Fourier-Based Universal Learning

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

Gil I. Shamir GSHAMIR@IEEE.ORG
Google Inc., Pittsburgh, PA, USA

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

Editor:

Abstract

We develop a Fourier-based approach for classification. We use the approach for the feature selection problem, which we study and analyze from a PAC learning perspective for non-uniform (product) probability distributions in deterministic and stochastic settings. In this problem, given the training instances and a parameter k, the objective is to select the best k out of d features along with a k-variable predictor that yield the minimum misclassification probability. We formulate this problem as an optimization in the Fourier domain to characterize the optimal features and the predictor. Our algorithm can be viewed as a Fourier-based implementation of L_1 polynomial regression — an L_1 counterpart of the well-known low-degree algorithm. We show that, under statistical independence of the features, our algorithm agnostically learns with respect to the class of k-variable predictors (a.k.a k-juntas) and outperforms other PAC learning approaches in terms of sample complexity and computational complexity. In addition, our results can be used in other fundamental problems under non-uniform distributions, such as learning Boolean k-junta and linear-threshold functions.

Keywords: Agnostic PAC Learning, Fourier Expansion, Polynomial Regression, Feature Selection Binary Classification

1. Introduction

Feature selection is critical to the design of learning systems impacting their performance and complexity. In the supervised learning paradigm, studied in this paper, good feature selection can reduce the training and utilization running time, as well as improve model interpretability (Guyon and Elisseeff, 2003). The objective of feature selection can be stated as finding a set of features (say k out of d) so that the prediction accuracy remains relatively unchanged. For that, the main challenge is formulating a measure to evaluate the feature subsets. Such a measure needs to be computationally efficient and theoretically justified. Several measures and methods have been introduced in the literature (Kohavi and John, 1997; Koller and Sahami, 1996; Battiti, 1994; Vergara and Estévez, 2014; Yu and Liu, 2004; Peng et al., 2005; Gretton et al., 2005; Chen et al., 2017). However, in general, provable relations between these measures and the prediction accuracy remain open. In this work we study the feature selection problem as a *Probably Approximately Correct (PAC)* learning problem. The PAC framework and its agnostic version were introduced by Valiant (1984) and Kearns et al. (1994), respectively. Learning under this model has been studied extensively (Linial et al., 1993; Valiant, 2015; Goel and Klivans, 2019).

We formulate feature selection as an *agnostic PAC* learning problem. The focus of this paper is on supervised binary classification with features taking values on the Boolean hyper-cube. Our model considers learning mappings from an example, which consists of d features, to a *binary* label. The mapping should map every possible combination of feature values in an example to a label. In particular, our model observes n training instances each of which contains d features $\mathbf{x} \in \{-1,1\}^d$ with labels $y \in \{-1,1\}$. The samples are generated independent and identically distributed (i.i.d.) according to an unknown, but fixed probability distribution $P_{\mathbf{X}Y}$. The 0-1 loss function is used to measure the prediction accuracy. The expectation of this loss over $\mathsf{P}_{\mathbf{X}Y}$ is referred to as *misclassification* probability.

More precisely, in feature selection, given a parameter $k < d$, the objective is to find the set of best k features with a k-variable predictor that minimizes the misclassification probability. Hence, the set of all k-variable predictors $g : \{-1,1\}^k \mapsto \{-1,1\}$ followed by the selected features $(j_1, j_2, ..., j_k)$ are considered as the target class. The size of this set is $O(d^k 2^{2^k})$ ($O(d^k)$ for selecting a k-features combination, and $O(2^{2^k})$ for assigning a dictionary to all possible feature vectors and for all possible dictionaries). The Vapnik–Chervonenkis (VC) dimension of the class is between 2^k and $2^k + O(k \log d)$ (VC dimension of a hypothesis class is the maximum size of the input set that can be shattered (Shalev-Shwartz and Ben-David, 2014)). For this target class, the *minimum attainable misclassification probability* is defined as

$$\mathsf{P}_{opt} \triangleq \min_{j_1,j_2,..,j_k \in [d]} \min_{g:\{-1,1\}^k \mapsto \{-1,1\}} \mathbb{P}_{\mathbf{X},Y}\Big\{Y \neq g(X_{j_1}, X_{j_2}, ..., X_{j_k})\Big\}.$$

In the PAC learning framework, upon observing the training instances, a learning algorithm outputs, with probability $(1-\delta)$, a feature subset of cardinality k with a predictor so that the resulted misclassification probability is at most $\mathsf{P}_{opt} + \epsilon$, where $\epsilon, \delta \in (0,1)$. The goals are to minimize both, the number of training examples needed to achieve such probabilities, and the complexity of the algorithm.

1.1 Summary of Our Contributions and Approach

In this work, we propose a Fourier-framework to study the feature selection problem. In our earlier work (Heidari et al., 2019), we characterized P_{opt} under the deterministic labeling, i.e., $Y = f(\mathbf{X})$, and under known statistics. In this paper, we extend that result to agnostic settings with stochastic mappings. Further, we propose an agnostic-PAC feature selection and learning algorithm and derive theoretical guarantees for the case where the features are statistically independent. For the feature selection problem, our algorithm agnostically PAC-learns with sample complexity $O(c_k^2 \frac{k^2 2^{2k}}{\epsilon^2} \log \frac{d}{\delta})$ and with $O(nk(2d)^k)$ arithmetic operations (see Theorem 4). Table 1 compares our approach with well-known PAC learning algorithms adopted to the above feature selection problem. To the best of our knowledge, both the sample and computational complexities of our algorithm improve upon previously known PAC-learning algorithms. In particular, we improve the misclassification probability of the low-degree algorithm (Linial et al., 1993), 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 Subsection 1.2. Our main contributions are itemized below.

Table 1: Comparison of our algorithm with other PAC-learning approaches

Approach	Sample Cmpx.	Computational Cmpx.	Misclassification Prob.
ERM (Naive Exhaustive)	$O(\frac{k2^k}{\epsilon^2}\log\frac{d}{\delta})$	$O(nd^k 2^{2^k})$	$\mathsf{P}_{opt}+\epsilon$
$\mathcal{L}_1$-Poly. Regression (Kalai et al., 2008)	$O(d^{\Theta(k)/\epsilon})$	$O(n^2 d^{(3+\omega)3k})$	$\mathsf{P}_{opt}+\epsilon$
Low-degree Algorithm (Linial et al., 1993)	$O(2^k\log\frac{1}{\delta})$	$O(nkd^k)$	$8\mathsf{P}_{opt}$ (*uniform dist.*)
Our Approach	$O(c_k^2\frac{k^2 2^{2k}}{\epsilon^2}\log\frac{d}{\delta})$	$O(nk(2d)^k)$	$\mathsf{P}_{opt}+\epsilon$ (*product dist., stochastic labeling*)

A Fourier Framework: We extend the range of applications of *Fourier expansion on Boolean cubes*, by adapting it to the more general space of stochastic mappings (e.g., mappings from one probability space to another). Then, we develop a framework that allows us to characterize P_{opt} in the Fourier domain and find the optimal predictor and the feature subset. Note that, applying this framework, we do not require any distributional assumption on the label other than taking values from $\{-1,1\}$.

We leverage the standard *Fourier expansion on the Boolean cube*, which has been central in a range of other applications such as noise sensitivity (O'Donnell, 2014; Kalai, 2005), and information-theoretic problems (Courtade and Kumar, 2014). In this expansion, (analogue to standard Fourier series on a periodic function) any real-valued function on the Boolean cube can be expressed as a linear combination of *parities* (O'Donnell, 2014; Wolf, 2008). The Fourier coefficients quantify the levels of "nonlinearities" in a function, and this property can be leveraged to our results.

Guaranteed Universal Learning for Independent Features: We consider *Universal PAC* learning, i.e., agnostic PAC learning under some distributional (class) restrictions without knowing the actual distribution in the class. From an information-theoretic standpoint, the term *universal* indicates that a data-processing scheme achieves certain theoretical optimalities while agnostic to the statistics of the data (Cover and Thomas, 2006)[1]. That said, we restrict ourselves to agnostic PAC learning under the distributional restrictions that the features are independent. We note that, this condition can be relaxed to being almost independent (Blais et al., 2010), that is $P_{\mathbf{X}}$ is close to a product probability distribution in *total variation* distance.

Fourier-Based Learning Algorithm: We propose a feature selection algorithm with PAC learning guarantees using lower sample and computations complexities than previously known PAC learning algorithms. For that, we propose a Fourier-based implementation of $\mathcal{L}_1$ polynomial regression — that is an $\mathcal{L}_1$ counterpart of the low-degree algorithm. Hence, our algorithm achieves P_{opt} with lower computational complexity as compared to $\mathcal{L}_1$ polynomial regression. The objective of $\mathcal{L}_1$ polynomial regression is to minimize the mean absolute error (MAE) over all polynomials of a given degree, that is $\min_{p(\mathbf{x})}\mathbb{E}[|Y-p(\mathbf{X})|]$, where $P(\mathbf{x})$ is a polynomial of fixed degree (say k). In $\mathcal{L}_2$ polynomial regression, the objective is to minimize the mean square error (MSE), that is $\min_{p(\mathbf{x})}\mathbb{E}[(Y-p(\mathbf{X}))^2]$. The low-degree algorithm can be viewed as a computationally more

1. Lempel-Ziv (Ziv and Lempel, 1977) is an example of a universal algorithm.

efficient way of implementing $\mathcal{L}_2$ polynomial regression (Kalai et al., 2008). However, due to $\mathcal{L}_2$ polynomial regression, the current PAC learning guarantees are $8\mathsf{P}_{opt}$ in small-error regions. We address this issue by implementing $\mathcal{L}_1$ regression instead of $\mathcal{L}_2$. Hence, benefiting from the learning capability of $\mathcal{L}_1$ regression as well as computational efficiency of the low-degree algorithm.

Via a large deviation analysis based on Azuma's inequality for the concentration of martingales (Azuma, 1967; Szpankowski, 2011), we provide bounds on the rate of the convergence of the algorithm's misclassification probability. More precisely, we show in Theorem 5 that the expected misclassification probability of the algorithm converges to P_{opt} with rate $O(n^{-\gamma})$ for some $\gamma < 1/2$.

Other Applications. Building upon the prior concentration results (Blais et al., 2010), our algorithm can be used in other fundamental problems in computational learning, such as *agnostic* PAC learning with respect to *linear threshold* functions and the class of $\alpha(\epsilon, k)$ concentrated functions, that is a collection of functions each of which approximated by a polynomial of degree k with MSE at most ϵ. Also, our Fourier-based measure can be used in feature selection problems to evaluate the feature subsets. One can adopt conventional search algorithms (e.g., the greedy algorithm or ranking methods) with our measure for feature selection.

1.2 Related Approaches

Several approaches have been proposed in the literature for the feature selection problem with PAC guarantees. We briefly review them below.

Naive Empirical Risk Minimization (ERM): This is an exhaustive search over all feature subsets and predictors to minimize the empirical misclassification rate. For our problem, ERM is a PAC learning algorithm with sample complexity of $O(\frac{k2^k}{\epsilon^2}\log\frac{d}{\delta})$ and with computational complexity $O(nd^k 2^{2^k})$ (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.

$\mathcal{L}_1$ Polynomial Regression and SVM. 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 a (k, ϵ)-concentrated hypothesis class. Recently, Blais et al. (2010) provided some generalizations of this class. Adopting this algorithm to our problem, with an exhaustive search over feature subsets, requires a sample complexity $O(d^{\Theta(k)/\epsilon})$. With a *linear programming* implementation, the computational complexity of this algorithm is $O(n^2 d^{(3+\omega)3k})$, where $\omega < 2.4$ is the matrix-multiplication exponent. A more efficient implementation is support-vector machine (SVM) with degree-k polynomial kernel and without any regularization (Kalai et al., 2008). This implementation PAC learns in the non-agnostic setting, that is when the target labeling function itself belongs to the hypothesis class. However, this is not the case in the agnostic setting and when P_{opt} is away from zero (Blais et al., 2010).

Low Degree Algorithm. Linial et al. (1993) investigated PAC learning from an alternative perspective under a distributional restriction on $\mathbf{X}$ 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 with an error up to $8\mathsf{P}_{opt} + \epsilon$. This algorithm is based on the Fourier expansion on the Boolean hyper-cube. Although computationally efficient, this algorithm has limited practical applications due to its distributional restrictions — uniform (and known) distribution is unrealistic in many applications. Further, the factor 8 in the accuracy bound is noticeable when P_{opt} is not close to zero (say ≈ 0.1). 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. However, the second issue (factor 8) remains to be resolved.

The Fourier estimation approach in the low-degree algorithm has interesting properties which makes it suitable for other applications (Mossel et al., 2003, 2004; Jackson, 1997). Such works, mainly, consider PAC learning in *non-agnostic* setting (in which the error probability is zero or sufficiently close to zero) and the under uniform distribution. In this paper, we extend the applications of Fourier estimation to *agnostic* setting involving stochastic mappings. Therefore, we significantly relax the distributional assumptions about the features as well as the target labeling. Also, we provide a rigorous theoretical analysis guaranteeing the PAC learning capabilities of our algorithm (see Theorem 4 and 5).

Notation: The input set of the features is denoted by $\mathcal{X}$, where, unless otherwise stated, $\mathcal{X} = \{-1,1\}^d$. For shorthand, the random vector of the features is denoted by $\mathbf{X} = (X_1, X_2, ..., X_d)$. We construct the vector space of real-valued functions on $\mathcal{X}$ with inner product denoted by $\langle f, g\rangle \triangleq \mathbb{E}[f(\mathbf{X})g(\mathbf{X})]$ for any real-valued function f, g on $\mathcal{X}$. For any bounded function $f : \mathcal{X} \mapsto \mathbb{R}$ in this space, the 1-norm and 2-norm are defined as $\|f\|_1 \triangleq \mathbb{E}\big[|f(\mathbf{X})|\big]$ and $\|f\|_2 \triangleq \sqrt{\mathbb{E}\big[f(\mathbf{X})^2\big]}$, respectively. As a shorthand, in this paper, for any natural number ℓ, the set $\{1, 2, \cdots, \ell\}$ is denoted by $[\ell]$. 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 $x \in \mathcal{X}$.

2. A Fourier-Based Framework

The Fourier expansion on the Boolean cube has been a powerful tool to characterize non-linear relations among the features and the labels. Such an expansion has been developed also on the Boolean cube with non-uniform distribution (O'Donnell, 2014). In what follows we present an overview of this Fourier expansion. A more detailed discussion on the properties of this Fourier expansion is available in Appendix A.

The Fourier expansion on the Boolean cube: Let $\mathbf{X} = (X_1, X_2, ..., X_d)$ be a vector of mutually independent random variables on the Boolean cube $\{-1,1\}^d$. Let μ_j and σ_j be the mean and standard-deviation of $X_j, j \in [d]$. Suppose that these random variables are non-trivial, that is $\sigma_j > 0$ for all $j \in [d]$. The Fourier expansion is defined via a set of basis functions called *parities*. The *parity* for a subset $\mathcal{S} \subseteq [d]$ is a function $\psi_{\mathcal{S}} : \mathbb{R}^d \mapsto \mathbb{R}$ defined as

$$\psi_{\mathcal{S}}(\mathbf{x}) \triangleq \prod_{i\in\mathcal{S}} \frac{x_i - \mu_i}{\sigma_i}, \quad \forall \mathbf{x} \in \mathbb{R}^d.$$

By construction, the above parities are orthonormal. In other words, $\mathbb{E}[\psi_{\mathcal{S}}(\mathbf{X})\ \psi_{\mathcal{T}}(\mathbf{X})] = 0$ for $\mathcal{S} \neq \mathcal{T}$ and $\mathbb{E}[\psi_{\mathcal{S}}(\mathbf{X})^2] = 1$.

