IEEE Transactions on Information Theory

Sufficiently Informative and Relevant Features: An Information-theoretic and Fourier-based Characterization

Journal:	*IEEE Transactions on Information Theory*
Manuscript ID	IT-21-0232
Manuscript Type:	Regular Manuscript
Date Submitted by the Author:	03-Apr-2021
Complete List of Authors:	Heidari, Mohsen; Purdue University Sreedharan, Jithin; Wadhwani AI Shamir, Gil Szpankowski, Wojciech; Purdue University, Department of Computer Science
Keywords:	Discrete Fourier transforms, Unsupervised learning, Supervised learning, Feature Selection

Sufficiently Informative and Relevant Features: An Information-theoretic and Fourier-based Characterization

Mohsen Heidari *, Jithin Sreedharan †, Gil Shamir ‡ and Wojciech Szpankowski *,
* CS Dept., Purdue University, {mheidari, szpan}@purdue.edu
† Wadhwani AI, jithin.k.s@gmail.com
‡ Google Inc. gshamir@google.com

Abstract

A fundamental challenge in learning is the presence of nonlinear redundancies and dependencies in the data. To address this, we propose a Fourier-based approach to characterize feature redundancies, in unsupervised learning, and feature-label dependencies, in the supervised variant of the problem. We first develop a novel Fourier expansion for functions (more generally stochastic mappings) of correlated binary random variables. This is a generalization of the standard Fourier expansion on the Boolean cube beyond product probability spaces. As an important application of this analysis, we investigate learning with feature subset selection. In the unsupervised variant of this problem, we characterize feature redundancies via the Shannon entropy and group the features into sufficiently informative and redundant. Then, we make a connection to the proposed Fourier expansion and derive an upper bound on the joint entropy. Based on that, we propose a measure to quantify feature redundancies and present an unsupervised learning algorithm. We test our method on various real-world and synthetic datasets and demonstrate improvements on conventional unsupervised feature selection techniques.

Then, we investigate the supervised feature subset selection and reformulate it in the Fourier domain. Bridging the Bayesian error rate with the Fourier coefficients, we demonstrate that the Fourier expansion provides a powerful tool to characterize nonlinear feature-label dependencies. Further, we introduce a computationally efficient measure for selecting relevant features. Via a theoretical analysis, we show that our proposed measure finds provably *asymptotically optimal* feature subsets. Lastly, we present an algorithm based on this measure and via numerical experiments demonstrate its improvements on various supervised feature selection algorithms.

I. Introduction

A central challenge in learning with feature selection is to jointly identify nonlinear *redundancies* within the features and the *dependencies* in the feature-label relation. Many well-known feature selection approaches (supervised or unsupervised) are based on measures that capture only linear relations or focus on the features individually [1]–[3]. Kernel-based methods are an exception; however, are prohibitive in large datasets as the computational complexity of computing a kernel grows super linearly with the number of the samples [4]. Alternatively, information-theoretic metrics are powerful candidates in quantifying nonlinear dependencies among the random variables. However, estimating such quantities usually requires high sample complexity.

In this work, we take an alternative approach and adapt discrete Fourier analysis with information-theoretic measures. Hence, capturing nonlinear relations with low sample complexity, while avoiding kernel computations. The discrete Fourier expansion (on the Boolean cube) provides an essential tool to characterize different levels of "nonlinearities" in a function. In this expansion, any real-valued function on the Boolean cube can be written as a linear combination of monomials (*parities*) [5], [6]. Highly nonlinear functions have Fourier expansion with large coefficients for high-degree parities. Thus, the Fourier expansion is potentially a powerful tool in learning problems. However, limitations are making it impractical. First, it is developed for product probability spaces (mutually independent input variables). Secondly, this expansion is defined only for deterministic functions. These assumptions are too strong, as learning problems, often, involve correlated features with stochastic labeling. In this work, we make a connection between the two approaches and aim to address these challenges.

A. Main Contributions

We address the above challenges by developing a novel Fourier expansion for stochastic mappings of *correlated* binary random variables. Using this framework, we study feature selection, where the objective is to remove as many features as possible without significantly increasing classification loss. We demonstrate that the Fourier expansion provides a powerful tool to characterize nonlinear *redundancies* in features and nonlinear *dependencies* in the features-label relation. We propose two Fourier-based algorithms: one for supervised and one for the unsupervised variant of the problem. We use the Fourier expansion to provide a theoretical analysis and derive conditions under which our algorithms find the optimal feature subset. Further, through numerical experiments, we show that our algorithms outperform several well-known feature selection techniques. That said, the contributions of this paper are summarized below:

