Donnell, and Rocco A Servedio. Learning functions of k relevant variables. *J. Comput. Syst. Sci*, 69(3):421–434, 2004.

Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

Rocco A. Servedio, Li-Yang Tan, and John Wright. Adaptivity helps for testing juntas. In *Proceedings of the 30th Conference on Computational Complexity*, CCC '15, page 264–279, Dagstuhl, DEU, 2015. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. ISBN 9783939897811.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA, 2014. ISBN 1107057132, 9781107057135.

J.A.K. Suykens and J. Vandewalle. *Neural Processing Letters*, 9(3):293–300, 1999. doi: 10.1023/a:1018628609742.

L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, nov 1984. doi: 10.1145/1968.1972.

Santosh S. Vempala and Ying Xiao. Structure from local optima: Learning subspace juntas via higher order pca. August 2011.

Nir Weinberger and Ofer Shayevitz. On the optimal Boolean function for prediction under quadratic loss. *IEEE Trans. Inf. Theory*, 63(7):4202–4217, 2017.

Nir Weinberger and Ofer Shayevitz. Self-predicting Boolean functions. In *Proc. IEEE ISIT*, 2018.

Ronald de Wolf. *A Brief Introduction to Fourier Analysis on the Boolean Cube*. Number 1 in Graduate Surveys. Theory of Computing Library, 2008. doi: 10.4086/toc.gs.2008.001. URL http://www.theoryofcomputing.org/library.html.

Appendix A. Proof of Proposition 3

Any polynomial p of degree upto k can be written in terms of the monomials $\chi_{\mathcal{S}}$ of degree up to k. From the ordering in (1) and the orthogonalization process, these monomials are themselves decompose in terms of the basis $\psi_{\mathcal{S}}$ with $|\mathcal{S}| \leq k$. This is done be reversing the recursive equations in the orthogonalization process. Consequently, the polynomial p decomposes as

$$p \equiv \sum_{\mathcal{S}:|\mathcal{S}|\leq k} p_{\mathcal{S}} \psi_{\mathcal{S}}.$$

From the Parseval's identity, we have that

$$\begin{aligned}\|Y - p\|_{2,D}^2 &= \sum_{\mathcal{S}} (\alpha_{\mathcal{S}} - p_{\mathcal{S}})^2 \\ &= \sum_{\mathcal{S}:|\mathcal{S}|\leq k} (\alpha_{\mathcal{S}} - p_{\mathcal{S}})^2 + \sum_{\mathcal{S}:|\mathcal{S}|>k} (\alpha_{\mathcal{S}})^2\end{aligned}$$

where $\alpha_{\mathcal{S}} = \langle Y, \psi_{\mathcal{S}} \rangle_D$ are the Fourier coefficients corresponding to Y. The second equality follows since the Fourier expansion of p is zero for $\mathcal{S}$ with $|\mathcal{S}| > k$. As the above equation holds for all $p \in \mathcal{P}_k$, only the first term will be changed. Therefore, the least square error is obtained by setting $p_{\mathcal{S}} = \alpha_{\mathcal{S}}$ for all $\mathcal{S}$ with $|\mathcal{S}| \leq k$. Hence, the optimal polynomial p^* decomposes as

$$p^* \equiv \sum_{\mathcal{S}:|\mathcal{S}|\leq k} \alpha_{\mathcal{S}} \psi_{\mathcal{S}}.$$

The proof is complete by noting that $\alpha_{\mathcal{S}} = \langle Y, \psi_{\mathcal{S}} \rangle_D$.