It is known that these parity functions form an orthonormal basis for the space of bounded functions on the Boolean cube (O'Donnell, 2014), that is when $\mathcal{X} = \{-1,1\}^d$. As a result, any bounded function $f : \{-1,1\}^d \mapsto \mathbb{R}$ can be written as a linear combination of the form

$$f(\mathbf{x}) = \sum_{\mathcal{S}\subseteq[d]} f_{\mathcal{S}}\, \psi_{\mathcal{S}}(\mathbf{x}), \quad \forall \mathbf{x} \in \{-1,1\}^d,$$

where $f_{\mathcal{S}} \in \mathbb{R}$ are called the *Fourier coefficients* of f. Due to the orthogonality of the parities, the Fourier coefficients can be computed as follows

$$f_{\mathcal{S}} = \mathbb{E}[f(\mathbf{X})\psi_{\mathcal{S}}(\mathbf{X})], \quad \forall \mathcal{S} \subseteq [d]. \tag{1}$$

As a special case, the standard Fourier expansion on the Boolean cube is obtained when X_j's are uniform random variables over $\{-1,1\}$. As a result, $\psi_{\mathcal{S}} \equiv \mathbf{x}^{\mathcal{S}}$ and $f \equiv \sum_{\mathcal{S}\subseteq[d]} f_{\mathcal{S}}\, \mathbf{x}^{\mathcal{S}}$.

Characterization of P_{opt} in the Fourier domain: Next, we characterize the misclassification probability in the Fourier domain. Consider the special case in which the label Y is generated according to an unknown, but fixed, function as $Y = f(\mathbf{X})$. A more general case in which Y is generated through $P_{Y|\mathbf{X}}$ is studied in Section 4. With the above assumption, the minimum misclassification probability becomes

$$\mathsf{P}_{opt} = \min_{\mathcal{J}\subset[d]:|\mathcal{J}|=k} \; \min_{g:\{-1,1\}^k\mapsto\{-1,1\}} \mathbb{P}_{\mathbf{X}}\{f(\mathbf{X}) \neq g(X^{\mathcal{J}})\}. \tag{2}$$

Here $\mathcal{J}$ represents the selected feature subset and g is the k-variable predictor of the label $Y = f(\mathbf{X})$. We provide an alternative representation of P_{opt} in Fourier domain and characterize the optimal predictor and subset $\mathcal{J}$. As a key ingredient in our characterization, we need to define the notion of *projection onto* $\mathcal{J}$.

Definition 1 (Projection onto a subset) *The projection of a function $f : \{-1,1\}^d \mapsto \mathbb{R}$ onto a feature subset $\mathcal{J} \subseteq [d]$ is defined as*

$$f^{\subseteq\mathcal{J}}(\mathbf{x}) \triangleq \sum_{\mathcal{S}\subseteq\mathcal{J}} \langle f, \psi_{\mathcal{S}}\rangle\, \psi_{\mathcal{S}}(\mathbf{x}), \qquad \forall \mathbf{x} \in \{-1,1\}^d.$$

As a special case, if $\mathcal{J} = [d]$, then $f^{\subseteq\mathcal{J}} \equiv f$. Using the above notion, we provide a characterization of P_{opt} in the following proposition.

Proposition 2 *The minimum attainable misclassification probability with deterministic labeling $Y = f(\mathbf{X})$, as defined in (2), equals to*

$$\mathsf{P}_{opt} = \frac{1}{2}\left[1 - \max_{\mathcal{J}\subseteq[d],\,|\mathcal{J}|=k} \|f^{\subseteq\mathcal{J}}\|_1\right]. \tag{3}$$

Further, an optimal k-variable predictor is given by $g^* = \mathsf{sign}[f^{\subseteq\mathcal{J}^*}]$, where $\mathcal{J}^*$ is an optimal feature subset that maximizes the 1-norm expression above.

Proof The proof follows from similar steps as in (Heidari et al., 2019). For completeness, we provide the proof. Fix a subset $\mathcal{J} \subseteq [d]$ and a predictor $g : \{-1,1\}^k \mapsto \{-1,1\}$. Since the range of f, g belongs to $\{-1,1\}$, we can write

$$\mathbb{P}\Big\{f(\mathbf{X}) \neq g(\mathbf{X}^{\mathcal{J}})\Big\} = \frac{1}{2} - \frac{1}{2}\mathbb{E}[f(\mathbf{X})g(\mathbf{X}^{\mathcal{J}})]. \tag{4}$$

Let $\tilde{g} : \{-1,1\}^d \mapsto \{-1,1\}$, with $\tilde{g}(\mathbf{x}) = g(\mathbf{x}^{\mathcal{J}})$ for all $\mathbf{x} \in \{-1,1\}^d$. The function $\tilde{g}$ is a representation of g in the d-dimensional space. Note that since $\tilde{g}$ depends only on the coordinates of $\mathcal{J}$, its Fourier coefficients for $\mathcal{S} \nsubseteq \mathcal{J}$ are zero. This give the Fourier expansion of the form $\tilde{g} \equiv \sum_{\mathcal{S}\subseteq\mathcal{J}} \tilde{g}_{\mathcal{S}}\psi_{\mathcal{S}}$. Therefore, the expectation in (4) can be written as

$$\begin{aligned}\mathbb{E}[f(\mathbf{X})g(\mathbf{X}^{\mathcal{J}})] = \mathbb{E}[f(\mathbf{X})\tilde{g}(\mathbf{X})] &= \sum_{\mathcal{S}\subseteq\mathcal{J}} \tilde{g}_{\mathcal{S}}\mathbb{E}[f(\mathbf{X})\psi_{\mathcal{S}}(\mathbf{X})] \\ &= \sum_{\mathcal{S}\subseteq\mathcal{J}} \tilde{g}_{\mathcal{S}} f_{\mathcal{S}} \overset{(a)}{=} \langle f^{\subseteq\mathcal{J}}, \tilde{g}\rangle \overset{(b)}{\leqslant} \langle |f^{\subseteq\mathcal{J}}|, |\tilde{g}|\rangle, \end{aligned} \tag{5}$$

where $f_{\mathcal{S}}$'s are the Fourier coefficients of f and $f^{\subseteq\mathcal{J}}$ is its projection onto $\mathcal{J}$, as in Definition 1. Equality (a) holds because of Fact 1 in Appendix A and that $f^{\subseteq\mathcal{J}} \equiv \sum_{\mathcal{S}\subseteq\mathcal{J}} f_{\mathcal{S}}\psi_{\mathcal{S}}$. Inequality (b) holds by taking the absolute value of $f^{\subseteq\mathcal{J}}$ and $\tilde{g}$. Since the range of $\tilde{g}$ is $\{-1,1\}$, then $|\tilde{g}| \equiv 1$. Therefore, $\langle |f^{\subseteq\mathcal{J}}|, |\tilde{g}|\rangle = \|f^{\subseteq\mathcal{J}}\|_1$. This together with (4) establishes the following lower bound

$$\mathsf{P}_{opt} \geqslant \frac{1}{2} - \frac{1}{2} \max_{\mathcal{J}:|\mathcal{J}|=k} \|f^{\subseteq\mathcal{J}}\|_1.$$

Next, we derive an upper bound on P_{opt} by constructing a predictor. For that fix a subset $\mathcal{J}$ and take $g \equiv \mathsf{sign}[f^{\subseteq\mathcal{J}}]$. Let $\tilde{g}$ be the representation of g in the d-dimensional space. Then for this choice,

$$\langle f^{\subseteq\mathcal{J}}, \tilde{g}\rangle = \mathbb{E}[|f^{\subseteq\mathcal{J}}(\mathbf{X}^{\mathcal{J}})|] = \|f^{\subseteq\mathcal{J}}\|_1.$$

Therefore, from (4) and the argument above, we obtain

$$\mathsf{P}_{opt} \leqslant \mathbb{P}\Big\{f(\mathbf{X}) \neq g(\mathbf{X}^{\mathcal{J}})\Big\} = \frac{1}{2} - \frac{1}{2}\|f^{\subseteq\mathcal{J}}\|_1.$$

This is an upper bound on P_{opt} for any k-element subset $\mathcal{J}$. Hence, the following is also an upper bound:

$$\mathsf{P}_{opt} \leqslant \frac{1}{2} - \frac{1}{2} \max_{\mathcal{J}:|\mathcal{J}|=k} \|f^{\subseteq\mathcal{J}}\|_1.$$

The proof is complete as the lower bound and the upper bound are matching. ∎

The optimization in (3) is over the feature-subsets of size k and, hence, the size of the search space is $\binom{d}{k}$, significantly lower than the original search space which is $O(d^k)2^{2^k}$. Once the optimal feature subset $\mathcal{J}^*$ is determined, the optimal k-variable predictor (k-junta) is obtained by taking the sign of the optimal projection, which is $\mathsf{sign}[f^{\subseteq\mathcal{J}^*}]$. Consequently, there is no need for further search in the space of k-letter functions.

Although the above formulation is characterizable only when the feature's distribution and the labeling function f are known, it gives intuitions about the structure of the optimal feature selection in the agnostic settings. Our objective is to design a feature selection method that selects the optimal feature subset ($\mathcal{J}^*$), and a learning algorithm that outputs a hypothesis close to the optimal predictor ($\mathsf{sign}[f^{\subseteq\mathcal{J}^*}]$) in the universal setting. We present our algorithm in the next section.

3. Fourier-Based Learning Algorithm

We build upon the characterization of optimal predictor (Proposition 2) and propose a Fourier-based supervised learning algorithm with an embedded feature selection (see Algorithm 1). In *agnostic* PAC learning, a learning algorithm achieves the minimum attainable misclassification probability under any feature-label distribution. In this paper, our theoretical guarantees hold under the condition that the features are independent. To emphasize this, we first present a notion of *agnostic* PAC learning which is restricted to a class of feature-label distributions.

Definition 3 ($\mathscr{P}$-Universality) *Given a class of feature-label distribution $\mathscr{P}$, a learning algorithm is said to be universal, if it agnostically learns with respect to a hypothesis class $\mathcal{H}$ under any feature distribution $P_{\mathbf{X},Y} \in \mathscr{P}$. More precisely, for every $\epsilon, \delta \in (0,1)$, every probability distribution $P_{\mathbf{X},Y} \in \mathscr{P}$, and at least $n(\epsilon,\delta)$ number of i.i.d. training samples generated by $P_{\mathbf{X},Y}$, the algorithm produces, with probability at least $(1-\delta)$, a hypothesis $g \in \mathcal{H}$ with misclassification probability at most $\mathsf{P}_{opt} + \epsilon$.*

The focus of this section is on $\mathscr{P}$-universality with $\mathscr{P}$ being the set of all $P_{\mathbf{X},Y}$ on $\{-1,1\}^d \times \{-1,1\}$ such that the marginal $P_{\mathbf{X}}$ is a product probability distribution and $Y = f(\mathbf{X})$ for some unknown function f. Further, the hypothesis class is the set of all functions g that depends on at most k inputs. In section 4 we extend our result to stochastic labeling, that is when the label is generated according to an unknown, but fixed, conditional probability distribution $P_{Y|\mathbf{X}}$.

Algorithm 1 Fourier-Based Learning

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

procedure FEATURE SELECTION

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

Compute $\mathrm{score}_1(\mathcal{J})$, as in (8), for all subsets $\mathcal{J} \subseteq [d]$ with size k.

Set $\hat{\mathcal{J}}$ as the feature subset that maximizes $\mathrm{score}_1(\mathcal{J})$.

return $\hat{\mathcal{J}}$

procedure PREDICTOR($\hat{\mathcal{J}}$)

Compute the empirical Fourier coefficients $\hat{f}_S$, as in (7), for all $\mathcal{S} \subseteq [d]$.

Construct the empirical projection function $\hat{f}^{\subseteq \hat{\mathcal{J}}}$ defined as

$$\hat{f}^{\subseteq \hat{\mathcal{J}}}(\mathbf{x}) \triangleq \sum_{\mathcal{S} \subseteq \hat{\mathcal{J}}} \hat{f}_S \prod_{j \in \mathcal{S}} \frac{x_j - \hat{\mu}_j}{\hat{\sigma}_j}.$$

Construct the predictor as $\hat{g} = \mathsf{sign}[\hat{f}^{\subseteq \hat{\mathcal{J}}}]$.

return $\hat{g}$

Recall from Proposition 2 that the optimal feature subset $\mathcal{J}^*$ maximizes the 1-norm expression $\|f^{\subseteq \mathcal{J}}\|_1$. Also, the optimal predictor g^* is the *sign* of the projection function $f^{\subseteq \mathcal{J}^*}$. That said, Algorithm 1 consists of two main processes: one for finding $\mathcal{J}^*$ and the other for estimating its projection function. In the first process, the training samples are used for estimating the 1-norm expression $\|f^{\subseteq \mathcal{J}}\|_1$ for all subsets $\mathcal{J}$ of size k. The estimation of $\|f^{\subseteq \mathcal{J}}\|_1$ is used as a measure for selecting the feature subsets $\mathcal{J}$. With that, the algorithm searches over all feature subsets with k elements and finds the one that maximizes it. Let $\hat{\mathcal{J}}$ denote the selected feature subset. In the second process, the algorithm constructs the predictor $\hat{g}$ by estimating $f^{\subseteq \hat{\mathcal{J}}}$ and taking its sign. This is summarized in Algorithm 1.

Before explaining our estimation methods, we argue that the estimations are accurate enough and the algorithm finds an asymptotically optimal feature subset. More precisely, we show in the following theorem that the algorithm is universal for memoryless features. We provide the proof of the theorem in Section 5 and appendices.

Theorem 4 *Given the parameters $k, d \in \mathbb{N}$, Algorithm 1 is a universal learning algorithm in the sense of Definition 3 for independent features and deterministic labeling. More precisely, if δ, $\epsilon \in (0,1)$ and f is the unknown labeling function, the misclassification rate of the algorithm satisfies $\mathbb{P}\{f(\mathbf{X}) \neq \hat{g}(X^{\hat{\mathcal{J}}})\} \leqslant \mathsf{P}_{opt} + \epsilon$ with probability at least $(1-\delta)$, provided that the training sample size is at least $n(\epsilon, \delta)$ with*

$$n(\epsilon, \delta) \leqslant O\Big(\frac{k^2 2^{2k} c_k^2}{\epsilon^2} \log \frac{d}{\delta}\Big), \tag{6}$$

where c_k is a constant bounded as $c_k \leqslant \big(\max_{j \in [d]}\{\frac{1+|\mu_j|}{\sigma_j}\}\big)^{2k}$.

To have a better insight on the performance of the algorithm, we also characterize the misclassification probability of the algorithm averaged over all realizations of the training samples. The asymptotic behavior of this quantity is provided in the following theorem which is proved in Appendix F.