1) Fourier expansion for correlated random variables: We develop a generalized Fourier expansion for functions of *correlated binary* random variables (Proposition 1). For this purpose, we adopt a Gram-Schmidt-type orthogonalization and construct a set of orthogonal basis functions. Further, we adapt our Fourier expansion to the more general space of stochastic mappings (e.g., mappings from one probability space to another). To the best of our knowledge, this is the first generalization of the Fourier expansion for correlated binary random variables. Although this Fourier expansion is defined on the Boolean cube, our algorithms apply to non-binary features too. We view the Binary Fourier as a framework that captures a special class of nonlinearities — those characterized via the *parities*. Alternatively, we could generalize our Fourier expansion to discrete features and, based on it, design feature selection algorithms. However, such a generalization requires character theory, which is beyond the scope of this paper. We note that there are other forms of orthogonal decomposition including the Hoeffding-Sobel decomposition [7]–[9] and its generalization [9]. However, such decompositions are basis-free. Our Fourier expansion is defined by constructing a set of *orthonormal* basis functions which makes it suitable for feature selection.

2) Sufficiently informative: In the unsupervised setting, we take an information-theoretic perspective, and group the features into *redundant* and *sufficiently informative*. All the accessible *information* about the data can be captured from the latter group. More precisely, we define a subset of features as sufficiently informative when the joint entropy of all the input features equals that of the chosen subset. The former group is statistically a function of the latter, hence can be removed without affecting the learning's performance. This approach extends the notion of *Markov blanket* for "redundant" features to the unsupervised setting [10]–[12]. We then develop a characterization of sufficiently informative features based on our Fourier expansion (Theorem 1). Built upon this, we design an Unsupervised Fourier Feature Selection (UFFS) algorithm, which captures the redundant features in our new formulation. Instead of ranking the features, the UFFS finds redundant features and declares the rest of the features as sufficiently informative. Through comprehensive numerical experiments in Section V, we show that the UFFS outperforms conventional methods for unsupervised feature selection.

Our unsupervised feature selection algorithm is multi-variate (evaluating the dependencies of features jointly) and identifies non-linear dependencies. Whereas many well-known existing works are considered univariate or focus on the linear dependencies among the features [3]. Some common approaches in unsupervised feature selection are pseudo-label based, "column subset selection", and spectral/manifold based. Methods in the first approach such as [13], [14] attempt to generate pseudo-labels via certain clustering methods. However, such methods focus on linear transformations between features and the pseudo labels and ignore the nonlinear relations. The second approach, "column subset selection", assumes only linear dependencies among the features and solves an optimization problem that is similar to principal component analysis (PCA) [15], [16]. Although methods in the third approach, such as [17], capture nonlinear relations, they ignore the interaction between the features.

3) Measure for feature subset selection: When the feature-label probability distribution is known, features are ideally selected based on the Bayes misclassification rate as the measure. In practice, without knowledge of this distribution, given the training set, one approach (wrapper method) is to select feature subsets that minimize the empirical error rate of a given classifier [1]. Alternatively, to reduce the computational complexity, the common approach is to select feature subsets based on similarity or correlation measures [2].

Unlike conventional wrapper methods whose performance criteria depend on the given classifier, our measure for feature subset selection is independent of the classifier. For that, we first formulate the feature selection in an ideal setting as follows: given a parameter k, the objective is to find k features such that the misclassification rate of the Bayes classifier, restricted to k features, is minimized. We, then, reformulate this problem in the Fourier domain

and characterize the optimal feature subset. Building upon such a formulation, we develop a measure to evaluate feature subsets. We prove that when the features are binary, an exhaustive search based on this measure finds an asymptotically optimal feature subset. That is a feature subset whose Bayes misclassification rate is at most $O(n^{-\gamma})$, $\gamma \in (0, 1/2)$, larger than that of the optimal feature subset (Theorem 3).

4) Search algorithm for Fourier-based measure: Since the exhaustive search in the Fourier characterization is computationally expensive, we develop a search algorithm with fixed depth – given a depth parameter t, the idea is to evaluate only the feature subsets of size at most t. For numerical results, we usually set $t \leq 3$. With this approach, we propose the Supervised Fourier Feature Selection (SFFS) algorithm with computational complexity $O(n(d+\tilde{d}^t))$, where n is the number of the samples, d is the number of the features and $\tilde{d}$ is the number of *non-redundant* features. Based on our numerical experiments, $\tilde{d}$ is typically much smaller than d. Through our numerical experiments, we show that SFFS, even with $t = 1$ or 2, perform consistently better on a variety of datasets unlike several well-known feature selection algorithm (e.g., mRMR [18],, MI, RFS [19], CCM [20], and ReliefF [21]) (See Section V). We thus overcome two well-known demerits of wrapper methods for feature selection that limits their usage in practice – heavy dependency on the predictive performance of the learning algorithm and huge search space.