Appendix B. Proof of Lemma 2

From (3) in the main text, the generalization error of $\text{sign}[h_{\mathcal{J}}]$ can be written as $\frac{1}{2} - \frac{1}{2}\langle Y, \text{sign}[h_{\mathcal{J}}]\rangle$. Since $\text{sign}[h_{\mathcal{J}}]$ is a function depending only on the coordinates of $\mathcal{J}$, then in the inner product, we can replace Y with $\Pi_Y^{\mathcal{J}}$. Hence, we obtain that

$$\begin{aligned}\mathbb{P}\Big\{Y \neq \text{sign}[h_{\mathcal{J}}(\mathbf{X})]\Big\} &= \frac{1}{2} - \frac{1}{2}\langle \Pi_Y^{\mathcal{J}}, \text{sign}[h_{\mathcal{J}}]\rangle \\ &= \frac{1}{4}\big(1 - \|\Pi_Y^{\mathcal{J}}\|_2^2 + \|\Pi_Y^{\mathcal{J}} - \text{sign}[h_{\mathcal{J}}]\|_2^2\big). \end{aligned} \tag{8}$$

The 2-norm quantity above is upper-bounded as follows

$$\begin{aligned}\|\Pi_Y^{\mathcal{J}} - \text{sign}[h_{\mathcal{J}}]\|_2^2 &\overset{(a)}{\leq} \Big(\|\Pi_Y^{\mathcal{J}} - h_{\mathcal{J}}\|_2 + \|h_{\mathcal{J}} - \text{sign}[h_{\mathcal{J}}]\|_2\Big)^2, \\ &= \Big(\|\Pi_Y^{\mathcal{J}} - h_{\mathcal{J}}\|_2^2 + \|h_{\mathcal{J}} - \text{sign}[h_{\mathcal{J}}]\|_2^2 + 2\|\Pi_Y^{\mathcal{J}} - h_{\mathcal{J}}\|_2\|h_{\mathcal{J}} - \text{sign}[h_{\mathcal{J}}]\|_2\Big),\end{aligned} \tag{9}$$

where (a) follows from the Minkowski's Inequality inequality for 2-norm. Note that $|h_{\mathcal{J}} - \text{sign}[h_{\mathcal{J}}]| = |1 - h_{\mathcal{J}}|$. Therefore,

$$\begin{aligned}\|h_{\mathcal{J}} - \text{sign}[h_{\mathcal{J}}]\|_2^2 &= \mathbb{E}\left[(1 - |h_{\mathcal{J}}(X^{\mathcal{J}})|)^2\right] \\ &= 1 + \|h_{\mathcal{J}}\|_2^2 - 2\|h_{\mathcal{J}}\|_1.\end{aligned} \tag{10}$$

From this relation and equations (8), (9), we obtain the following upper bound

$$4\mathbb{P}\Big\{Y \neq \text{sign}[h_{\mathcal{J}}(\mathbf{X})]\Big\} \leq 2 - 2\|h_{\mathcal{J}}\|_1 + \underbrace{\|h_{\mathcal{J}}\|_2^2 - \|\Pi_Y^{\mathcal{J}}\|_2^2}_{\text{(I)}}$$
$$+ \|\Pi_Y^{\mathcal{J}} - h_{\mathcal{J}}\|_2^2 + 2\|\Pi_Y^{\mathcal{J}} - h_{\mathcal{J}}\|_2 \underbrace{\|h_{\mathcal{J}} - \text{sign}[h_{\mathcal{J}}]\|_2}_{\text{(II)}}. \tag{11}$$

In what follows, we bound the terms denoted by (I) and (II).

Bounding (I): From the Minkowski's inequality for 2-norm, we have

$$\begin{aligned}\|h_{\mathcal{J}}\|_2^2 &\leq \Big(\|\Pi_Y^{\mathcal{J}}\|_2 + \|h_{\mathcal{J}} - \Pi_Y^{\mathcal{J}}\|_2\Big)^2 \\ &= \|\Pi_Y^{\mathcal{J}}\|_2^2 + \|h_{\mathcal{J}} - \Pi_Y^{\mathcal{J}}\|_2^2 + 2\|\Pi_Y^{\mathcal{J}}\|_2\|h_{\mathcal{J}} - \Pi_Y^{\mathcal{J}}\|_2 \\ &\leq \|\Pi_Y^{\mathcal{J}}\|_2^2 + \|h_{\mathcal{J}} - \Pi_Y^{\mathcal{J}}\|_2^2 + 2\|h_{\mathcal{J}} - \Pi_Y^{\mathcal{J}}\|_2\end{aligned}$$

where the second inequality is due Bessel's inequality implying that $\|\Pi_Y^{\mathcal{J}}\|_2 \leq 1$. Hence, the term (I) in (11) is upper bounded as

$$\text{(I)} \leq \lambda_1 \triangleq \|h_{\mathcal{J}} - \Pi_Y^{\mathcal{J}}\|_2^2 + 2\|h_{\mathcal{J}} - \Pi_Y^{\mathcal{J}}\|_2. \tag{12}$$