Theorem 5 *Let $\mathcal{D}_n$ denote the training set consisting of the instances $(\mathbf{x}(i), y(i)), i = 1, 2, ..., n$. For a fixed k, the expected misclassification probability of Algorithm 1 converges to P_{opt} as n grows. More precisely, the following inequality*

$$\mathbb{E}_{\mathcal{D}_n}\big[\mathbb{P}_{\mathbf{X},Y}\{Y \neq \hat{g}(\mathbf{X}^{\hat{\mathcal{J}}})\}\big] \leqslant \mathsf{P}_{opt} + O\big(n^{-\gamma}\big)$$

holds for any $\gamma \in (0, \frac{1}{2})$.

3.1 Estimation Processes in Algorithm 1

As discussed, the optimal feature subset and the predictor are obtained by maximizing the 1-norm quantity $\|f^{\subseteq \mathcal{J}}\|_1$. Since f and the feature's distribution $P_{\mathbf{X}}$ are unknown, only an estimate of $\|f^{\subseteq \mathcal{J}}\|_1$ is possible. For that, we need to estimate $f^{\subseteq \mathcal{J}}$ and compute its empirical 1-norm. As for the estimation of the projections, we use the fact that $f^{\subseteq \mathcal{J}}$ is constructed from a collection of the Fourier coefficients of f as the summation $f^{\subseteq \mathcal{J}}(\mathbf{x}) \triangleq \sum_{\mathcal{S} \subseteq \mathcal{J}} f_{\mathcal{S}}\ \psi_{\mathcal{S}}(\mathbf{x})$. Using this structure, the estimation of $f^{\subseteq \mathcal{J}}$ is obtained by estimating the parity functions $\psi_{\mathcal{S}}$, and the Fourier coefficients $f_{\mathcal{S}}$. That said, there are three estimation processes in the algorithm which are described in the following.

Estimation of the parities. For approximation of the parity functions, first, the mean and the standard deviation of the features are estimated. Let $(\hat{\mu}_j, \hat{\sigma}_j)$ denote the empirical mean and standard deviation of the jth feature. The quantities $(\hat{\mu}_j, \hat{\sigma}_j)$ are computed using conventional estimation methods. Next, the estimation of the parity function $\psi_{\mathcal{S}}$ is given by $\widehat{\psi}_{\mathcal{S}}(\mathbf{x}) \triangleq \prod_{j \in \mathcal{S}} \frac{x_j - \hat{\mu}_j}{\hat{\sigma}_j}$.

Estimating The Projection Functions. Using the estimated parities, the empirical Fourier coefficient f_S is calculated as

$$\hat{f}_S \triangleq \frac{1}{n}\sum_{i=1}^{n} y(i)\ \widehat{\psi}_{\mathcal{S}}(\mathbf{x}(i)), \qquad \widehat{\psi}_{\mathcal{S}}(\mathbf{x}) \triangleq \prod_{j\in\mathcal{S}} \frac{x_j - \hat{\mu}_j}{\hat{\sigma}_j}, \tag{7}$$

where $(\mathbf{x}(i), y(i)) \in \mathcal{D}_n, i = 1, 2, ..., n$ are the training samples. Note that the estimated parity functions are no longer orthonormal and, hence, amount to a level of inaccuracy in the estimation of $f_{\mathcal{S}}$. Once $\hat{f}_S$ are computed, the estimation of the projection function $f^{\subseteq\mathcal{J}}$ is obtained by the equation $\hat{f}^{\subseteq\mathcal{J}}(\mathbf{x}) \triangleq \sum_{\mathcal{S}\subseteq\mathcal{J}} \hat{f}_S\ \widehat{\psi}_{\mathcal{S}}(\mathbf{x})$.

Estimating the 1-norm. When $\hat{f}^{\subseteq\mathcal{J}}$ is obtained, the next step is to approximate $\|\hat{f}^{\subseteq\mathcal{J}}\|_1$ which is needed to obtain $\hat{\mathcal{J}}$ as an approximation to $\mathcal{J}^*$. By definition, this 1-norm operation equals $\|\hat{f}^{\subseteq\mathcal{J}}\|_1 \triangleq \mathbb{E}_{\mathbf{X}}[|\hat{f}^{\subseteq\mathcal{J}}(\mathbf{X})|]$. Hence, naturally, the estimation of this quantity is obtained by the empirical averaging

$$\frac{1}{n}\sum_{i=1}^{n} \big|\hat{f}^{\subseteq\mathcal{J}}(\mathbf{x}(i))\big|.$$

Since we use the same training samples to obtain both $\hat{f}^{\subseteq\mathcal{J}}$ and its empirical 1-norm, these two quantities are correlated. Hence, the above estimation is possibly biased.

Making the Estimations Unbiased. That said, to ensure that the estimation is unbiased, we compute the estimator as follows

$$\text{score}_1(\mathcal{J}) = \widehat{\|f^{\subseteq\mathcal{J}}\|}_1 \triangleq \frac{1}{n-1}\sum_{i=1}^{n} \Bigg| \sum_{\mathcal{S}\subseteq\mathcal{J}} \hat{f}_S \widehat{\psi}_{\mathcal{S}}(\mathbf{x}(i)) - \frac{1}{n} y(i)\big(\widehat{\psi}_{\mathcal{S}}(\mathbf{x}(i))\big)^2 \Bigg|. \tag{8}$$

This correction is done by subtracting the quantity $\frac{1}{n}y(i)\big(\widehat{\psi}_{\mathcal{S}}(\mathbf{x}(i))\big)^2$. We use $\text{score}_1(\mathcal{J})$ as an estimate of $\|f^{\subseteq\mathcal{J}}\|_1$. We show in the following lemma that this estimator is asymptotically unbiased; that is $\big|\ \mathbb{E}[\text{score}_1(\mathcal{J})] - \|f^{\subseteq\mathcal{J}}\|_1\ \big| \to 0$ as $n \to \infty$. We start with the following lemma which is proved in Appendix D.

Lemma 1 *Suppose $\hat{\mu}_j = \mu_j$ and $\hat{\sigma}_j = \sigma_j$ for all $j \in [d]$. The measure $\text{score}_1(\mathcal{J}) = \widehat{\|f^{\subseteq\mathcal{J}}\|}_1$ as in (8) is an asymptotically unbiased estimate of $\|f^{\subseteq\mathcal{J}}\|_1$. More precisely*

$$\Big|\mathbb{E}\big[\ \text{score}_1(\mathcal{J})\big] - \|f^{\subseteq\mathcal{J}}\|_1\Big| \leqslant \frac{2^{k/2}}{\sqrt{n-1}}.$$

The idea behind the proof of the lemma is to rewrite score_1 as a summation of the form $\text{score}_1(\mathcal{J}) = \frac{1}{n}\sum_i |\hat{f}^{\subseteq\mathcal{J}}_{(i)}|$, where $\hat{f}^{\subseteq\mathcal{J}}_{(i)}$ is the term in the bracket in (8). These quantities are estimates of $f^{\subseteq\mathcal{J}}$ and are identically distributed random variables depending on the training instances. This is possible because of the additional term we added in (8) to make the estimate of $\|f^{\subseteq\mathcal{J}}\|_1$ unbiased. With this approach, we show that the expectation of score_1 equals to $\mathbb{E}\big[\|\hat{f}^{\subseteq\mathcal{J}}_{(1)}\|_1\big]$. Then, we relate this quantity to the square root of the MSE of estimating Fourier coefficients; that is $\sqrt{\mathbb{E}[(\hat{f}_S - f_S)^2]}$. Since $\hat{f}_S$ is the empirical average of $y(i)\widehat{\psi}_{\mathcal{S}}(\mathbf{x}(i))$ for $i \in [n]$, then the MSE is $O(1/n)$. For convenience, in the lemma we assumed there is no error in estimating feature's mean and variance. A more general version of this lemma without such assumption is Lemma 9 in Appendix B.

With the above method, one can verify that the estimation of $\|f^{\subseteq \mathcal{J}}\|_1$ for a subset $\mathcal{J}$ with $|\mathcal{J}| = k$ is computed in $O(nk2^k)$ arithmetic operations. As a result, the computational complexity of the feature selection method in our algorithm is $O(d^k nk2^k)$.

4. Extensions to Stochastic Labeling

The Fourier framework in Section 2 is developed for deterministic labeling, where $Y = f(\mathbf{X})$ for some function f. In this section, we address this restriction and extend the Fourier framework to stochastic mappings. We show that our results in Proposition 2, Theorem 4, and 5 still hold when the labels are generated according to an arbitrary unknown probability distribution $P_{Y|\mathbf{X}}$. More precisely, based on Definition 3, we consider $\mathscr{P}$-universal learning where $\mathscr{P}$ is the set of all distributions $P_{\mathbf{X},Y}$ on $\{-1,1\}^d \times \{-1,1\}$ such that $P_{\mathbf{X}}$ is a product probability distribution.

In what follows, we prove necessary statements enabling us to extend our Fourier framework to stochastic mappings. We start with generalizing our notion of projection given in Definition 1.

Definition 6 (Projection onto a subset) *Given a joint probability distribution $P_{\mathbf{X},Y}$ on $\{-1,1\}^d \times \{-1,1\}$, the projection of Y onto a subset $\mathcal{J} \subseteq [d]$ is defined as*

$$f^{\subseteq \mathcal{J}}(\mathbf{x}) \triangleq \sum_{\mathcal{S} \subseteq \mathcal{J}} \mathbb{E}[Y \psi_{\mathcal{S}}(\mathbf{X})] \psi_{\mathcal{S}}(\mathbf{x}), \qquad \forall \mathbf{x} \in \{-1,1\}^d.$$

When Y is a deterministic function of the features $\mathbf{X}$, then the above notions reduces to the one in Definition 1. In the following lemma, we show that the projection function $f^{\subseteq \mathcal{J}}$ provides a proxy to analyze the misclassification probability.

Lemma 2 *Given any subset $\mathcal{J} \subseteq [d]$, let $g : \{-1,1\}^d \mapsto \{-1,1\}$ be a function whose output depends only on the coordinates of $\mathcal{J}$. Then $\mathbb{E}[Y\, g(\mathbf{X})] = \langle f^{\subseteq \mathcal{J}}, g \rangle$, where $f^{\subseteq \mathcal{J}}$ is the projection of Y onto $\mathcal{J}$ as in Definition 6. Further, the resulted misclassification probability satisfies*

$$\mathbb{P}\Big\{Y \neq g(\mathbf{X})\Big\} = \frac{1}{2} - \frac{1}{2}\langle f^{\subseteq \mathcal{J}}, g\rangle = \frac{1}{4}\big(\|f^{\subseteq \mathcal{J}} - g\|_2^2 + 1 - \|f^{\subseteq \mathcal{J}}\|_2^2\big). \tag{9}$$

Proof Since g depends only on $\mathbf{x}^{\mathcal{J}}$, then, from Fact 3, its Fourier expansion is of the form $g \equiv \sum_{\mathcal{S} \subseteq \mathcal{J}} g_{\mathcal{S}} \psi_{\mathcal{S}}$, where $g_{\mathcal{S}}$'s are the Fourier coefficients. Using this summation we have

$$\mathbb{E}[Y g(\mathbf{X})] = \sum_{\mathcal{S} \subseteq \mathcal{J}} g_{\mathcal{S}}\, \mathbb{E}[Y \psi_{\mathcal{S}}(\mathbf{X})] = \sum_{\mathcal{S} \subseteq \mathcal{J}} \mathbb{E}[g(\mathbf{X}) \psi_{\mathcal{S}}(\mathbf{X})]\, \mathbb{E}[Y \psi_{\mathcal{S}}(\mathbf{X})] = \langle f^{\subseteq \mathcal{J}}, g\rangle. \tag{10}$$

where the second equality holds as $g_{\mathcal{S}} = \langle g, \psi_{\mathcal{S}}\rangle$. Hence, the first statement of the lemma is proved. Next, we prove the equalities in (9). Since Y and $g(\mathbf{X})$ take values from $\{-1,1\}$, then

$$\mathbb{P}\Big\{Y \neq g(\mathbf{X})\Big\} = \frac{1}{2} - \frac{1}{2}\, \mathbb{E}[Y g(\mathbf{X})].$$

Hence, with the above equation and (10), we establish the first equality in (9). Next, we prove the second equality. Form the definition of 2-norm, we have

$$\|f^{\subseteq \mathcal{J}} - g\|_2^2 = \langle (f^{\subseteq \mathcal{J}} - g), (f^{\subseteq \mathcal{J}} - g)\rangle = \|f^{\subseteq \mathcal{J}}\|_2^2 + \|g\|_2^2 - 2\,\langle f^{\subseteq \mathcal{J}}, g\rangle.$$

Since the range of g belongs to $\{-1,1\}$, then $\|g\|_2^2 = 1$. Therefore, by rewriting the above equality we have

$$\langle f^{\subseteq \mathcal{J}}, g\rangle = \frac{1}{2}\big(1 + \|f^{\subseteq \mathcal{J}}\|_2^2 - \|f^{\subseteq \mathcal{J}} - g\|_2^2\big).$$

The proof is complete as this equation implies the second equality in the statement of the lemma. ■

Using this lemma, we can easily extend our results to stochastic labeling. For instance, Proposition 2 extends to non-deterministic labeling. To see this, note that due to Lemma 2, equation (5) in the proof of the proposition still holds for stochastic Y. The rest of the proof of the proposition only depends on $f^{\subseteq \mathcal{J}}$, hence holds for stochastic Y.

5. Theoretical Analysis

In this section, we present an overview of our analysis for Algorithm 1 and give a road map to prove Theorem 4. For simplicity of presenting the proof, it is assumed that $\hat{\mu}_j = \mu_j$ and $\hat{\sigma}_j = \sigma_j, j \in [d]$, that is the mean and standard deviation of the features are known. In Appendix B, we take into account the effect of the estimation error in features' mean and standard deviation. We characterize the changes in the misclassification probability as function of the estimation error.

5.1 Steps for Proving Theorem 4

For a fixed training set with n instances $\mathcal{D}_n = \{(\mathbf{x}(i), y(i))\}_{i\in[n]}$, the algorithm outputs a feature subset $\hat{\mathcal{J}}$ and a predictor $\hat{g}$ of the form $\hat{g} = \mathsf{sign}[\hat{f}^{\subseteq \hat{\mathcal{J}}}]$ with $\hat{f}^{\subseteq \hat{\mathcal{J}}}$ being an empirical estimate of $f^{\subseteq \hat{\mathcal{J}}}$. The misclassification rate of this predictor is denoted by

$$P_e(\hat{g}) \triangleq \mathbb{P}\Big\{Y \neq \hat{g}(X^{\hat{\mathcal{J}}})\Big\}.$$

In our analysis, we take a probabilistic approach and treat the training samples as a realization of random variables. Hence, the quantities $\hat{\mathcal{J}}, \hat{g}$ and the misclassification rate $P_e(\hat{g})$ are, also, realizations of random variables. Recall from Proposition 2 that P_{opt} is the minimum attainable misclassification rate and, thus, $P_e(\hat{g}) \geqslant \mathsf{P}_{opt}$. Our objective is to prove that with high probability $P_e(\hat{g}) \leqslant \mathsf{P}_{opt} + \epsilon$, when at least $n(\epsilon, \delta)$ number of training instances are available with $n(\epsilon,\delta)$ satisfying (6). We show this statement and find bounds on $n(\epsilon,\delta)$ in three steps discussed next.