B. Related Works

Feature selection methods are usually classified into three main groups: wrappers, Filter, and embedded [1]. In the wrapper method, the feature subsets are evaluated directly by an induction algorithm. In embedded methods, feature selection is performed during the training process of the given learning algorithm. Such approaches are usually computationally expensive and, hence, prohibitive in large data sets. An alternative solution is the *filter* approach in which an intermediate measure, independent of the induction learning algorithm, is used to evaluate the feature subsets. Filter methods are preferred as they are computationally more efficient and relatively robust against overfitting. The challenge in this area, that remains open, is to design a computationally efficient measure that is provably related to the generalization loss. Several measures have been introduced in the literature. Well-known criteria for feature selection can be grouped into similarity-based measures (e.g., Pearson correlation, Fisher Score), information-theoretic measures [10], [18], [22]–[24], and Kernel-based measures [20], [25], [26]. Although correlation criteria are computationally more efficient, they usually are not able to detect *nonlinear* dependencies in features-label relations. Methods based on kernels can detect nonlinear dependencies. However, the computational complexity of computing a kernel grows super linearly, if not quadratic, with the number of the samples [4]. Mutual Information (MI) criteria, on the other hand, can detect nonlinear dependencies with lower computational complexity [24]. Also, mutual information can be used to bound the Bayes misclassification rate [27], [28]. However, estimating multi-variate mutual information is known to be a difficult task with high sample complexity.

Notations: As a shorthand, in this paper, for any natural number m, the set $\{1, 2, \cdots, m\}$ is denoted by $[m]$. Also, for any subset $\mathcal{J} \subseteq [d]$ with ordered elements $\{j_1, j_2, \cdots, j_k\}$, the vectors $(X_{j_1}, X_{j_2}, \cdots, X_{j_k})$, and $(x_{j_1}, x_{j_2}, \cdots, x_{j_k})$ are denoted, respectively, by $\mathbf{X}^{\mathcal{J}}$ and $\mathbf{x}^{\mathcal{J}}$. For any pair of functions g_1, g_2 denote $\langle g_1, g_2 \rangle \triangleq \mathbb{E}[g_1(\mathbf{X})g_2(\mathbf{X})]$. Further, we write $g_1 \equiv g_2$, when $g_1(x) = g_2(x)$ for all $x \in \mathcal{X}$.

II. Fourier Expansion for Correlated Random Variables

As an ingredient for our results, we first propose a Fourier expansion for functions of *correlated* binary features. We built upon this framework and propose our results in Section III and IV for feature selection.

A. Boolean Fourier expansion on product probability spaces

We start with a brief overview of the well-known Fourier expansion on Boolean cube [5], [6] that has applications in computational learning [29]–[34].

Let $\mathbf{X} = (X_1, X_2, ..., X_d)$ be a vector of mutually independent random variables taking values from a subset $\mathcal{X} \subset \mathbb{R}^d$. 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 defined as

$$\phi_{\mathcal{S}}(\mathbf{x}) \triangleq \prod_{i \in \mathcal{S}} \frac{x_i - \mu_i}{\sigma_i}, \qquad \text{for all } \mathbf{x} \in \mathbb{R}^d.$$

Since X_i's are mutually independent, the parities are orthonormal, that is $\mathbb{E}[\phi_{\mathcal{S}}(\mathbf{X})^2] = 1$ for any subset $\mathcal{S}$, and $\mathbb{E}[\phi_{\mathcal{S}}(\mathbf{X})\ \phi_{\mathcal{T}}(\mathbf{X})] = 0$ when $\mathcal{T} \neq \mathcal{S}$ (that is $\exists x \in \mathcal{T} \bigcup \mathcal{S}$ such that $x \notin \mathcal{T} \bigcap \mathcal{S}$). Under the assumption that $\mathcal{X} = \{-1, 1\}^d$, the parities form an orthonormal basis for the space of bounded functions [5]. That is, any bounded function $f : \{-1, 1\}^d \mapsto \mathbb{R}$ can be written as a linear combination of the form

$$f(\mathbf{x}) = \sum_{\mathcal{S} \subseteq [d]} \mathtt{f}_{\mathcal{S}}\ \phi_{\mathcal{S}}(\mathbf{x}), \quad \text{for all } \mathbf{x} \in \{-1, 1\}^d,$$

where $\mathtt{f}_{\mathcal{S}} \in \mathbb{R}$ are called the *Fourier coefficients* of f with respect to $D_{\mathbf{X}}$, the distribution of $\mathbf{X}$. Further, the Fourier coefficients can be computed as $\mathtt{f}_{\mathcal{S}} = \mathbb{E}[f(\mathbf{X})\phi_{\mathcal{S}}(\mathbf{X})]$, for all subsets $\mathcal{S} \subseteq [d]$.