Bounding (II): From (10), we have

$$\begin{aligned}\|h_{\mathcal{J}} - \text{sign}[h_{\mathcal{J}}]\|_2^2 &= 1 + \|h_{\mathcal{J}}\|_2^2 - 2\|h_{\mathcal{J}}\|_1 \\ &\overset{(a)}{\leq} 1 + 2(\|\Pi_Y^{\mathcal{J}}\|_2^2 + \|\Pi_Y^{\mathcal{J}} - h_{\mathcal{J}}\|_2^2) - 2\|h_{\mathcal{J}}\|_1 \\ &\overset{(b)}{=} 1 + 2(\|\Pi_Y^{\mathcal{J}}\|_2^2 + \|\Pi_Y^{\mathcal{J}} - h_{\mathcal{J}}\|_2^2) - 2\big(\|\Pi_Y^{\mathcal{J}}\|_1 + (\|h_{\mathcal{J}}\|_1 - \|\Pi_Y^{\mathcal{J}}\|_1)\big) \\ &= 1 + 2(\|\Pi_Y^{\mathcal{J}}\|_2^2 - \|\Pi_Y^{\mathcal{J}}\|_1) + 2\|\Pi_Y^{\mathcal{J}} - h_{\mathcal{J}}\|_2^2 - 2\big(\|h_{\mathcal{J}}\|_1 - \|\Pi_Y^{\mathcal{J}}\|_1\big) \\ &\overset{(c)}{\leq} 1 + 2\|\Pi_Y^{\mathcal{J}} - h_{\mathcal{J}}\|_2^2 - 2\big(\|h_{\mathcal{J}}\|_1 - \|\Pi_Y^{\mathcal{J}}\|_1\big) \\ &\overset{(d)}{\leq} 1 + 2\|\Pi_Y^{\mathcal{J}} - h_{\mathcal{J}}\|_2^2 + 2\|\Pi_Y^{\mathcal{J}} - h_{\mathcal{J}}\|_2,\end{aligned} \tag{13}$$

where (a) follows from the Minkowski's inequality for 2-norm and the inequality $(x+y)^2 \leq 2(x^2+y^2)$. Equality (b) follows by adding and subtracting $\|\Pi_Y^{\mathcal{J}}\|_1$. Inequality (c) holds, since from Jensen's inequality $\|\Pi_Y^{\mathcal{J}}\|_2^2 \leq \|\Pi_Y^{\mathcal{J}}\|_1$. Lastly, inequality (d) holds because of the following chain of inequalities

$$\Big|\|\Pi_Y^{\mathcal{J}}\|_1 - \|h_{\mathcal{J}}\|_1\Big| \leq \|\Pi_Y^{\mathcal{J}} - h_{\mathcal{J}}\|_1 \leq \|\Pi_Y^{\mathcal{J}} - h_{\mathcal{J}}\|_2, \tag{14}$$

where the first inequality is due to the Minkowski's inequality for 1-norm and the second inequality is due to Holder's inequality.