Step 1. (Characterization of $P_e(\hat{g})$): We first derive an upper-bound on $P_e(\hat{g})$ in terms of 1-norm and 2-norm expressions. To get the desired expression, we exploit the fact that the predictor $\hat{g}$ is constructed by taking the sign of the real-valued function $\hat{f}^{\subseteq \hat{\mathcal{J}}}$ (see Algorithm 1). For that, we prove the following lemma in Appendix C.

Lemma 3 *Given a subset $\mathcal{J} \subseteq [d]$, let $h_{\mathcal{J}}$ denote an arbitrary bounded real-valued function on $\{-1,1\}^d$ that depends only on the coordinates of $\mathcal{J}$. Then,*

$$\mathbb{P}\Big\{Y \neq \mathsf{sign}[h_{\mathcal{J}}(\mathbf{X})]\Big\} \leqslant \frac{1}{2}(1 - \|f^{\subseteq \mathcal{J}}\|_1) + U(\|f^{\subseteq \mathcal{J}} - h_{\mathcal{J}}\|_2),$$

where $f^{\subseteq \mathcal{J}}$ is the projection of Y onto $\mathcal{J}$ as in Definition 6 and U is defined as $U(x) = x^3 + \frac{3}{2}x^2 + \frac{5}{4}x$, for all $x \geqslant 0$.

Hence, applying this lemma to $h_{\mathcal{J}} \equiv \hat{f}^{\subseteq \hat{\mathcal{J}}}$ gives

$$P_e(\hat{g}) \leqslant \frac{1}{2}(1 - \|f^{\subseteq \hat{\mathcal{J}}}\|_1) + U(\|f^{\subseteq \hat{\mathcal{J}}} - \hat{f}^{\subseteq \hat{\mathcal{J}}}\|_2). \tag{11}$$

Recall from Proposition 2 that P_{opt} can be written as $\mathsf{P}_{opt} = \frac{1}{2}(1 - \|f^{\subseteq \mathcal{J}^*}\|_1)$. Hence, we get

$$\mathbb{P}\Big\{P_e(\hat{g}) \geqslant \mathsf{P}_{opt} + \epsilon\Big\} \leqslant \mathbb{P}\Big\{\|f^{\subseteq \mathcal{J}^*}\|_1 - \|\hat{f}^{\subseteq \hat{\mathcal{J}}}\|_1 + 2U(\|f^{\subseteq \hat{\mathcal{J}}} - \hat{f}^{\subseteq \hat{\mathcal{J}}}\|_2) \geqslant 2\epsilon\Big\}. \tag{12}$$

With this inequality, we argue that the misclassification rate depends on two processes. The first process is the feature selection in which the subset $\hat{\mathcal{J}}$ is selected. For the selected $\hat{\mathcal{J}}$, the second process involves an estimation of the projection $f^{\subseteq \hat{\mathcal{J}}}$. That said, using the above inequality, we separate the effects of these processes on the misclassification rate. The performance of the feature selection, with no estimation taken into account, is measured as $\|f^{\subseteq \mathcal{J}^*}\|_1 - \|f^{\subseteq \hat{\mathcal{J}}}\|_1$. This measure is always non-negative as $\|f^{\subseteq \mathcal{J}^*}\|_1$ is the maximum value. The accuracy of the estimation process, on its own, is measured as $\|f^{\subseteq \hat{\mathcal{J}}} - \hat{f}^{\subseteq \hat{\mathcal{J}}}\|_2$. In the next two steps, we show that these two measures are sufficiently small with high probability.

Step 2 (Optimality of the Feature Selection): As for the performance of the feature selection process in the algorithm, we provide a bound on $\|f^{\subseteq \mathcal{J}^*}\|_1 - \|f^{\subseteq \hat{\mathcal{J}}}\|_1$. For that, we establish the following lemma.

Lemma 4 *Suppose $\hat{\mu}_j = \mu_j$ and $\hat{\sigma}_j = \sigma_j$ for all $j \in [d]$. Given $\epsilon_1, \delta_1 \in (0,1)$, with probability at least $(1-\delta_1)$, the following inequalities on score_1, as in (8),*

$$\left| \text{score}_1(\mathcal{J}) - \|f^{\subseteq \mathcal{J}}\|_1 \right| \leqslant \epsilon_1$$

hold for all subsets $\mathcal{J} \subseteq [d]$ with size k, provided that the number of training samples are at least $n_1(\epsilon_1, \delta_1) \triangleq \frac{72\ 2^{2k} c_k^2}{(\epsilon_1 - \frac{2^{k/2}}{\sqrt{n-1}})^2} \log(\frac{\binom{d}{k}}{2\delta_1})$, where c_k is the same constant as in Theorem 4.

A more general version of the lemma, incorporating the mean and variance estimations, is provided in Appendix B as Lemma 12. The argument for the proof of this lemma follows from Lemma 1 and Azuma's inequality which is presented here:

Lemma 5 ((Azuma, 1967)) *Suppose $\{X_i\}_{i \geqslant 1}$ is a sequence of i.i.d. random variables. For every $n \geqslant 1$ $Z_n = f_n(X_1, X_2, \cdots, X_n)$, where f_n is function such that for every $i \in [n]$, there exist constant α_i*

$$\left|f_n(X_1, X_2, \ldots, X_i, ..., X_n) - f_n(X_1, X_2, ..., \tilde{X}_i, ..., X_n)\right| \leqslant \alpha_i,$$

where $\tilde{X}_i$ is independent of X_i and has the same distribution as X_i. Then, for any $\epsilon > 0$,

$$\mathbb{P}\Big\{\left|Z_n - \mathbb{E}[Z_n]\right| \geqslant \epsilon\Big\} \leqslant 2\exp\left(-\frac{\epsilon^2}{2\sum_i \alpha_i^2}\right).$$

Proof of Lemma 4: We apply Azuma's inequality for $\text{score}_1(\mathcal{J})$ which is a function of the random training samples. For that we need to calculate the constants α_i. This is done in the following lemma which is proved in Appendix G.1.

Lemma 6 *The constants α_i, as in Azuma's inequality, for* score_1 *are equal to*

$$\alpha_i = \frac{6\ 2^k c_k}{n}, \qquad c_k \triangleq \max_{\mathcal{S}\subseteq[d],|\mathcal{S}|\leqslant k} \|\psi_{\mathcal{S}}\|_\infty^2. \tag{13}$$

Therefore, from Azuma's inequality, for a fixed subset $\mathcal{J} \subseteq [d]$ with $|\mathcal{J}| = k$

$$\mathbb{P}\Big\{\big|\,\mathrm{score}_1(\mathcal{J}) - \mathbb{E}[\mathrm{score}_1(\mathcal{J})]\big| \leqslant \epsilon'\Big\} \leqslant 2\exp\Big\{-\frac{n\epsilon'^2}{72\ 2^{2k}c_k^2}\Big\}.$$

Hence, using the union bound, the inequalities

$$\big|\,\mathrm{score}_1(\mathcal{J}) - \mathbb{E}[\mathrm{score}_1(\mathcal{J})]\big| \leqslant \epsilon', \quad \forall \mathcal{J} \subseteq [d],\ |\mathcal{J}| = k \tag{14}$$

hold with probability $(1-\delta_1)$ provided that $n \geqslant \tilde{n}_1(\epsilon,\delta)$, where

$$\tilde{n}_1(\epsilon',\delta_2) = \frac{72\ 2^{2k}c_k^2}{\epsilon'^2}\log\Big(\frac{\binom{d}{k}}{2\delta_1}\Big).$$

Next, from Lemma 1, we have that

$$\Big|\mathbb{E}\big[\,\mathrm{score}_1(\mathcal{J})\big] - \|f^{\subseteq\mathcal{J}}\|_1\Big| \leqslant \frac{2^{k/2}}{\sqrt{n-1}}. \tag{15}$$

Lastly, we combine this inequality to the one in (15). That said, from the triangle inequality, we have that

$$\big|\,\mathrm{score}_1(\mathcal{J}) - \|f^{\subseteq\mathcal{J}}\|_1\big| \leqslant \epsilon' + \frac{2^{k/2}}{\sqrt{n-1}}, \quad \forall \mathcal{J} \subseteq [d],\ |\mathcal{J}| = k$$

hold with probability $(1-\delta_1)$ provided that $n \geqslant \tilde{n}_1(\epsilon',\delta_1)$. Hence, setting $\epsilon_1 = \epsilon' + \frac{2^{k/2}}{\sqrt{n-1}}$ and $n_1(\epsilon_1,\delta_1) = \tilde{n}_1(\epsilon',\delta_1)$ complete the proof. ■

Thus, from the lemma and the fact that $\hat{\mathcal{J}}$ maximizes score_1, we obtain

$$\|f^{\subseteq\hat{\mathcal{J}}}\|_1 \geqslant \mathrm{score}_1(\hat{\mathcal{J}}) - \epsilon_1 \geqslant \mathrm{score}_1(\mathcal{J}^*) - \epsilon_1 \geqslant \|f^{\subseteq\mathcal{J}^*}\|_1 - 2\epsilon_1,$$

which implies that $\|f^{\subseteq\mathcal{J}^*}\|_1 - \|f^{\subseteq\hat{\mathcal{J}}}\|_1 \leqslant 2\epsilon_1$, with probability at least $(1-\delta_1)$, when at least $n_1(\epsilon_1,\delta_1)$, as in Lemma 4, number of training samples are available.

Step 3 (Accuracy of the Estimations): In this step, we show that the estimation of $f^{\subseteq\hat{\mathcal{J}}}$ is accurate enough; that is $\|f^{\subseteq\hat{\mathcal{J}}} - \hat{f}^{\subseteq\hat{\mathcal{J}}}\|_2 \leqslant \epsilon_2$ with high probability. Note that $\hat{\mathcal{J}}$ and $\hat{f}^{\subseteq\hat{\mathcal{J}}}$ are correlated. Hence, to show the desired result, we establish a stronger statement in the following lemma which is proved in Appendix E.

Lemma 7 *Suppose $\hat{\mu}_j = \mu_j$ and $\hat{\sigma}_j = \sigma_j$ for all $j \in [d]$. Given $\epsilon_2, \delta_2 \in (0,1)$, with probability at least $(1-\delta_2)$, the inequalities*

$$\|f^{\subseteq\mathcal{J}} - \hat{f}^{\subseteq\mathcal{J}}\|_2 \leqslant \epsilon_2$$

hold for all subsets $\mathcal{J} \subseteq [d]$ with size k, provided that the number of training samples are at least $n_2(\epsilon_2,\delta_2) = \frac{8\ 2^{k/2}c_k}{\epsilon_2^2}\log\Big(\frac{2kd^k}{\delta_2}\Big)$, where c_k is the same constant as in Theorem 4.

A more general version of the lemma, incorporating the mean and variance estimations, is provided in Appendix B as Lemma 13. As a result of this lemma, we have

$$\mathbb{P}\Big\{\|f^{\subseteq\hat{\mathcal{J}}} - \hat{f}^{\subseteq\hat{\mathcal{J}}}\|_2 \geqslant \epsilon_2\Big\} \leqslant \mathbb{P}\Big\{\bigcup_{\mathcal{J}:|\mathcal{J}|=k}\{\|f^{\subseteq\mathcal{J}} - \hat{f}^{\subseteq\mathcal{J}}\|_2 \geqslant \epsilon_2\}\Big\} \leqslant \delta_2, \tag{16}$$

where the first inequality holds as $|\hat{\mathcal{J}}| = k$.

Putting together (12) and (16), and using the identity $\mathbb{P}(A) \leqslant \mathbb{P}(A\bigcap B) + \mathbb{P}(B^c)$, we can show that

$$\mathbb{P}\Big\{P_e(\hat{g}) \geqslant \mathsf{P}_{opt} + \epsilon\Big\} \leqslant \mathbb{P}\Big\{\|f^{\subseteq\mathcal{J}^*}\|_1 - \|f^{\subseteq\hat{\mathcal{J}}}\|_1 \geqslant 2\epsilon - 2U(\epsilon_2)\Big\} + \delta_2, \tag{17}$$

under the condition that $n \geqslant n_2(\epsilon_2, \delta_2)$. Lastly, from Step 2 and by an appropriate choice of ϵ_1 and ϵ_2, we obtain that

$$\mathbb{P}\Big\{P_e(\hat{g}) \geqslant \mathsf{P}_{opt} + \epsilon\Big\} \leqslant \delta_1 + \delta_2,$$

under the condition that $n \geqslant \max\big\{n_1(\epsilon_1, \delta_1), n_2(\epsilon_2, \delta_2)\big\}$.

Acknowledgment

This work was supported by NSF Center for Science of Information (CSoI) Grant CCF-0939370, and in addition by NSF Grants CCF-1524312, CCF-2006440, and CCF-2007238.

Appendices

A. Fourier Analysis in Product Probability Spaces

The following facts summarize some basic properties of the Fourier expansion. These statements are derived from the orthogonality of the parities. Hence, we omit the proofs.

Fact 1 *For any bounded pair of functions $f, g : \{-1,1\}^d \mapsto \mathbb{R}$, the following statements hold:*

- *Plancherel Identity:* $\mathbb{E}[f(\mathbf{X})g(\mathbf{X})] = \sum_{\mathcal{S}\subseteq[d]} f_{\mathcal{S}} g_{\mathcal{S}}$.
- *Parseval's identity* $\|f\|_2^2 = \sum_{\mathcal{S}\subseteq[d]} f_{\mathcal{S}}^2$.
- *Jensen's Inequality:* $\|f\|_1 \leqslant \|f\|_2$.

Fact 2 *Let $f : \{-1,1\}^d \mapsto \{-1,1\}$ and $\mathcal{J} \subseteq [d]$, then the following holds*

- $\|f\|_1 = \|f\|_2 = 1$ *and* $\|f^{\subseteq\mathcal{J}}\|_2 \leqslant 1$.
- $\|f^{\subseteq\mathcal{J}}\|_2^2 \leqslant \|f^{\subseteq\mathcal{J}}\|_1 \leqslant \|f^{\subseteq\mathcal{J}}\|_2$.

Fact 3 *If $g : \{-1,1\}^d \mapsto \{-1,1\}$ is a function whose output depends only on the coordinates of a subset $\mathcal{J} \subseteq [d]$, then $g_{\mathcal{S}} = 0$ for all $\mathcal{S} \nsubseteq \mathcal{J}$. Further, for any $f : \{-1,1\}^d \mapsto \{-1,1\}$*

$$\|f - g\|_2^2 = 1 - \|f^{\subseteq\mathcal{J}}\|_2^2 + \|f^{\subseteq\mathcal{J}} - g\|_2^2$$

Fact 4 *The misclassification probability between any pair of functions $f, g : \mathcal{X} \mapsto \{-1, 1\}$ can be written as*

$$\mathbb{P}\Big\{f(\mathbf{X}) \neq g(\mathbf{X})\Big\} = \frac{1}{2} - \frac{1}{2}\langle f, g\rangle = \frac{1}{4}\|f - g\|_2^2.$$

B. The Effect of Estimating Features' Mean and Variance

In our analysis it was assumed that mean and variance are estimated accurately, that is $\hat{\mu}_j = \mu_j$ and $\hat{\sigma}_j = \sigma_j$ for all $j \in [d]$. In this section, we take into account the effect of the imperfections in mean and variance estimation. We characterize the changes in the misclassification probability as function of the estimation error. First, we compute the estimations errors as a function of the number of the samples.