B. Orthogonalization

We proceed by constructing a Fourier expansion for functions of correlated random variables. Note that, in a general probability space with correlated features, the standard Fourier expansion is no longer well-defined. Because the parities $\phi_{\mathcal{S}}$ are not necessarily orthogonal. That said, we construct our Fourier expansion by adopting a Gram-Schmidt-type procedure to make the parities *orthogonal*. Then, we use this basis to develop our Fourier expansion for functions of correlated random variables. Based on that, we establish the following result.

Proposition 1 (Correlated Fourier Expansion). *Let $D_{\mathbf{X}}$ be any probability distribution on $\{-1, 1\}^d$. Then there are a set of orthonormal parity functions $\psi_{\mathcal{S}}, \mathcal{S} \subseteq [d]$ such that any bounded function $g : \{-1, 1\}^d \mapsto \mathbb{R}$ is decomposed as*

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

for all for all $\mathbf{x} \in \{-1, 1\}^d$ except a measure-zero subset. Further, the coefficients $g_{\mathcal{S}}$ are unique and obtained from $g_{\mathcal{S}} = \mathbb{E}[g(\mathbf{X})\psi_{\mathcal{S}}(\mathbf{X})]$.

Proof: Let μ_j and σ_j be the mean and standard deviation of each input random variable $X_j, j \in [d]$. Suppose that these random variables are non-trivial, that is $\sigma_j > 0$ for all $j \in [d]$. The proof is based on an orthogonalization process that is explained in the following.

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

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

We apply the Gram-Schmidt process on $\phi_{\mathcal{S}_i}$ with the above ordering. The first orthogonalized parity is trivially given by $\psi_{\emptyset} \equiv 1$. The orthogonalized parity corresponding to the ith subset is obtained from the following operation:

$$\tilde{\psi}_{\mathcal{S}_i} \equiv \phi_{\mathcal{S}_i} - \sum_{j=1}^{i-1} \langle \psi_{\mathcal{S}_j}, \phi_{\mathcal{S}_i} \rangle\ \psi_{\mathcal{S}_j}, \qquad \psi_{\mathcal{S}_i} \equiv \begin{cases} \frac{\tilde{\psi}_{\mathcal{S}_i}}{\|\tilde{\psi}_{\mathcal{S}_i}\|_2} & \text{if } \|\tilde{\psi}_{\mathcal{S}_i}\|_2 > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

where $\|\tilde{\psi}_{\mathcal{S}_i}\|_2 = \sqrt{\langle \tilde{\psi}_{\mathcal{S}_i}, \tilde{\psi}_{\mathcal{S}_i} \rangle}$. By construction, the resulted nontrivial parities $\psi_{\mathcal{S}_i}$'s are orthonormal, that is $\langle \psi_{\mathcal{S}_i}, \psi_{\mathcal{S}_j} \rangle = 0$ for $i \neq j$ and $\langle \psi_{\mathcal{S}_i}, \psi_{\mathcal{S}_i} \rangle = 1$ if $\psi_{\mathcal{S}_i}$ is not trivial. The rest of the argument, given in Appendix A, follows by showing that $\psi_{\mathcal{S}}$'s span the space of all bounded functions. ■

Depending on the statistics of the features, the number of non-trivial parities $\psi_{\mathcal{S}_i}$ ranges from 1 to 2^d. On one extreme, if the features are mutually independent, then $\psi_{\mathcal{S}_i} = \phi_{\mathcal{S}_i}$. On the other extreme, if the features are trivial, then $\psi_{\mathcal{S}_i} = 0$ for $i > 1$, and hence there is only one non-trivial parity. Note also that different orderings for the subsets of $[d]$ result in different orthogonalized parities. We show in Section III-B that this ordering is beneficial to remove "redundant" features. Hence, unless otherwise stated, we use the ordering in (1).

Remark 1. *In contrary to our Fourier expansion, which is established only for binary features, the orthogonalization process is not restricted to such an assumption. Because, by construction, the orthogonalized parities are orthonormal for any value domain $\mathcal{X} \subset \mathbb{R}^d$. If $\mathcal{X} = \{-1, 1\}^d$, then the parities span the space of all functions on $\mathcal{X}$; otherwise they span a subspace of such functions. We clarify this in the following example.*