Next, we show that the quantity $\big\|h_{\mathcal{J}} - \text{sign}[h_{\mathcal{J}}]\big\|_2$ without the square is upper bounded by the same term as in the right-hand side of (13). That is

$$\text{(II)} = \big\|h_{\mathcal{J}} - \text{sign}[h_{\mathcal{J}}]\big\|_2 \leq \lambda_2 \triangleq 1 + 2\|\Pi_Y^{\mathcal{J}} - h_{\mathcal{J}}\|_2^2 + 2\|\Pi_Y^{\mathcal{J}} - h_{\mathcal{J}}\|_2. \tag{15}$$

The argument is as follows: if $\big\|h_{\mathcal{J}} - \text{sign}[h_{\mathcal{J}}]\big\|_2$ is less than one, then the upper bound holds trivially as $\lambda_2 \geq 1$; otherwise, this quantity is less than its squared and, hence, the upper-bound holds.

As a result of the bounds in (11), (12), and (15) we obtain that

$$\begin{aligned}
4\mathbb{P}\Big\{Y \neq \text{sign}[h_{\mathcal{J}}(\mathbf{X})]\Big\} &\leq 2 - 2\|h_{\mathcal{J}}\|_1 + \lambda_1 + \|\Pi_Y^{\mathcal{J}} - h_{\mathcal{J}}\|_2^2 + 2\lambda_2\|\Pi_Y^{\mathcal{J}} - h_{\mathcal{J}}\|_2 \\
&= 2 - 2\|\Pi_Y^{\mathcal{J}}\|_1 + \Big(\|\Pi_Y^{\mathcal{J}}\|_1 - \|h_{\mathcal{J}}\|_1\Big) + \lambda_1 + \|\Pi_Y^{\mathcal{J}} - h_{\mathcal{J}}\|_2^2 + 2\lambda_2\|\Pi_Y^{\mathcal{J}} - h_{\mathcal{J}}\|_2 \\
&\leq 2 - 2\|\Pi_Y^{\mathcal{J}}\|_1 + \|\Pi_Y^{\mathcal{J}} - h_{\mathcal{J}}\|_2 + \lambda_1 + \|\Pi_Y^{\mathcal{J}} - h_{\mathcal{J}}\|_2^2 + 2\lambda_2\|\Pi_Y^{\mathcal{J}} - h_{\mathcal{J}}\|_2,
\end{aligned}$$

where the last inequality is due to (14). Therefore, from the definition of λ_1 and λ_2, and the function U in the statement of the lemma, we obtain

$$4\mathbb{P}\Big\{Y \neq \text{sign}[h_{\mathcal{J}}(\mathbf{X})]\Big\} \leq 2 - 2\|\Pi_Y^{\mathcal{J}}\|_1 + 4U(\|\Pi_Y^{\mathcal{J}} - h_{\mathcal{J}}\|_2).$$

This completes the proof.

Appendix C. Proof of Lemma 3

For any subset $\mathcal{S} \subseteq [d]$, let $\mathbf{x}^{\mathcal{S}} = \prod_{r\in\mathcal{S}} x_r$. Let $b_{j\ell} = \mathbb{E}_D[\mathbf{X}^{\mathcal{S}_j}\mathbf{X}^{\mathcal{S}_\ell}]$ and $\hat{b}_{j,\ell} = \frac{1}{n}\sum_{l=1}^{n} \mathbf{x}^{\mathcal{S}_j}\mathbf{x}^{\mathcal{S}_\ell}$ for all j, ℓ as in Procedure 1. Note that if $\hat{b}_{j,\ell} = b_{j,\ell}$ then $\psi_{\mathcal{S}} = \widehat{\psi}_{\mathcal{S}}$. Let $\bar{\Pi}_Y^{\mathcal{J}}$ denote the version of $\hat{\Pi}_Y^{\mathcal{J}}$ under this assumption. From Minkowsky's inequality, by adding and subtracting $\bar{\Pi}_Y^{\mathcal{J}}$ we have

$$\|\Pi_Y^{\mathcal{J}} - \hat{\Pi}_Y^{\mathcal{J}}\|_2 \leq \|\Pi_Y^{\mathcal{J}} - \bar{\Pi}_Y^{\mathcal{J}}\|_2 + \|\bar{\Pi}_Y^{\mathcal{J}} - \hat{\Pi}_Y^{\mathcal{J}}\|_2. \tag{16}$$

Note that the first term above measures the error in estimating the Fourier coefficients; while the second term measures the error is creating the basis. The first term is bounded as follows.