Mean and variance estimations: 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 8 *Given $\epsilon_0, \delta_0 \in (0, 1)$ the following inequalities hold with probability at least $(1 - \delta_0)$*

$$\big|\hat{\mu}_j - \mu_j\big| \leqslant \epsilon_0, \qquad\qquad \big|1 - \frac{\sigma_j}{\hat{\sigma}_j}\big| \leqslant \frac{2\epsilon_0}{\sigma_j^2}, \tag{18}$$

for all $j \in [d]$, provided that atleast $n_0(\epsilon_0, \delta_0) = \frac{8}{\epsilon_0^2} \log \frac{2d}{\delta_0}$ samples are available.

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

$$\mathbb{P}\{|\hat{\mu}_j - \mu_j| \geqslant \epsilon_0\} \leqslant 2\exp\{-\frac{n\epsilon_0^2}{8}\}.$$

Therefore, applying the union bound gives

$$\mathbb{P}\Big\{\bigcup_{j=1}^{d}\big\{|\hat{\mu}_j - \mu_j| \geqslant \epsilon_0\big\}\Big\} \leqslant 2d\exp\{-\frac{n\epsilon_0^2}{8}\}.$$

Thus, the right-hand side of the above inequality is less than δ_0, if $n \geqslant \frac{8}{\epsilon_0^2}\log(\frac{2d}{\delta_0})$. 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. ■

Our technical analysis in Subsection 5.1 is under the assumption that $\epsilon_0 = 0$. In what follows, we adjust our results in Lemma 1, 4 and 7 by removing this condition. As a result we prove the following lemmas, incorporating the error is mean and variance estimations.

Lemma 9 (Generalizing Lemma 1) *The measure* $\text{score}_1 = \widehat{\|f^{\subseteq\mathcal{J}}\|_1}$ *which is defined in* (8) *is an asymptotically unbiased estimate of* $\|f^{\subseteq\mathcal{J}}\|_1$. *More precisely, for any* $\gamma \in (0, 1/2)$

$$\Big|\mathbb{E}\big[\,\text{score}_1(\mathcal{J})\big] - \|f^{\subseteq\mathcal{J}}\|_1\Big| \leqslant O(n^{-\gamma})$$

as $n \to \infty$.

Proof Let B be the event that the inequalities in (18) hold, that is $|\hat{\mu}_j - \mu_j| \leqslant \epsilon_0$ and $|1 - \frac{\sigma_j}{\hat{\sigma}_j}| \leqslant \frac{2\epsilon_0}{\sigma_j^2}$ for all $j \in [d]$. From Lemma 8, $\mathbb{P}(B) \geqslant 1 - \delta_0$. By conditioning on B we have

$$\mathbb{E}\big[\,\text{score}_1(\mathcal{J})\big] = \mathbb{P}(B)\mathbb{E}\big[\,\text{score}_1(\mathcal{J})|B\big] + (1 - \mathbb{P}(B))\mathbb{E}\big[\,\text{score}_1(\mathcal{J})|B^c\big].$$

Therefore, from triangle inequality we obtain

$$\begin{aligned}\Big|\mathbb{E}\big[\text{score}_1(\mathcal{J})\big] - \|f^{\subseteq\mathcal{J}}\|_1\Big| &\leqslant \Big|\mathbb{P}(B)\mathbb{E}\big[\,\text{score}_1(\mathcal{J})|B\big] - \|f^{\subseteq\mathcal{J}}\|_1\Big| + \Big|(1 - \mathbb{P}(B))\mathbb{E}\big[\,\text{score}_1(\mathcal{J})|B^c\big]\Big| \\ &\leqslant \underbrace{\Big|\mathbb{P}(B)\mathbb{E}\big[\,\text{score}_1(\mathcal{J})|B\big] - \|f^{\subseteq\mathcal{J}}\|_1\Big|}_{\text{(I)}} + \underbrace{\delta_0 \max_{(x(i),y(i))}\big|\,\text{score}_1(\mathcal{J})\big|}_{\text{(II)}},\end{aligned}$$

where the last inequality holds from $\mathbb{P}(B) \geqslant 1 - \delta_0$ and by upper-bounding the expectation with maximization over all realizations of the training samples.

Bounding (I): Let $\overline{\text{score}_1}$ be the score_1 measure under the assumption that $\hat{\mu}_j = \mu_j$ and $\hat{\sigma}_j = \sigma_j$ for all $j \in [d]$. From triangle inequality we have that

$$\begin{aligned}\text{(I)} &\leqslant \Big|\mathbb{P}(B)\mathbb{E}\big[\,\text{score}_1(\mathcal{J})|B\big] - \mathbb{E}\big[\overline{\text{score}_1}(\mathcal{J})\big]\Big| + \Big|\mathbb{E}\big[\overline{\text{score}_1}(\mathcal{J})\big] - \|f^{\subseteq\mathcal{J}}\|_1\Big| \\ &\overset{(a)}{\leqslant} \Big|\mathbb{P}(B)\mathbb{E}\big[\,\text{score}_1(\mathcal{J})|B\big] - \mathbb{E}\big[\overline{\text{score}_1}(\mathcal{J})\big]\Big| + \frac{2^{k/2}}{\sqrt{n-1}} \\ &\leqslant \mathbb{P}(B)\Big|\mathbb{E}\big[\,\text{score}_1(\mathcal{J})|B\big] - \mathbb{E}\big[\overline{\text{score}_1}(\mathcal{J})\big]\Big| + (1 - \mathbb{P}(B))\Big|\mathbb{E}\big[\overline{\text{score}_1}(\mathcal{J})\big]\Big| + \frac{2^{k/2}}{\sqrt{n-1}} \\ &\leqslant \Big|\mathbb{E}\big[\,\text{score}_1(\mathcal{J})|B\big] - \mathbb{E}\big[\overline{\text{score}_1}(\mathcal{J})\big]\Big| + \delta_0 \max_{(x(i),y(i))}\Big|\overline{\text{score}_1}(\mathcal{J})\Big| + \frac{2^{k/2}}{\sqrt{n-1}},\end{aligned}$$

where (a) follows from Lemma 1. Note that the conditions $\hat{\mu}_j = \mu_j$ and $\hat{\sigma}_j = \sigma_j$ in this lemma are automatically satisfied for $\overline{\text{score}_1}$. We proceed with the following lemma which is proved in Appendix G.2.

Lemma 10 *Conditioned on B the inequalities $\big|\ \overline{\text{score}_1}(\mathcal{J}) - \text{score}_1(\mathcal{J})\ \big| \leqslant \lambda(\epsilon_0)$ hold, almost surely, for all k-element subsets $\mathcal{J}$, where λ is a function satisfying $\lambda(\epsilon_0) = O(k2^k c_k \epsilon_0)$ as $\epsilon \to 0$.*

As a result of the lemma, we obtain the following bound on (I)

$$\text{(I)} \leqslant \lambda(\epsilon_0) + \delta_0 \max_{(x(i),y(i))} \left|\overline{\text{score}_1}(\mathcal{J})\right| + \frac{2^{k/2}}{\sqrt{n-1}}.$$

Bounding (II): As explained in the proof of Lemma 1, score_1 can be written as

$$\text{score}_1(\mathcal{J}) = \frac{1}{n}\sum_i \left|\hat{f}^{\subseteq \mathcal{J}}_{(i)}(\mathbf{X}(i))\right|$$

where $\hat{f}^{\subseteq \mathcal{J}}_{(i)}$ is defined as in (28) which is repeated below

$$\hat{f}^{\subseteq \mathcal{J}}_{(i)}(\mathbf{x}) \triangleq \frac{n}{n-1}\sum_{\mathcal{S}\subseteq\mathcal{J}}\Big(\hat{f}_S - \frac{1}{n}Y(i)\prod_{j\in\mathcal{S}}\frac{X_j(i)-\hat{\mu}_j}{\hat{\sigma}_j}\Big)\widehat{\psi}_{\mathcal{S}}(\mathbf{x}).$$

As a result,

$$\begin{aligned}
\text{score}_1(\mathcal{J}) \leqslant \|\hat{f}^{\subseteq\mathcal{J}}_{(1)}\|_\infty &\leqslant \frac{n}{n-1}\sum_{\mathcal{S}\subseteq\mathcal{J}}|\hat{f}_S|\|\widehat{\psi}_{\mathcal{S}}\|_\infty + \frac{1}{n}\|\widehat{\psi}_{\mathcal{S}}\|^2_\infty \\
&\overset{(a)}{\leqslant} \frac{n}{n-1}\sum_{\mathcal{S}\subseteq\mathcal{J}}\|\widehat{\psi}_{\mathcal{S}}\|^2_\infty + \frac{1}{n}\|\widehat{\psi}_{\mathcal{S}}\|^2_\infty \\
&\leqslant \frac{n+1}{n-1}\sum_{\mathcal{S}\subseteq\mathcal{J}}\|\widehat{\psi}_{\mathcal{S}}\|^2_\infty,
\end{aligned}$$

where (a) holds because $\hat{f}_S = \frac{1}{n}\sum_i Y(i)\widehat{\psi}_{\mathcal{S}}(\mathbf{X}(i)) \leqslant \|\widehat{\psi}_{\mathcal{S}}\|_\infty$. We proceed with the following lemma which is proved in Appendix G.3.

Lemma 11 *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 $\epsilon_0 \to 0$.*

Therefore, from Lemma 11 and the inequality $(x+y)^2 \leqslant 2(x^2+y^2)$, we obtain

$$\begin{aligned}
\text{score}_1(\mathcal{J}) &\leqslant \frac{n+1}{n-1}2\sum_{\mathcal{S}\subseteq\mathcal{J}}\Big(\|\psi_{\mathcal{S}}\|^2_\infty + \gamma^2(\epsilon_0)\Big) \\
&\overset{(c)}{\leqslant} 6\ 2^k\big(c_k + \gamma^2(\epsilon_0)\big) = O(2^k c_k),
\end{aligned} \tag{19}$$

where (c) follows from the definition of c_k as in (13), which implies that $\|\psi_{\mathcal{S}}\|^2_\infty \leqslant c_k$.

Combining the bounds and tuning (ϵ_0, δ_0)**:** Now, by combining the bound on (I) and (II), we have that

$$\begin{aligned}\Big|\mathbb{E}\big[\mathrm{score}_1(\mathcal{J})\big] - \|f^{\subseteq\mathcal{J}}\|_1\Big| &\leqslant \frac{2^{k/2}}{\sqrt{n-1}} + \lambda(\epsilon_0) + \delta_0 \max_{(x(i),y(i))} \Big|\overline{\mathrm{score}_1}(\mathcal{J})\Big| + \delta_0 O(2^k c_k)\\ &\leqslant \frac{2^{k/2}}{\sqrt{n-1}} + \lambda(\epsilon_0) + \delta_0(O(2^k c_k) + \lambda(\epsilon_0)) + \delta_0 O(2^k c_k),\end{aligned}$$

where the last inequality holds, because from Lemma 10 and inequality (19) we can write

$$\overline{\mathrm{score}_1}(\mathcal{J}) \leqslant \mathrm{score}_1(\mathcal{J}) + \lambda(\epsilon_0) \leqslant O(2^k c_k) + \lambda(\epsilon_0).$$

Now, we tune (ϵ_0, δ_0). For $\gamma \in (0, 1/2)$, set $\epsilon_0 = \big(k 2^k c_k n^\gamma\big)^{-1}$. Hence, $\delta_0 = 2d \exp\{-\frac{n^{1-2\gamma}}{8k^2 2^{2k} c_k^2}\}$. With this choice $\lambda(\epsilon_0) = O(n^{-\gamma})$ and plugging it into the above inequality implies

$$\Big|\mathbb{E}\big[\mathrm{score}_1(\mathcal{J})\big] - \|f^{\subseteq\mathcal{J}}\|_1\Big| \leqslant \frac{2^{k/2}}{\sqrt{n-1}} + O(n^{-\gamma}) = O(n^{-\gamma}).$$

■

Lemma 12 (Generalizing Lemma 4) *given $\epsilon_1, \delta_1 \in (0,1)$, with probability at least $(1-\delta_1)$, the inequalities $\Big|\ \mathrm{score}_1(\mathcal{J}) - \|f^{\subseteq\mathcal{J}}\|_1\ \Big| \leqslant \epsilon_1$ hold for all subsets $\mathcal{J} \subseteq [d]$ with size k, provided that the number of training samples are atleast $O(\frac{k^2 2^{2k} c_k^2}{\epsilon_1^2} \log \frac{d}{\delta_1})$.*

Proof Let $\overline{\mathrm{score}_1}$ be the score_1 measure 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 (18) hold. From triangle inequality we have that

$$\Big|\ \mathrm{score}_1(\mathcal{J}) - \|f^{\subseteq\mathcal{J}}\|_1\ \Big| \leqslant \underbrace{\Big|\overline{\mathrm{score}_1}(\mathcal{J}) - \|f^{\subseteq\mathcal{J}}\|_1\ \Big|}_{V} + \underbrace{\Big|\ \mathrm{score}_1(\mathcal{J}) - \overline{\mathrm{score}_1}(\mathcal{J})\Big|}_{W}.$$

Let V and W denote the first and the second term above, respectively. We know that W is measurable with respect to B. In particular, from Lemma 10, given B, $W \leqslant \lambda(\epsilon_0)$ almost surely. Therefore, we have

$$\begin{aligned}\mathbb{P}\Big\{\Big|\ \mathrm{score}_1(\mathcal{J}) - \|f^{\subseteq\mathcal{J}}\|_1\ \Big| \leqslant \epsilon_1 + \lambda(\epsilon_0)\Big\} &\geqslant \mathbb{P}\Big\{V + W \leqslant \epsilon_1 + \lambda(\epsilon_0)\Big\}\\ &\geqslant \mathbb{P}\Big\{V \leqslant \epsilon_1, W \leqslant \lambda(\epsilon_0)\Big\}\\ &\geqslant \mathbb{P}\Big\{V \leqslant \epsilon_1, W \leqslant \lambda(\epsilon_0), B\Big\}\\ &= \mathbb{P}\Big\{V \leqslant \epsilon_1, B\Big\}\\ &\overset{(a)}{=} \mathbb{P}\Big\{V \leqslant \epsilon_1\Big\}\mathbb{P}\Big\{B\Big\}\\ &\geqslant (1-\delta_1)(1-\delta_0),\end{aligned}$$

where (a) holds as B is independent of V. Now set $\epsilon_0 = \frac{\epsilon_1}{k2^k c_k}$, and $\delta_0 = \delta_1$. With this choice, $n_0(\epsilon_0, \delta_0) = \frac{8k^2 2^{2k} c_k^2}{\epsilon_1^2} \log \frac{2d}{\delta_1}$ and $\lambda(\epsilon_0) = O(\epsilon_1)$. Hence, by appropriate choice of ϵ_1, δ_1, the following inequality

$$\Big|\, \mathrm{score}_1(\mathcal{J}) - \|f^{\subseteq \mathcal{J}}\|_1 \,\Big| \leqslant \epsilon_1,$$

holds with probability $(1-\delta_1)$ for for all k-element subsets $\mathcal{J}$, provided that there are atleast $n_1(\epsilon_1, \delta_1) + O(\frac{k^2 2^{2k} c_k^2}{\epsilon_1^2} \log \frac{d}{\delta_1})$ samples. The proof is complete by noting that $n_1 \leqslant O(\frac{k^2 2^{2k} c_k^2}{\epsilon_1^2} \log \frac{d}{\delta_1})$. ∎

Lemma 13 (Generalizing Lemma 7) *Given $\epsilon_2, \delta_2 \in (0,1)$, with probability at least $(1-\delta_2)$, the inequalities*

$$\|f^{\subseteq \mathcal{J}} - \hat{f}^{\subseteq \mathcal{J}}\|_2 \leqslant \epsilon_2$$

hold for all subsets $\mathcal{J} \subseteq [d]$ with size k, provided that the number of training samples are atleast $O(\frac{k^2 2^{2k} c_k^2}{\epsilon_1^2} \log \frac{d}{\delta_1})$.