Example 1. *Set $d = 3$ and let X_1 and X_2 be independent random variables with Gaussian distribution $N(0, 1)$. Suppose $X_3 = X_1X_2$ with probability one. There are eight standard parities, one for each subset, as*

$$(1, x_1, x_2, x_1x_2, x_3, x_1x_3, x_2x_3, x_1x_2x_3).$$

By performing the orthogonalization process, as in (2), there are only four non-trivial orthogonalized parities as $\psi_\emptyset = 1, \psi_{\{1\}} = x_1, \psi_{\{2\}} = x_2$, and $\psi_{\{1,2\}} = x_1x_2$. The rest of the parities are zero, because $\|\tilde{\psi}_\mathcal{S}\|_2 = 0$ for any of the subsets $\{3\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}$. Now, suppose we change the relation of X_3 to $X_3 = \mathsf{sign}[X_1X_2]$. In this case, after the orthogonalization process, there are eight non-trivial parities. For instance, it is not difficult to check that $\tilde{\psi}_{\{3\}} = x_3 - \frac{2}{\pi}x_1x_2$. Hence, $\|\tilde{\psi}_{\{3\}}\|_2 > 0$, implying that $\psi_{\{3\}}$ is not redundant.

We view our binary Fourier as a framework that captures a special class of nonlinearities — those characterized via orthogonalized parities. Our numerical experiments confirm that such an approximation is sufficient to outperform many unsupervised feature selection methods on several data sets (see Table II). We further note that the orthogonalization process is different from the dimension reduction methods such as Kernel PCA. Because, unlike Kernel PCA, our process does not mix the features– it is a feature selection method.

III. Feature Selection in Known Statistics

In this section, we analyze the optimal feature selection for binary classification with $0-1$ loss function and in the ideal setting, where the underlying statistics of the features and the label are known. We first study the unsupervised variant of the problem in Section III-B. Then, in Section III-C, we analyze supervised feature selection in the Fourier domain.

A. Problem Formulation

The feature selection problem is formulated as follows. There are d features $\mathbf{X} = (X_1, X_2, ..., X_d)$ with label Y taking values from $\mathcal{X}^d$ and $\{-1, 1\}$, respectively. Suppose that $(\mathbf{X}, Y)$ are generated according to a fixed, but unknown, probability distribution D. Available are n independent and identically distributed (i.i.d.) instances $\mathcal{S}_n = \big\{(\mathbf{x}(i), y(i)), i = 1, 2, ..., n\big\}$ generated from D. We describe the feature selection problem by first defining the optimum feature subset and the minimum *misclassification* probability in the ideal setting, where D is known. For a feature subset $\mathcal{J} \subseteq [d]$, the minimum attainable mislabeling probability is obtained from

$$L_D(\mathcal{J}) = \min_{g \in \mathcal{G}_k} \mathbb{P}_{(\mathbf{X},Y)\sim D}\big\{Y \neq g(X^\mathcal{J})\big\}, \tag{3}$$

where $\mathcal{G}_k$ is the collection of all functions on $\mathcal{X}^k$. Given $k \leq d$, the optimum feature subset $\mathcal{J}^*$ and the minimum loss are defined as

$$\mathcal{J}^* = \underset{\mathcal{J}\subseteq[d],\,|\mathcal{J}|=k}{\arg\min}\; L_D(\mathcal{J}), \qquad\qquad L_D^*(k) = L_D(\mathcal{J}^*). \tag{4}$$

In agnostic settings, where only a training data set is available, the above optimization is infeasible to solve. Instead, an intermediate measure M_n is defined to evaluate feature subsets using the training instances. Then, feature selection using the measure M_n is modeled by the following optimization

$$\hat{\mathcal{J}}_n = \underset{\mathcal{T}\in\mathsf{T}_k}{\arg\min}\, M_n(\mathcal{T}),$$

where T_k is a collection of feature subsets with at most k-elements. Our objective is to propose a measure M_n so that mislabeling probability based on $\hat{\mathcal{J}}_n$ be as close as possible to that using $\mathcal{J}^*$.

B. Information Sufficiensy in Unsupervised Setting

We take an information-theoretic approach to identify feature redundancies in unsupervised feature selection. The goal is to define a measure to identify the features as "sufficiently informative" and "redundant". Intuitively, the former group contains all the *information* accessible from the features. The latter consists of the features that are a function of the "informative" features, and hence, can be removed from the data set.

Suppose $\mathcal{X}$ is a finite set. Then, we call a feature subset $\mathcal{J} \subseteq [d]$ "sufficiently informative", if $H(\mathbf{X}) = H(X^\mathcal{J})$, where H is the Shannon entropy. From the notion of Markov Blanket [23], one can show that $\mathcal{J}$ is a Markov blanket

for any feature in $\mathcal{J}^c$. Further, $\mathcal{J}$ being sufficiently informative immediately leads to $\mathcal{J}^c$ being redundant. Because, the condition $H(\mathbf{X}) = H(X^{\mathcal{J}})$ implies that there exists a mapping T, such that $X^{\mathcal{J}^c} = T(X^{\mathcal{J}})$, with probability one [28]. Hence, all the features not included in $\mathcal{J}$ can be removed. As there are multiple such $\mathcal{J}$'s, the objective is to find the smallest one[1]. Tolerating small amounts of imperfections, we formalize the above notion in the following.