Note that $\bar{\Pi}_Y^{\mathcal{J}}$ equals to $\bar{\Pi}_Y^{\mathcal{J}}(x^d) = \sum_{\mathcal{S}\subseteq\mathcal{J}} \bar{f}_{\mathcal{S}}\psi_{\mathcal{S}}(x^d)$, where the Fourier-estimates $\bar{f}_{\mathcal{S}}$ are given as $\bar{f}_{\mathcal{S}} = \frac{1}{n}\sum_i Y(i)\psi_{\mathcal{S}}(X(i))$. In addition, by definition of the projection function $\Pi_Y^{\mathcal{J}}$, we have that $\Pi_Y^{\mathcal{J}}(\mathbf{x}) = \sum_{\mathcal{S}\subseteq\mathcal{J}} f_{\mathcal{S}}\ \psi_{\mathcal{S}}(\mathbf{x})$. Therefore, from Parseval's identity, the 2-norm factors as

$$\|\Pi_Y - \bar{\Pi}_Y^{\mathcal{J}}\|_2^2 = \sum_{\mathcal{S}\subseteq\mathcal{J}} |f_{\mathcal{S}} - \bar{f}_{\mathcal{S}}|^2.$$

In what follows, we show that $|f_{\mathcal{S}} - \bar{f}_{\mathcal{S}}| \leq \epsilon$ for all subsets $\mathcal{S} \subseteq [d]$ with $|\mathcal{S}| \leq k$. For that, we use the standard arguments using McDiarmid's inequality.

Note that $\bar{f}_{\mathcal{S}}$ is a function of the training random samples $(X(i), Y(i)), i = 1, 2, ..., n$. Observe that $\mathbb{E}[\bar{f}_{\mathcal{S}}] = f_{\mathcal{S}}$ which implies that $\bar{f}_{\mathcal{S}}$ is an unbiased estimation of $f_{\mathcal{S}}$. Since the samples are drawn i.i.d., we apply McDiarmid's inequality to bound the probability of the event $|f_{\mathcal{S}} - \bar{f}_{\mathcal{S}}| \geq \epsilon'$.

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

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

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

$$\mathbb{P}\Big\{\max_{\mathcal{S}:|\mathcal{S}|\le k}\big|\bar{f}_{\mathcal{S}}-f_{\mathcal{S}}\big|\ge\epsilon'\Big\}\le 2\Big[\sum_{m=0}^{k}\binom{d}{m}\Big]\exp\Big\{-\frac{n\epsilon'^2}{2c_k}\Big\}, \tag{17}$$

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

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

As a result, with probability at least $(1-\delta)$, $\max_{\mathcal{S}:|\mathcal{S}|\le k}\big|\bar{f}_{\mathcal{S}}-f_{\mathcal{S}}\big|\le\sqrt{\frac{2c_k}{n}\log\frac{2d^k}{(k-1)!\delta}}$. Hence, for any k-element subset $\mathcal{J}$, with probability at least $(1-\delta)$, we have that

$$\|\Pi_Y^{\mathcal{J}}-\bar{\Pi}_Y^{\mathcal{J}}\|_2^2\le 2^k\frac{2c_k}{(k-1)!n}\log\frac{2d^k}{(k-1)!\delta}. \tag{18}$$

Next, we bound the second term in (16). The argument follows from having $\hat{b}_{j,\ell}\approx b_{\ell,j}$ that itself follows again from McDiarmid's inequality. Since there are at most $O(d^k)$ such coefficients, we obtain a similar bound as in (18) for the second term. That said, combining all the bounds, we obtain the desired result

$$\|\Pi_Y^{\mathcal{J}}-\hat{\Pi}_Y^{\mathcal{J}}\|_2\le O\Bigg(\sqrt{\frac{2^kc_k}{(k-1)!n}\log\frac{d^k}{(k-1)!\delta}}\Bigg),$$

for any k-element subset $\mathcal{J}$.