Proof Let $\bar{f}^{\subseteq J}$ denote the version of $\hat{f}^{\subseteq \mathcal{J}}$ 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 (18) hold. From Minkowsky's inequality, by adding and subtracting $\bar{f}^{\subseteq J}$ we have

$$\|f^{\subseteq \mathcal{J}} - \hat{f}^{\subseteq \mathcal{J}}\|_2 \leqslant \underbrace{\|f^{\subseteq \mathcal{J}} - \bar{f}^{\subseteq J}\|_2}_{V} + \underbrace{\|\bar{f}^{\subseteq J} - \hat{f}^{\subseteq \mathcal{J}}\|_2}_{W}.$$

Let V and W denote the first and the second term above, respectively. We proceed by the following lemma which is proved in Appendix G.4.

Lemma 14 *Conditioned on B, the inequalities $\|\bar{f}^{\subseteq J} - \hat{f}^{\subseteq \mathcal{J}}\|_\infty \leqslant \lambda(\epsilon)$ hold, almost surely, for all k-element subsets $\mathcal{J} \subset [d]$, where λ is a function satisfying $\lambda(\epsilon_0) = O(k2^k c_k \epsilon_0)$ as $\epsilon_0 \to 0$.*

From Lemma 14, we know that W is measurable with respect to B. In particular, conditioned on B, $W \leqslant \lambda(\epsilon_0)$. Therefore, using the inequality $\|\cdot\|_2 \leqslant \|\cdot\|_\infty$, we have

$$\begin{aligned}
\mathbb{P}\Big\{\|f^{\subseteq \mathcal{J}} - \hat{f}^{\subseteq \mathcal{J}}\|_2 \leqslant \epsilon_2 + \lambda(\epsilon_0)\Big\} &\geqslant \mathbb{P}\Big\{V + W \leqslant \epsilon_2 + \lambda(\epsilon_0)\Big\} \\
&\geqslant \mathbb{P}\Big\{V \leqslant \epsilon_2, W \leqslant \lambda(\epsilon_0)\Big\} \\
&\geqslant \mathbb{P}\Big\{V \leqslant \epsilon_2, W \leqslant \lambda(\epsilon_0), B\Big\} \\
&= \mathbb{P}\Big\{V \leqslant \epsilon_2, B\Big\} \\
&\overset{(a)}{=} \mathbb{P}\Big\{V \leqslant \epsilon_2\Big\}\mathbb{P}\Big\{B\Big\} \\
&\geqslant (1-\delta_2)(1-\delta_0),
\end{aligned}$$

where (a) holds as B is independent of V. Now set $\epsilon_0 = \frac{\epsilon_2}{k 2^k c_k}$, and $\delta_0 = \delta_2$. With this choice, $n_0(\epsilon_0, \delta_0) = \frac{8k^2 2^{2k} c_k^2}{\epsilon_2^2} \log\frac{2d}{\delta_2}$ and $\lambda(\epsilon_0) = O(\epsilon_2)$. Hence, by appropriate choice of ϵ_2, δ_2, the inequality $\|f^{\subseteq\mathcal{J}} - \hat{f}^{\subseteq\mathcal{J}}\|_2 \leqslant \epsilon_2$ holds with probability $(1-\delta_2)$ for for all k-element subsets $\mathcal{J}$, provided that there are atleast $n_2(\epsilon_2, \delta_2) + O(\frac{k^2 2^{2k} c_k^2}{\epsilon_2^2} \log\frac{d}{\delta_2})$ samples. Lastly, the proof is complete by noting that $n_2 \leqslant O(\frac{k^2 2^{2k} c_k^2}{\epsilon_2^2} \log\frac{d}{\delta_2})$.

■

C. Proof of Lemma 3

Since the range of f belongs to $\{-1, 1\}$, then from Lemma2 the misclassification probability can be written as

$$\mathbb{P}\Big\{Y \neq \mathsf{sign}[h_{\mathcal{J}}(\mathbf{X})]\Big\} = \frac{1}{4}\big(1 - \|f^{\subseteq\mathcal{J}}\|_2^2 + \|f^{\subseteq\mathcal{J}} - \mathsf{sign}[h_{\mathcal{J}}]\|_2^2\big). \tag{20}$$

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

$$\begin{aligned}\|f^{\subseteq\mathcal{J}} - \mathsf{sign}[h_{\mathcal{J}}]\|_2^2 &\overset{(a)}{\leqslant} \Big(\|f^{\subseteq\mathcal{J}} - h_{\mathcal{J}}\|_2 + \|h_{\mathcal{J}} - \mathsf{sign}[h_{\mathcal{J}}]\|_2\Big)^2, \\ &= \Big(\|f^{\subseteq\mathcal{J}} - h_{\mathcal{J}}\|_2^2 + \|h_{\mathcal{J}} - \mathsf{sign}[h_{\mathcal{J}}]\|_2^2 \\ &\qquad + 2\|f^{\subseteq\mathcal{J}} - h_{\mathcal{J}}\|_2 \|h_{\mathcal{J}} - \mathsf{sign}[h_{\mathcal{J}}]\|_2\Big),\end{aligned} \tag{21}$$

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

$$\|h_{\mathcal{J}} - \mathsf{sign}[h_{\mathcal{J}}]\|_2^2 = \mathbb{E}\left[(1 - |h_{\mathcal{J}}(X^{\mathcal{J}})|)^2\right] = 1 + \|h_{\mathcal{J}}\|_2^2 - 2\|h_{\mathcal{J}}\|_1. \tag{22}$$

From this relation and equations (20), (21), we obtain the following upper bound

$$\begin{aligned}4\mathbb{P}\Big\{Y \neq \mathsf{sign}[h_{\mathcal{J}}(\mathbf{X})]\Big\} \leqslant 2 - 2\|h_{\mathcal{J}}\|_1 + \underbrace{\|h_{\mathcal{J}}\|_2^2 - \|f^{\subseteq\mathcal{J}}\|_2^2}_{\text{(I)}} + \|f^{\subseteq\mathcal{J}} - h_{\mathcal{J}}\|_2^2 \\ + 2\|f^{\subseteq\mathcal{J}} - h_{\mathcal{J}}\|_2 \underbrace{\|h_{\mathcal{J}} - \mathsf{sign}[h_{\mathcal{J}}]\|_2}_{\text{(II)}}.\end{aligned} \tag{23}$$

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

Bounding (I): From the triangle inequality for 2-norm, we have

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

where the second inequality is due to Fact 2 that $\|f^{\subseteq\mathcal{J}}\|_2 \leqslant 1$. Hence, the term (I) in (23) is upper bounded as

$$\text{(I)} \leqslant \lambda_1 \triangleq \|h_{\mathcal{J}} - f^{\subseteq\mathcal{J}}\|_2^2 + 2\|h_{\mathcal{J}} - f^{\subseteq\mathcal{J}}\|_2. \tag{24}$$

Bounding (II): From (22), we have

$$\begin{aligned}
\|h_{\mathcal{J}} - \mathsf{sign}[h_{\mathcal{J}}]\|_2^2 &= 1 + \|h_{\mathcal{J}}\|_2^2 - 2\|h_{\mathcal{J}}\|_1 \\
&\overset{(a)}{\leqslant} 1 + 2(\|f^{\subseteq\mathcal{J}}\|_2^2 + \|f^{\subseteq\mathcal{J}} - h_{\mathcal{J}}\|_2^2) - 2\|h_{\mathcal{J}}\|_1 \\
&\overset{(b)}{=} 1 + 2(\|f^{\subseteq\mathcal{J}}\|_2^2 + \|f^{\subseteq\mathcal{J}} - h_{\mathcal{J}}\|_2^2) - 2\big(\|f^{\subseteq\mathcal{J}}\|_1 + (\|h_{\mathcal{J}}\|_1 - \|f^{\subseteq\mathcal{J}}\|_1)\big) \\
&= 1 + 2(\|f^{\subseteq\mathcal{J}}\|_2^2 - \|f^{\subseteq\mathcal{J}}\|_1) + 2\|f^{\subseteq\mathcal{J}} - h_{\mathcal{J}}\|_2^2 - 2\big(\|h_{\mathcal{J}}\|_1 - \|f^{\subseteq\mathcal{J}}\|_1\big) \\
&\overset{(c)}{\leqslant} 1 + 2\|f^{\subseteq\mathcal{J}} - h_{\mathcal{J}}\|_2^2 - 2\big(\|h_{\mathcal{J}}\|_1 - \|f^{\subseteq\mathcal{J}}\|_1\big) \\
&\overset{(d)}{\leqslant} 1 + 2\|f^{\subseteq\mathcal{J}} - h_{\mathcal{J}}\|_2^2 + 2\|f^{\subseteq\mathcal{J}} - h_{\mathcal{J}}\|_2
\end{aligned} \tag{25}$$

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

$$\Big|\|f^{\subseteq\mathcal{J}}\|_1 - \|h_{\mathcal{J}}\|_1\Big| \leqslant \|f^{\subseteq\mathcal{J}} - h_{\mathcal{J}}\|_1 \leqslant \|f^{\subseteq\mathcal{J}} - h_{\mathcal{J}}\|_2. \tag{26}$$

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

Next, we show that the quantity $\big\|h_{\mathcal{J}} - \mathsf{sign}[h_{\mathcal{J}}]\big\|_2$ without the square is upper bounded by the same term as in the right-hand side of (25). That is

$$\text{(II)} = \big\|h_{\mathcal{J}} - \mathsf{sign}[h_{\mathcal{J}}]\big\|_2 \leqslant \lambda_2 \triangleq 1 + 2\|f^{\subseteq\mathcal{J}} - h_{\mathcal{J}}\|_2^2 + 2\|f^{\subseteq\mathcal{J}} - h_{\mathcal{J}}\|_2. \tag{27}$$

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

As a result of the bounds in (23), (24), and (27) we obtain that

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

where the last inequality is due to (26). Therefore, from the definition of λ_1 and λ_2, and the function U in the statement of the lemma, we obtain

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

This completes the proof.

D. Proof of Lemma 1

Proof We start with rewriting score_1. Define, the function

$$\hat{f}^{\subseteq \mathcal{J}}_{(i)}(\mathbf{x}) \triangleq \frac{n}{n-1} \sum_{\mathcal{S}\subseteq \mathcal{J}} \Big(\hat{f}_{\mathcal{S}} - \frac{1}{n} Y(i) \prod_{j\in\mathcal{S}} \frac{X_j(i) - \hat{\mu}_j}{\hat{\sigma}_j}\Big)\widehat{\psi}_{\mathcal{S}}(\mathbf{x}), \tag{28}$$

for all $x \in \{-1,1\}^d$. With this definition, given any $\mathbf{x}$, the quantity $\hat{f}^{\subseteq \mathcal{J}}_{(i)}()$ is independent of $(X^d(i), Y(i))$. Further, we can write score_1 as the summation $\mathrm{score}_1(\mathcal{J}) = \frac{1}{n}\sum_i |\hat{f}^{\subseteq \mathcal{J}}_{(i)}(\mathbf{X}(i))|$. Hence, the exception of score_1 taken over the training samples gives

$$\begin{aligned}
\mathbb{E}[\mathrm{score}_1(\mathcal{J})] &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathbf{X}(1),\dots,\mathbf{X}(n)}\Big[\ \Big|\ \hat{f}^{\subseteq \mathcal{J}}_{(i)}(\mathbf{X}(i))\ \Big|\ \Big] \\
&= \mathbb{E}_{\mathbf{X}(1),\dots,\mathbf{X}(n)}\Big[\ \Big|\ \hat{f}^{\subseteq \mathcal{J}}_{(1)}(\mathbf{X}(1))\ \Big|\ \Big] \\
&= \mathbb{E}_{\mathbf{X}(2),\dots,\mathbf{X}(n)}\mathbb{E}_{\mathbf{X}(1)}\Big[\ \Big|\ \hat{f}^{\subseteq \mathcal{J}}_{(1)}(\mathbf{X}(1))\ \Big|\ \Big] \\
&= \mathbb{E}_{\mathbf{X}(2),\dots,\mathbf{X}(n)}\big[\|\hat{f}^{\subseteq \mathcal{J}}_{(1)}\|_1\big],
\end{aligned} \tag{29}$$

where the first equality is due to the symmetry with respect to the index i of the training samples. The last equality is due to the definition of 1-norm and the property that the function $\hat{f}^{\subseteq \mathcal{J}}_{(1)}$ is independent of $(\mathbf{X}(1), Y(1))$. Note that $\hat{f}^{\subseteq \mathcal{J}}_{(1)}$ is as an estimation of the projection $f^{\subseteq \mathcal{J}}$ using the $(n-1)$ training samples $(\mathbf{X}(i), Y(i)), i = 2, 3, \dots, n$. Next, we bound the difference $\Big|\mathbb{E}\|\hat{f}^{\subseteq \mathcal{J}}_{(1)}\|_1 - \|f^{\subseteq \mathcal{J}}\|_1\Big|$. Observe that

$$\begin{aligned}
\Big|\mathbb{E}\big[\|\hat{f}^{\subseteq \mathcal{J}}_{(1)}\|_1\big] - \|f^{\subseteq \mathcal{J}}\|_1\Big| &\leqslant \mathbb{E}\big[\|f^{\subseteq \mathcal{J}} - \hat{f}^{\subseteq \mathcal{J}}_{(1)}\|_1\big] \\
&\leqslant \mathbb{E}\big[\|f^{\subseteq \mathcal{J}} - \hat{f}^{\subseteq \mathcal{J}}_{(1)}\|_2\big] \\
&\leqslant \sqrt{\mathbb{E}\big[\|f^{\subseteq \mathcal{J}} - \hat{f}^{\subseteq \mathcal{J}}_{(1)}\|_2^2\big]}
\end{aligned}$$

where the first inequality is obtained by applying the triangle inequality twice, one for $\|\hat{f}^{\subseteq \mathcal{J}}_{(1)}\|_1$ and once for $\|f^{\subseteq \mathcal{J}}\|_1$. The second inequality is from the identity $\|\cdot\|_1 \leqslant \|\cdot\|_2$ as in Fact 1. The third inequality is due to the Jensen's inequality. Note that as $\hat{\mu}_j = \mu_j$ and $\hat{\sigma}_j = \sigma_j$, then $\widehat{\psi}_{\mathcal{S}} \equiv \psi_{\mathcal{S}}$. Therefore, by Parseval's identity in Fact 1, we have

$$\mathbb{E}\big[\|f^{\subseteq \mathcal{J}} - \hat{f}^{\subseteq \mathcal{J}}_{(1)}\|_2^2\big] = \sum_{S\subseteq J} \mathbb{E}\big[|f_{\mathcal{S}} - \hat{f}_{(1),\mathcal{S}}|^2\big] = \sum_{S\subseteq J} \mathrm{var}\left(\hat{f}_{(1),\mathcal{S}}\right).$$

Note that $\hat{f}_{(1),\mathcal{S}}$ is the empirical average of i.i.d. random variables $Y(i)\psi_{\mathcal{S}}(\mathbf{X}(i))$ for $i = 2, 3, \dots, n$. Thus,

$$\begin{aligned}
\mathrm{var}\left(\hat{f}_{(1),\mathcal{S}}\right) &= \frac{1}{n-1}\mathrm{var}\left(Y\psi_{\mathcal{S}}(\mathbf{X})\right) \\
&= \frac{1}{n-1}\left(\mathbb{E}\big[Y^2\psi^2_{\mathcal{S}}(\mathbf{X})\big] - f^2_{\mathcal{S}}\right) \\
&= \frac{1}{n-1}(1 - f^2_{\mathcal{S}}),
\end{aligned}$$

where the last equality holds because of of the following chain of equalities:

$$\mathbb{E}\big[Y^2\psi_{\mathcal{S}}^2(\mathbf{X})\big] = \mathbb{E}\big[\psi_{\mathcal{S}}^2(\mathbf{X})\big] = \langle\psi_{\mathcal{S}},\psi_{\mathcal{S}}\rangle = 1,$$

where the first equality holds because $Y^2 = 1$ which is due to the fact that $Y \in \{-1,1\}$. The last equality is due to orthonormality of the parities.