Definition 1 (Sufficiently Informative). *For discrete features and $0 \leq \epsilon \leq 1$, a feature subset $\mathcal{J}$ is said to be ϵ-sufficiently informative, if $H(\mathbf{X}|X^{\mathcal{J}}) \leq \epsilon$. The feature subset $\mathcal{J}$ is sufficiently informative, if $H(\mathbf{X}|X^{\mathcal{J}}) = 0$. Such $\mathcal{J}$ is called minimal, if it has the minimum cardinality among all sufficiently informative feature subsets.*

We make a connection between the above definition and the orthogonalization process in (2). We employ this process to extract a sufficiently informative feature subset.

Theorem 1. *For $\epsilon \in [0, 1]$, let $\mathcal{J}_\epsilon \subseteq [d]$ be the set of all $j \in [d]$ such that $\|\tilde{\psi}_{\{j\}}\|_2 > \epsilon$. Then, $H(\mathbf{X}|X^{\mathcal{J}_\epsilon}) = dO(\epsilon)$ as $\epsilon \to 0$. Further, if the features take values from $\{-1, 1\}^d$, then there exists a permutation of the features so that $\mathcal{J}_\epsilon$ with $\epsilon = 0$ is a sufficiently informative subset that is minimal.*

Proof: We start by deriving an upper-bound on $H(\mathbf{X}|X^{\mathcal{J}_\epsilon})$ in terms of the orthogonalized parities in (2). Note that $H(\mathbf{X}|X^{\mathcal{J}_\epsilon}) = H(X^{\mathcal{J}_\epsilon^c}|X^{\mathcal{J}_\epsilon})$, where $\mathcal{J}_\epsilon^c$ is the complement of $\mathcal{J}_\epsilon$. Thus, from the chain rule, this quantity equals to $\sum_{i\in\mathcal{J}_\epsilon^c} H(X_i|X^{\mathcal{J}_\epsilon}, X^{i-1})$. As X_i is a discrete random variable and $\phi_{\{i\}}(\mathbf{x}) = (x_i - \mu_i)/\sigma_j$, then

$$H(X_i|X^{\mathcal{J}_\epsilon}, X^{i-1}) = H\big(\phi_{\{i\}}(\mathbf{X})|X^{\mathcal{J}_\epsilon}, X^{i-1}\big).$$

Since $\phi_{\{i\}}$ is the standard parity as in Section II, then, from the orthogonalization process in (2), we can write

$$\phi_{\{i\}} = \tilde{\psi}_{\{i\}} + \sum_{\mathcal{S}\subseteq[i-1]} \alpha_{\mathcal{S}}\psi_{\mathcal{S}},$$

where $\alpha_{\mathcal{S}} = \langle\phi_{\{i\}}, \psi_{\mathcal{S}}\rangle$. In this decomposition, the terms in the summation depend only on X^{i-1}. This is due to the particular ordering in (1). Therefore, we get the following upper-bound

$$H(\phi_{\{i\}}|X^{\mathcal{J}_\epsilon}, X^{i-1}) = H(\tilde{\psi}_{\{i\}}|X^{\mathcal{J}_\epsilon}, X^{i-1}) \leq H(\tilde{\psi}_{\{i\}}),$$

where the last inequality follows by removing the conditioning in the entropy. Lastly, adapting this bound for all $i \in \mathcal{J}_\epsilon^c$, we get the following upper-bound

$$H(X^d|X^{\mathcal{J}_\epsilon}) \leq \sum_{i\in\mathcal{J}_\epsilon^c} H(\tilde{\psi}_{\{i\}}). \tag{5}$$

Now for any $i \in [d]$ let

$$a_i = \min\{|\tilde{\psi}_{\{i\}}(\mathbf{x})| : \mathbf{x} \in \mathcal{X}^d, \tilde{\psi}_{\{i\}}(\mathbf{x}) \neq 0\},$$

and define $a_{\min} \triangleq \min_i a_i$. From the assumption in the statement of the theorem, $\|\tilde{\psi}_{\{i\}}\|_2^2 \leq \epsilon^2$, for all $i \in \mathcal{J}_\epsilon^c$. Then, from Markov's inequality, we obtain that

$$\mathbb{P}\{\tilde{\psi}_{\{i\}}(\mathbf{X}) \neq 0\} = \mathbb{P}\{|\tilde{\psi}_{\{i\}}(\mathbf{X})|^2 \geq a_i^2\} \leq \frac{\epsilon^2}{a_i^2} \leq \frac{\epsilon^2}{a_{\min}^2}.$$