Appendix D. Proof of Theorem 4

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 use Lemma 4 in Section 4.1. Consequently, from the lemma and due the fact that θ in the algorithm is selected to minimize the empirical error, we obtain that

$$\mathbb{P}_{\hat{D}}\Big\{Y\neq\hat{g}(\mathbf{X})\Big\}\le\frac{1}{2}\|Y-\hat{p}\|_{2,\hat{D}}^2, \tag{19}$$

where $\hat{p}$ is the output of $\mathcal{L}_2$-polynomial regression and $\hat{g} \equiv \text{sign}[\hat{p} - \theta]$, 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 $\|c^* - p\|_2 \leq \epsilon$. Since $\hat{p}$ minimizes the empirical 2-norm, then the right-hand side of (19) satisfies

$$\frac{1}{2}\|Y - \hat{p}\|^2_{2,\hat{D}} \leq \frac{1}{2}\|Y - p^*\|^2_{2,\hat{D}}. \tag{20}$$

We proceed by taking the expected error of the empirical error with respect to the random training samples. From (19) and (20) we obtain the following inequalities

$$\begin{aligned}\mathbb{E}\Big[\mathbb{P}_{\hat{D}}\Big\{Y \neq \hat{g}(\mathbf{X})\Big\}\Big] &\leq \frac{1}{2}\mathbb{E}\Big[\|Y - p^*\|^2_{2,\hat{D}}\Big] = \frac{1}{2}\|Y - p^*\|^2_{2,D} \\ &\overset{(a)}{\leq} \frac{1}{2}\Big(\|Y - c^*\|_{2,D} + \|p^* - c^*\|_{2,D}\Big)^2 \\ &\leq \frac{1}{2}\Big(\|Y - c^*\|_{2,D} + \epsilon\Big)^2 \\ &\overset{(b)}{\leq} \frac{1}{2}\Big(\|Y - c^*\|^2_{2,D} + 4\epsilon + \epsilon^2\Big) \\ &\overset{(c)}{\leq} 2P_{opt} + \frac{5}{2}\epsilon, \end{aligned} \tag{21}$$

where (a) holds from Minkowski's inequality for 2-norm, (b) holds as $\|Y - c^*\|_{2,D} \leq 2$, and (c) holds because of the second equality in (3) and that $P_{opt} = \mathbb{P}\{Y \neq c^*(\mathbf{X})\}$.

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 sign$[p]$ for some polynomial of degree upto k does not exceed d^{k+1}. Therefore, from VC theory (See Corollary 3.19 in (Mohri et al., 2018)) for any δ, with probability at least $(1 - \delta)$, the following inequality holds

$$\begin{aligned}\mathbb{P}\Big\{Y \neq \hat{g}(\mathbf{X})\Big\} \leq \mathbb{P}_{\hat{D}}\Big\{Y \neq \hat{g}(\mathbf{X})\Big\} &+ \sqrt{\frac{2\, d^{k+1}}{n}\log\frac{en}{d^{k+1}}} \\ &+ \sqrt{\frac{\log\frac{1}{\delta}}{2n}}. \end{aligned} \tag{22}$$

Therefore, the proof is complete by taking the expectation and combining it with the last bound in (21).

Appendix E. Proof of Lemma 4

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

$$\begin{aligned}&\mathbb{E}_\theta\Big[\mathbb{P}_{\hat{D}}\Big\{Y \neq \text{sign}[p(\mathbf{X}) - \theta]\Big\}\Big] \\ &\qquad = \frac{1}{n}\sum_i \mathbb{E}_\theta\Big[\mathbb{1}\big\{y_i \neq \text{sign}(p(\mathbf{x}_i) - \theta)\big\}\Big] \\ &\qquad = \frac{1}{n}\sum_i \underbrace{\mathbb{P}\Big\{\theta \in [p(\mathbf{x}_i), y_i] \bigcup [y_i, p(\mathbf{x}_i)]\Big\}}_{\mathbb{P}_i}. \end{aligned} \tag{23}$$

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

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

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

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

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