As a result of the above argument, we can write

$$\mathbb{E}\big[\|f^{\subseteq\mathcal{J}} - \hat{f}_{(1)}^{\subseteq\mathcal{J}}\|_2^2\big] = \frac{1}{n-1}\sum_{S\subseteq J}(1-f_{\mathcal{S}}^2) = \frac{1}{n-1}(2^{|\mathcal{J}|} - \|f^{\subseteq\mathcal{J}}\|_2^2)$$
$$\leqslant \frac{1}{n-1}2^k$$

Putting all together we get that

$$\Big|\mathbb{E}\big[\|\hat{f}_{(1)}^{\subseteq\mathcal{J}}\|_1\big] - \|f^{\subseteq\mathcal{J}}\|_1\Big| \leqslant \frac{2^{k/2}}{\sqrt{n-1}}$$

The proof is complete by the above inequality and (29).

■

E. Proof of Lemma 7

Proof Assuming that $\hat{\mu}_j = \mu_j$ and $\hat{\sigma}_j = \sigma_j$ and from the definition of $\hat{f}^{\subseteq\mathcal{J}}$, we obtain that

$$\hat{f}^{\subseteq\mathcal{J}}(\mathbf{x}) = \sum_{\mathcal{S}\subseteq\mathcal{J}} \hat{f}_S\ \psi_{\mathcal{S}}(\mathbf{x}), \qquad \forall \mathbf{x}\in\mathcal{X}^d.$$

In addition, by definition of the projection function $f^{\subseteq\mathcal{J}}$, we have

$$f^{\subseteq\mathcal{J}}(\mathbf{x}) = \sum_{\mathcal{S}\subseteq\mathcal{J}} f_{\mathcal{S}}\ \psi_{\mathcal{S}}(\mathbf{x}), \qquad \forall \mathbf{x}\in\mathcal{X}^d.$$

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

$$\|f^{\subseteq\mathcal{J}} - \hat{f}^{\subseteq\mathcal{J}}\|_2^2 = \sum_{S\subseteq J}|f_{\mathcal{S}} - \hat{f}_S|^2.$$

In what follows, we show that $|f_{\mathcal{S}} - \hat{f}_S| \leqslant \epsilon$ for all subsets $\mathcal{S}\subseteq[d]$ with $|\mathcal{S}|\leqslant k$.

Note that $\hat{f}_S$ is a function of the training random samples $(X(i),Y(i)), i = 1,2,...,n$. Observe that $\mathbb{E}[\hat{f}_S] = f_{\mathcal{S}}$ which implies that $\hat{f}_S$ is an unbiased estimation of $f_{\mathcal{S}}$. Since the samples are drawn i.i.d., we apply Azuma's inequality (Lemma 5) to bound the probability of the event $|f_{\mathcal{S}} - \hat{f}_S| \geqslant \epsilon'$.

For that, we first find the constants c_i as in Lemma 5. Fix $i\in[d]$ and suppose $(X^d(i),Y(i))$ in the training set is replaced with an i.i.d. copy $(\tilde{X}^d(i),\tilde{Y}(i))$. With this replacement $\hat{f}_S$ is changed to another random variable denoted by $\tilde{f}_S$. Then

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

where $\|\psi_S\|_\infty = \max_{x^d}|\psi_S(x^d)|$. Therefore, from Azuma's inequality, for any $\epsilon' \in (0,1)$

$$\mathbb{P}\Big\{|\hat{f}_S - f_S| \geqslant \epsilon'\Big\} \leqslant 2\exp\Big\{-\frac{n\epsilon'^2}{8\|\psi_{\mathcal{S}}\|_\infty^2}\Big\}.$$

Applying the union bound for all subsets $S \subset [d]$ with cardinality at most k, gives the following upper-bound

$$\begin{aligned}\mathbb{P}\Big\{\bigcup_{\substack{s\subset[d],\\ |s|\leqslant k}}\{|\hat{f}_S - f_S| \geqslant \epsilon'\}\Big\} &\leqslant 2\Big[\sum_{m=0}^{k}\binom{d}{m}\Big]\exp\Big\{-\frac{n\epsilon'^2}{8c_k}\Big\} \\ &\leqslant 2kd^k\exp\Big\{-\frac{n\epsilon'^2}{8c_k}\Big\}\Big\} \qquad (30)\end{aligned}$$

where $c_k = \max_{\mathcal{S}\subseteq[d],|\mathcal{S}|\leqslant k}\|\psi_{\mathcal{S}}\|_\infty^2$ and the last inequality holds because for $k \ll d/2$

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

We find n for which the right hand side of (30) is less than δ. For that we have

$$n(\epsilon,\delta) \geqslant \frac{8c_k}{\epsilon'^2}\log\left(\frac{2kd^k}{\delta}\right).$$

Next, note that

$$\|f^{\subseteq\mathcal{J}} - \hat{f}^{\subseteq\mathcal{J}}\|_2^2 = \sum_{S\subseteq J}|f_{\mathcal{S}} - \hat{f}_S|^2 \leqslant \epsilon'^2 2^k.$$

Therefore, $\|f^{\subseteq\mathcal{J}} - \hat{f}^{\subseteq\mathcal{J}}\|_2 \leqslant \epsilon' 2^{k/2}$, and the proof is complete by setting $\epsilon' = \epsilon 2^{-k/2}$. ∎

F. Proof of Theorem 5

Note that $\hat{g} \equiv \mathsf{sign}[\hat{f}^{\subseteq\hat{\mathcal{J}}}]$ and by $P_e(\hat{g})$ denote its misclassification probability. Then, from Lemma 3, we have that

$$P_e(\hat{g}) \triangleq \mathbb{P}_{\mathbf{X},Y}\{Y \neq \hat{g}(\mathbf{X}^{\hat{\mathcal{J}}})\} \leqslant \frac{1}{2}(1 - \|f^{\subseteq\hat{\mathcal{J}}}\|_1) + U(\|f^{\subseteq\hat{\mathcal{J}}} - \hat{f}^{\subseteq\hat{\mathcal{J}}}\|_2),$$

where U is a polynomial of the form $U(x) = x^3 + \frac{3}{2}x^2 + \frac{5}{4}x$. Taking the expectation with respect to the training samples $\mathcal{D}_n$ gives,

$$\mathbb{E}_{\mathcal{D}_n}[P_e(\hat{g})] \leqslant \frac{1}{2}(1 - \mathbb{E}_{\mathcal{D}_n}[\|f^{\subseteq\hat{\mathcal{J}}}\|_1]) + \mathbb{E}_{\mathcal{D}_n}[U(\|f^{\subseteq\hat{\mathcal{J}}} - \hat{f}^{\subseteq\hat{\mathcal{J}}}\|_2)].$$

Therefore, subtracting P_{opt} gives

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_n}[P_e(\hat{g})] - \mathsf{P}_{opt} &\leqslant \frac{1}{2}(1 - \mathbb{E}_{\mathcal{D}_n}[\|f^{\subseteq\hat{\mathcal{J}}}\|_1]) + \mathbb{E}_{\mathcal{D}_n}\big[U(\|f^{\subseteq\hat{\mathcal{J}}} - \hat{f}^{\subseteq\hat{\mathcal{J}}}\|_2)\big] - \mathsf{P}_{opt}\\ &= \frac{1}{2}\underbrace{(\|f^{\subseteq\mathcal{J}^*}\|_1 - \mathbb{E}_{\mathcal{D}_n}[\|f^{\subseteq\hat{\mathcal{J}}}\|_1])}_{\text{(I)}} + \underbrace{\mathbb{E}_{\mathcal{D}_n}\big[U(\|f^{\subseteq\hat{\mathcal{J}}} - \hat{f}^{\subseteq\hat{\mathcal{J}}}\|_2)\big]}_{\text{(II)}}. \end{aligned} \tag{31}$$

Next, we provide upper bounds for the terms (I) and (II).

Bounding (I): Note that $\hat{\mathcal{J}}$ is a feature subset maximizing score_1 as in (8). Whereas, $\mathcal{J}^*$ maximizes $\|f^{\subseteq\mathcal{J}}\|_1$. Therefore, by adding and subtracting $\mathbb{E}[\text{score}_1(\hat{\mathcal{J}})]$ and $\mathbb{E}[\text{score}_1(\mathcal{J}^*)]$ we have:

$$\begin{aligned}(I) &= \big(\|f^{\subseteq\mathcal{J}^*}\|_1 - \mathbb{E}[\text{score}_1(\mathcal{J}^*)]\big) + \big(\mathbb{E}[\text{score}_1(\mathcal{J}^*)] - \mathbb{E}[\text{score}_1(\hat{\mathcal{J}})]\big)\\ &\quad + \big(\mathbb{E}[\text{score}_1(\hat{\mathcal{J}})] - \mathbb{E}[\|f^{\subseteq\hat{\mathcal{J}}}\|_1]\big)\\ &\overset{(a)}{\leqslant} \mathbb{E}\big[\|f^{\subseteq\mathcal{J}^*}\|_1 - \text{score}_1(\mathcal{J}^*)\big] + \mathbb{E}\big[\text{score}_1(\hat{\mathcal{J}}) - \|f^{\subseteq\hat{\mathcal{J}}}\|_1\big]\\ &\leqslant 2 \max_{\mathcal{J}:|\mathcal{J}|=k} \big|\mathbb{E}[\text{score}_1(\mathcal{J})] - \|f^{\subseteq\mathcal{J}}\|_1\big|\\ &\overset{(b)}{\leqslant} O(n^{-\gamma}), \end{aligned} \tag{32}$$

where (a) holds as $\text{score}_1(\hat{\mathcal{J}}) \geqslant \text{score}_1(\mathcal{J}^*)$ and inequality (b) follows from Lemma 9 with $\gamma \in (0, 1/2)$.

Bounding (II): We start by removing the effect of $\hat{\mathcal{J}}$ by maximizing over all feature subsets $\mathcal{J}$:

$$\text{(II)} \leqslant \max_{\mathcal{J}:|\mathcal{J}|=k} \mathbb{E}_{\mathcal{D}_n}\big[U(\|f^{\subseteq\mathcal{J}} - \hat{f}^{\subseteq\mathcal{J}}\|_2)\big]$$

Fix a k-element subset $\mathcal{J} \subseteq [d]$ and let $Z \triangleq \|f^{\subseteq\mathcal{J}} - \hat{f}^{\subseteq\mathcal{J}}\|_2$. Note that Z is a random variables which is a function of the training samples $\mathcal{D}_n$. From Lemma 7 we know that given ϵ_2, δ_2, the inequality $\mathbb{P}\{Z > \epsilon_2\} \leqslant \delta_2$ holds if $n \geqslant n_2(\epsilon_2, \delta_2)$, where $n_2(\cdot)$ is defined in the lemma. Therefore, by conditioning on the event $\{Z \leqslant \epsilon_2\}$ and its complement, we have

$$\begin{aligned}\mathbb{E}\big[U(Z)\big] &= \mathbb{P}\big\{Z \leqslant \epsilon_2\big\}\mathbb{E}\big[U(Z)\big|Z \leqslant \epsilon_2\big] + \mathbb{P}\big\{Z > \epsilon_2\big\}\mathbb{E}\big[U(Z)\big|Z > \epsilon_2\big]\\ &\overset{(a)}{\leqslant} \mathbb{E}\big[U(Z)\big|Z \leqslant \epsilon_2\big] + \delta_2\mathbb{E}\big[U(Z)\big|Z > \epsilon_2\big]\\ &\overset{(b)}{\leqslant} U(\epsilon_2) + \delta_2\mathbb{E}\big[U(Z)\big|Z > \epsilon_2\big]\\ &\overset{(c)}{\leqslant} U(\epsilon_2) + \delta_2 \max_{(\mathbf{x}(i),y(i))} U(\|f^{\subseteq\mathcal{J}} - \hat{f}^{\subseteq\mathcal{J}}\|_2), \end{aligned} \tag{33}$$

where (a) is due to the inequalities $\mathbb{P}\{Z \leqslant \epsilon_2\} \leqslant 1$ and $\mathbb{P}\{Z > \epsilon_2\} \leqslant \delta_2$. Inequality (b) holds due to the conditioning $Z \leqslant \epsilon_2$ and the fact that U is a monotone function. Inequality (c) follows by replacing the expectation with maximization over all realizations of the training samples. Next, we upper-bound the term inside $U(\cdot)$. From the triangle inequality, we obtain that

$$\|f^{\subseteq\mathcal{J}} - \hat{f}^{\subseteq\mathcal{J}}\|_2 \leqslant \|f^{\subseteq\mathcal{J}}\|_2 + \|\hat{f}^{\subseteq\mathcal{J}}\|_2 \leqslant 1 + \|\hat{f}^{\subseteq\mathcal{J}}\|_\infty,$$

where the last inequality holds due to $\|f^{\subseteq\mathcal{J}}\|_2 \leqslant 1$ and the identity $\|\cdot\|_2 \leqslant \|\cdot\|_\infty$. Not that $\hat{f}^{\subseteq\mathcal{J}} \equiv \sum_{\mathcal{S}\subseteq\mathcal{J}} \hat{f}_{\mathcal{S}}\widehat{\psi}_{\mathcal{S}}$ and that the Fourier coefficients $\hat{f}_{\mathcal{S}}$ can be written as a linear combination of the parities as in (7). Hence, from the definition of $\|\cdot\|_\infty$, we obtain that

$$\|\hat{f}^{\subseteq\mathcal{J}}\|_\infty = \max_{\mathbf{x}} \big|\hat{f}^{\subseteq\mathcal{J}}(\mathbf{x})\big| \leqslant \Big|\sum_{\mathcal{S}\subset\mathcal{J}} \|\psi_{\mathcal{S}}\|_\infty^2\Big| \leqslant 2^k c_k,$$

where c_k is the same term as in Theorem 4. As a result of the above equations and (33), we get the upper bound $\mathbb{E}[U(Z)] \leqslant U(\epsilon_2) + \delta_2\,(1 + 2^k c_k)$. Since this bound is independent of the choice of the k-element subset $\mathcal{J}$, then the inequality

$$\text{(II)} \leqslant U(\epsilon_2) + \delta_2\, U(1 + 2^k c_k).$$

holds as long as $n \geqslant n_2(\epsilon_2, \delta_2)$.