Therefore, from grouping axiom for entropy [28], we have that

$$H(\tilde{\psi}_{\{i\}}) \leq h_b(\frac{\epsilon^2}{a_{\min}^2}) + \frac{\epsilon^2}{a_{\min}^2}\log_2|\mathcal{X}|, \tag{6}$$

where $h_b(\cdot)$ is the binary entropy function. When $\epsilon \leq a_{min}$, by combining (5) and (6), the following inequality holds:

$$H(X^d|X^{\mathcal{J}_\epsilon}) \leq (d - |\mathcal{J}_\epsilon|)\Big(h_b(\frac{\epsilon^2}{a_{min}^2}) + \frac{\epsilon^2}{a_{min}^2}\log_2|\mathcal{X}|\Big). \tag{7}$$

Note that the binary entropy satisfies the inequality $h_b(p) \leq 2\sqrt{p(1-p)}$. Hence, for fixed d, a_{min} and $\mathcal{X}$, the right-hand side of the above inequality is $dO(\epsilon)$ as $\epsilon \to 0$. This completes the proof for the first statement of the theorem.

[1]The set of all features is a trivial example of a sufficiently informative feature subset.

Next, we prove the second statement: "if the features take values from $\{-1,1\}^d$, then there exist a permutation of the features so that $\mathcal{J}_\epsilon$ with $\epsilon = 0$ is a *sufficiently informative* subset with minimum cardinality." Note that, from Definition 1, the subset $\mathcal{J}_0 = \mathcal{J}_{\epsilon=0}$ is sufficiently informative. This is because for any $j \notin \mathcal{J}_0$ the parity $\tilde{\psi}_{\{j\}}(\mathbf{X}) = 0$ with probability one. Therefore, from (2), the standard parity $\phi_{\{j\}}$ is a function of $(X_1, X_2, ..., X_{j-1})$. Implying that X_j is a function of $(X_1, X_2, ..., X_{j-1})$. Hence, $H(X^d|X^{\mathcal{J}_0}) = 0$. It remains to prove that $\mathcal{J}_0$ is minimal when the features take values from $\{-1,1\}^d$ and are permuted appropriately.

Let $\mathcal{A} \subseteq [d]$ be a sufficiently informative subset with minimum cardinality as in Definition 1. Consider a permutation of the features such that the first $|\mathcal{A}|$ features are from $\mathcal{A}$. We perform the orthogonalization process on the permuted features. Let $\mathcal{J}_0$ be the subset for which $\|\tilde{\psi}_{\{i\}}\|_2 > 0$. We show that $|\mathcal{J}_0| = |\mathcal{A}|$, implying that $\mathcal{J}_0$ is minimal. By contradiction, suppose $|\mathcal{A}| < |\mathcal{J}_0|$. Consider the Hilbert space of bounded functions $f : \{-1,1\}^d \mapsto \mathbb{R}$ with inner product defined as $\langle f, g\rangle = \mathbb{E}[f(\mathbf{X})g(\mathbf{X})]$. Since $\mathcal{A}$ is sufficiently informative, there exists a mapping Γ such that $X^d = \Gamma(X^{\mathcal{A}})$ with probability one. Therefore, any function $f(X^d)$ can be written as $f(X^d) = f(\Gamma(X^{\mathcal{A}}))$. This implies that the effective dimension of the Hilbert space is at most $2^{|\mathcal{A}|}$. However, there are $|\mathcal{J}_0|$ non-trivial parities $\psi_{\mathcal{S}_i}$ that, from Proposition 1, form an orthonormal basis for this Hilbert space. Hence, the dimension of this space cannot be less than $2^{|\mathcal{J}_0|}$. This is a contradiction and completes the proof. ■

As a result of Theorem 1, $\|\tilde{\psi}_{\{j\}}\|_2$ can be viewed as a measure of the *redundancy* of each feature and that the orthogonalization procedure can remove them. We use this measure for unsupervised feature selection, where n i.i.d. instances $\{\mathbf{x}(i), i \in [n]\}$, with possible correlations across feature dimension, are available. The idea is to perform the orthogonalization process as in (2) and find the features j for which $\|\tilde{\psi}_{\{j\}}\|_2$ is smaller than a threshold ϵ. These features are declared as redundant. Based on this idea we present our unsupervised algorithm (UFFS) in Section IV-A.

C. Supervised Feature Selection

For the supervised version of the problem, we proceed by developing a representation of $L_D^*(k)$ in the Fourier domain. For that, we characterize the Bayes predictor via the Fourier expansion. The Fourier expansion in the previous section is defined for deterministic functions. The label Y in the feature selection problem is not necessarily a function of the features, rather it is a stochastic mapping. Therefore, we need to extend the Fourier expansion to such mappings.