Tuning (ϵ_2, δ_2)**:** Now let $\gamma \in (0, \frac{1}{2})$ and set

$$\epsilon_2 = 2^{k/4} n^{-\gamma}, \qquad \delta_2 = 2kd^k \exp\{\frac{-n^{1-2\gamma}}{8c_k}\}.$$

With this choice $n_2(\epsilon_2, \delta_2) = n$. Further, we obtain in the following that

$$\text{(II)} \leqslant U(2^{k/4} n^{-\gamma}) + 2kd^k U(1 + 2^k c_k) \exp\{\frac{-n^{1-2\gamma}}{8c_k}\} \overset{(a)}{=} O(n^{-\gamma}),$$

where (a) holds, since the exponential term on the left-hand side converges to zero faster than $n^{-\gamma}$. Consequently, from the above equation, (32), and the inequality (31), we get

$$\mathbb{E}_{\mathcal{D}_n}[P_e(\hat{g})] - \mathsf{P}_{opt} \leqslant O(n^{-\gamma}) + O(n^{-\gamma}) = O(n^{-\gamma}).$$

This completes the proof.

G. Proof of the Technical Lemmas

G.1 Proof of Lemma 6

First, as score_1 is symmetric with respect to i, then α_i's are equal. Therefore, we need only to calculate α_1. Suppose $(\tilde{\mathbf{X}}(1), \tilde{Y}(1))$ is an i.i.d. copy of the first sample in the training data set $(\mathbf{X}(1), Y(1))$. Let $\widetilde{\text{score}}_1$ be the same as score_1 but with $(\mathbf{X}(1), Y(1))$, replaced with its i.i.d. copy.

Then, we need to find α_1 such that $|\text{score}_1(\mathcal{J}) - \widetilde{\text{score}}_1(\mathcal{J})| \leqslant \alpha_1$. Note that $\widehat{\psi}_{\mathcal{S}} \equiv \psi_{\mathcal{S}}$ which follows from the assumption that $\hat{\mu}_j = \mu_j$ and $\hat{\sigma}_j = \sigma_j$. From (8), and by replacing $\hat{f}_S = \frac{1}{n}\sum_j Y(j)\psi_{\mathcal{S}}(\mathbf{X}(j))$ we can write

$$\text{score}_1(\mathcal{J}) = \frac{1}{n-1}\sum_{i=1}^{n}\Big|\sum_{\mathcal{S}\subseteq\mathcal{J}}\sum_{j\neq i}\frac{1}{n}Y(j)\psi_{\mathcal{S}}(\mathbf{X}(j))\psi_{\mathcal{S}}(\mathbf{X}(i))\Big|.$$

Depending whether $i = 1$ or $j = 1$, the right-hand side of the above equation is a sum of the following three terms

$$\frac{1}{n(n-1)}\Big|\sum_{\mathcal{S}\subseteq\mathcal{J}}\sum_{j\neq 1} Y(j)\psi_{\mathcal{S}}(\mathbf{X}(j))\psi_{\mathcal{S}}(\mathbf{X}(1))\Big| + \frac{1}{n(n-1)}\sum_{i\neq 1}\Big|\sum_{\mathcal{S}\subseteq\mathcal{J}} Y(1)\psi_{\mathcal{S}}(\mathbf{X}(1))\psi_{\mathcal{S}}(\mathbf{X}(i))\Big|$$
$$+\frac{1}{n(n-1)}\sum_{i\neq 1}\Big|\sum_{\mathcal{S}\subseteq\mathcal{J}}\sum_{j\neq i,1} Y(j)\psi_{\mathcal{S}}(\mathbf{X}(j))\psi_{\mathcal{S}}(\mathbf{X}(i))\Big|.$$

The third term int he above equation is the same in score_1 and $\widetilde{\mathrm{score}}_1$. Therefore, using the triangle inequality, we obtain that

$$|\mathrm{score}_1(\mathcal{J}) - \widetilde{\mathrm{score}}_1(\mathcal{J})| \leqslant \frac{1}{n(n-1)}\sum_{\mathcal{S}\subseteq\mathcal{J}}\sum_{j\neq 1}\Big|Y(j)\psi_{\mathcal{S}}(\mathbf{X}(j))\Big| \underbrace{\Big|\psi_{\mathcal{S}}(\mathbf{X}(1)) - \psi_{\mathcal{S}}(\tilde{\mathbf{X}}(1))\Big|}_{\text{(I)}}$$
$$+\frac{1}{n(n-1)}\sum_{\mathcal{S}\subseteq\mathcal{J}}\sum_{i\neq 1}\underbrace{\Big|Y(1)\psi_{\mathcal{S}}(\mathbf{X}(1)) - \tilde{Y}(1)\psi_{\mathcal{S}}(\tilde{\mathbf{X}}(1))\Big|}_{\text{(II)}} \Big|\psi_{\mathcal{S}}(\mathbf{X}(i))\Big|. \tag{34}$$

Note that $\big|Y(j)\psi_{\mathcal{S}}(\mathbf{X}(j))\big| = \big|\psi_{\mathcal{S}}(\mathbf{X}(j))\big| \leqslant \|\psi_{\mathcal{S}}\|_\infty$, where we used the definition of ∞-norm and the fact that $Y(j) \in \{-1, 1\}$. Thus, from the triangle inequality, the term (I) in (34) satisfies (I) $\leqslant 2\|\psi_{\mathcal{S}}\|_\infty$. As for (II), we add an subtract $Y(1)\psi_{\mathcal{S}}(\tilde{\mathbf{X}}(1))$ and apply the triangle inequality. As a result, we have that

$$\text{(II)} \leqslant \big|Y(1)\big| \ \Big|\psi_{\mathcal{S}}(\mathbf{X}(1)) - \psi_{\mathcal{S}}(\tilde{\mathbf{X}}(1))\Big| + \big|Y(1) - \tilde{Y}(1)\big| \ \Big|\psi_{\mathcal{S}}(\tilde{\mathbf{X}}(1))\Big| \leqslant 4\|\psi_{\mathcal{S}}\|_\infty.$$

With the above argument and the inequality in (34), we obtain that

$$|\mathrm{score}_1(\mathcal{J}) - \widetilde{\mathrm{score}}_1(\mathcal{J})| \leqslant \frac{1}{n(n-1)}\sum_{\mathcal{S}\subseteq\mathcal{J}}\sum_{j\neq 1} 6\|\psi_{\mathcal{S}}\|_\infty^2 \leqslant \frac{6\ 2^k}{n}\|\psi_{\mathcal{S}}\|_\infty^2,$$

where the last inequality follows since $|\mathcal{J}| \leqslant k$.

G.2 Proof of Lemma 10

Recall from the proof of Lemma 1 that score_1 can be written as

$$\mathrm{score}_1(\mathcal{J}) = \frac{1}{n}\sum_i \Big|\hat{f}_{(i)}^{\subseteq\mathcal{J}}(\mathbf{X}(i))\Big|,$$

where $\hat{f}_{(i)}^{\subseteq\mathcal{J}}$ is defined as in (28) which is repeated below

$$\hat{f}_{(i)}^{\subseteq\mathcal{J}}(\mathbf{x}) \triangleq \frac{n}{n-1}\sum_{\mathcal{S}\subseteq\mathcal{J}}\Big(\hat{f}_{\mathcal{S}} - \frac{1}{n}Y(i)\prod_{j\in\mathcal{S}}\frac{X_j(i) - \hat{\mu}_j}{\hat{\sigma}_j}\Big)\widehat{\psi}_{\mathcal{S}}(\mathbf{x}).$$

Further, note that $\overline{\mathrm{score}}_1$ is the same as score_1 but with $\hat{\mu}_j = \mu_j$ and $\hat{\sigma}_j = \sigma_j$. Therefore, from the above relation, $\overline{\mathrm{score}}_1$ can also be written as $\overline{\mathrm{score}_1}(\mathcal{J}) = \frac{1}{n}\sum_i \Big|\bar{f}_{(i)}^{\subseteq\mathcal{J}}(\mathbf{X}(i))\Big|$, where

$$\bar{f}_{(i)}^{\subseteq\mathcal{J}}(\mathbf{x}) \triangleq \frac{n}{n-1}\sum_{\mathcal{S}\subseteq\mathcal{J}}\Big(\bar{f}_{\mathcal{S}} - \frac{1}{n}Y(i)\prod_{j\in\mathcal{S}}\frac{X_j(i) - \mu_j}{\sigma_j}\Big)\psi_{\mathcal{S}}(\mathbf{x}),$$

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

With the above definitions, from triangle inequality and the fact that $||a| - |b|| \leqslant |a-b|$, we obtain

$$\left|\overline{\text{score}_1}(\mathcal{J}) - \text{score}_1(\mathcal{J})\right| \leqslant \frac{1}{n}\sum_i \left|\bar{f}_{(i)}^{\subseteq\mathcal{J}}(\mathbf{X}(i)) - \hat{f}_{(i)}^{\subseteq\mathcal{J}}(\mathbf{X}(i))\right| \leqslant \|\bar{f}_{(1)}^{\subseteq\mathcal{J}} - \hat{f}_{(1)}^{\subseteq\mathcal{J}}\|_\infty,$$

where the last inequality follows by maximizing over all realizations of $\mathbf{X}(i)$ and the symmetricity with respect to i. Note that, $\bar{f}_{(1)}^{\subseteq\mathcal{J}}$ and $\hat{f}_{(1)}^{\subseteq\mathcal{J}}$ are, respectively, equal to $\bar{f}^{\subseteq J}$ and $\hat{f}^{\subseteq\mathcal{J}}$ when the first sample $(\mathbf{X}(1), Y(1))$ is removed from the training samples. Hence, Lemma 14 of Appendix B applies here and gives

$$\|\bar{f}_{(1)}^{\subseteq\mathcal{J}} - \hat{f}_{(1)}^{\subseteq\mathcal{J}}\|_\infty \leqslant \lambda(\epsilon_0),$$

where $\lambda(\epsilon_0) = O(k2^k c_k \epsilon_0)$ as $\epsilon_0 \to 0$. This completes the proof.

G.3 Proof of Lemma 11

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 $\|\psi_{\mathcal{S}}\|_\infty = \prod_{j\in\mathcal{S}}\frac{1+|\mu_j|}{\sigma_j}$.

$$\|\psi_{\mathcal{S}} - \widehat{\psi}_{\mathcal{S}}\|_\infty \leqslant |1 - b_{\mathcal{S}}|\|\psi_{\mathcal{S}}\|_\infty + \frac{k\epsilon}{\sigma_{\min}} b_{\mathcal{S}}(1+\epsilon)^k\|\psi_{\mathcal{S}}\|_\infty. \tag{35}$$

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

$$1-e^{k\epsilon} \leqslant 1-b_{\mathcal{S}} \leqslant k\epsilon \leqslant e^{k\epsilon}-1. \tag{36}$$

Lastly, combining (35) and (36) 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 $\|\psi_{\mathcal{S}}\|_\infty \leqslant \sqrt{c_k}$.

G.4 Proof of Lemma 14

Recall that the function $\bar{f}^{\subseteq J}$ is defined as

$$\bar{f}^{\subseteq J}(x^d) \triangleq \sum_{\mathcal{S}\subseteq\mathcal{J}} \bar{f}_S \psi_{\mathcal{S}}(x^d),$$

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 $\hat{f}^{\subseteq\mathcal{J}}$ and $\bar{f}^{\subseteq J}$ we obtain

$$\|\hat{f}^{\subseteq\mathcal{J}}-\bar{f}^{\subseteq J}\|_\infty \leqslant \sum_{S\subseteq J}\|\hat{f}_S\,\widehat{\psi}_{\mathcal{S}}-\bar{f}_S\,\psi_{S}\|_\infty. \tag{37}$$

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

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

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,

$$\|\hat{f}_S\,\widehat{\psi}_{\mathcal{S}}-\bar{f}_S\,\psi_{S}\|_\infty \leqslant \big(\|\widehat{\psi}_{\mathcal{S}}\|_\infty + |\bar{f}_S|\big)\|\widehat{\psi}_{\mathcal{S}}-\psi_{\mathcal{S}}\|_\infty. \tag{38}$$

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 Lemma 11 of Appendix B, which is restated as follows: Conditioned on B, $\|\psi_{\mathcal{S}} - \widehat{\psi}_{\mathcal{S}}\|_\infty \leqslant \gamma(\epsilon_0)$, almost surely, where $\gamma(\epsilon_0) = O(k\epsilon\sqrt{c_k})$ as $\epsilon_0 \to 0$.

Recall from (13), that c_k is defined as $c_k = \max_{\mathcal{S}:|\mathcal{S}|\leqslant k}\|\psi_{\mathcal{S}}\|_\infty^2$. Therefore, combining these bounds for the terms in (38) gives the following bound

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

Lastly, as a result of the above bound and the inequality (37),

$$\|\hat{f}^{\subseteq \mathcal{J}} - \bar{f}^{\subseteq J}\|_\infty \leqslant \lambda(\epsilon_0) \triangleq 2^k\big(2\sqrt{c_k}\gamma(\epsilon_0) + \gamma^2(\epsilon_0)\big).$$

It is not difficult to check that $\lambda(\epsilon_0) = O(k2^k c_k \epsilon_0)$ as $\epsilon_0 \to 0$.

References

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 19(3):357–367, 1967. doi: 10.2748/tmj/1178243286.

R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4):537–550, July 1994. doi: 10.1109/72.298224.

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

Jianbo Chen, Mitchell Stern, Martin J Wainwright, and Michael I Jordan. Kernel feature selection via conditional covariance minimization. In *Advances in Neural Information Processing Systems*, pages 6946–6955, 2017.

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.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. Wiley-Interscience, 2006.

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

Surbhi Goel and Adam R. Klivans. Learning neural networks with two nonlinear layers in polynomial time. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1470–1499, Phoenix, USA, 25–28 Jun 2019. PMLR. URL http://proceedings.mlr.press/v99/goel19b.html.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Lecture Notes in Computer Science*, pages 63–77. Springer Berlin Heidelberg, 2005. doi: 10.1007/11564089_7.

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.

Jeffrey C Jackson. An efficient membership-query algorithm for learning DNF with respect to the uniform distribution. *Journal of Computer and System Sciences*, 55(3):414–440, dec 1997. doi: 10.1006/jcss.1997.1533.

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.

Ron Kohavi and George H John. Wrappers for feature subset selection. *Artificial intelligence*, 97 (1-2):273–324, 1997.

Daphne Koller and Mehran Sahami. Toward optimal feature selection. Technical report, Stanford InfoLab, 1996.

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

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.

Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238, 2005.

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.

Wojciech Szpankowski. *Average case analysis of algorithms on sequences*, volume 50. John Wiley & Sons, 2011.

Gregory Valiant. Finding correlations in subquadratic time, with applications to learning parities and the closest pair problem. *Journal of the ACM*, 62(2):1–45, may 2015. doi: 10.1145/2728167.

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

Jorge R Vergara and Pablo A Estévez. A review of feature selection methods based on mutual information. *Neural computing and applications*, 24(1):175–186, 2014.

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.

Lei Yu and Huan Liu. Efficient feature selection via analysis of relevance and redundancy. *Journal of machine learning research*, 5(Oct):1205–1224, 2004.

J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, may 1977. doi: 10.1109/tit.1977.1055714.