Let $\mathcal{J}$ be a subset of the selected features. Note that the Bayes predictor of Y from the observation $\mathbf{x}^{\mathcal{J}}$ is given by $g^*(\mathbf{x}^{\mathcal{J}}) = \mathsf{sign}\big[\mathbb{E}[Y|\mathbf{x}^{\mathcal{J}}]\big]$. As a key ingredient in our characterization, we need to define the notion of *projection onto a feature subset*.

Denote the elements of $\mathcal{J}$, in the ascending order, as $j_1 < j_2 < \cdots < j_k$. Fix the following ordering of subsets of $\mathcal{J}$:

$$\emptyset, \{j_1\}, \{j_2\}, \{j_1, j_2\}, \{j_3\}, \{j_1, j_3\}, \{j_2, j_3\}, \{j_1, j_2, j_3\}, \cdots, \{j_1, j_2, ..., j_k\}.$$

Apply the orthogonalization process with respect to this ordering and to all the parities $\psi_{\mathcal{S}}$ with $\mathcal{S} \subseteq \mathcal{J}$. Let $\psi_{\mathcal{S}_i}$, $i = 1, 2, ..., 2^k$ be the resulted orthogonalized parities. This process is called orthogonalization with respect to the feature subset $\mathcal{J}$.

Lemma 1. *Given $\mathcal{J} \subseteq [d]$, the Bayes predictor of the label Y from observation $\mathbf{x}^{\mathcal{J}}$ is given by* $\mathsf{sign}[f^{\subseteq\mathcal{J}}(\mathbf{x}^{\mathcal{J}})]$, *where $f^{\subseteq\mathcal{J}}$ is a function admitting the following Fourier expansion*

$$f^{\subseteq\mathcal{J}}(\mathbf{x}^{\mathcal{J}}) \triangleq \sum_{\mathcal{S}\subseteq\mathcal{J}} f_{\mathcal{S}}\psi_{\mathcal{S}}(\mathbf{x}^{\mathcal{J}}), \tag{8}$$

where $\psi_{\mathcal{S}}$'s are the parities resulted from the orthogonalization with respect to $\mathcal{J}$ and $f_{\mathcal{S}} = \mathbb{E}_D[Y\psi_{\mathcal{S}}(\mathbf{X})]$.

Proof: Note that the MMSE estimator of Y from $\mathbf{X}^{\mathcal{J}}$ is $\mathbb{E}[Y|\mathbf{x}^{\mathcal{J}}]$. Since, Y take values from $\{-1,1\}$, then the Bayes predictor is obtained from $\mathsf{sign}[\mathbb{E}[Y|\mathbf{x}^{\mathcal{J}}]]$. Define the function $e(\mathbf{x}) = \mathbb{E}[Y|\mathbf{x}^{\mathcal{J}}]$. Note that e can be viewed as a real-valued function on $\{-1,1\}^k$. In addition, we can apply Proposition 1 on coordinates $j \in \mathcal{J}$ and with $d = k$. As a result, e has a Fourier expansion of the form

$$e(\mathbf{x}) = \sum_{\mathcal{S}\subseteq\mathcal{J}} \hat{e}_{\mathcal{S}}\psi_{\mathcal{S}}(\mathbf{x}),$$

where $\psi_{\mathcal{S}}$'s are the orthogonalized parities w.r.t $\mathcal{J}$, and $\hat{e}_{\mathcal{S}} = \langle e, \psi_{\mathcal{S}} \rangle$. Then, for each $\mathcal{S} \subseteq \mathcal{J}$, we obtain that

$$\begin{aligned}\hat{e}_{\mathcal{S}} = \mathbb{E}[e(\mathbf{X})\psi_{\mathcal{S}}(\mathbf{X})] &= \mathbb{E}\Big[\mathbb{E}[Y|X^{\mathcal{J}}]\psi_{\mathcal{S}}(\mathbf{X})\Big] \\ &= \mathbb{E}\Big[\mathbb{E}[Y\psi_{\mathcal{S}}(\mathbf{X})|X^{\mathcal{J}}]\Big] \\ &= \mathbb{E}\Big[Y\psi_{\mathcal{S}}(\mathbf{X})\Big] \\ &= f_{\mathcal{S}}\end{aligned}$$

where the second equality holds as $\psi_{\mathcal{S}}(\mathbf{X})$ depends only on $X_j, j \in \mathcal{S}$, and the last equality follows from the definition of $f_{\mathcal{S}}$ as in the statement of the Lemma. Therefore, e admits the same Fourier expansion as $f^{\subseteq\mathcal{J}}$. With that the proof is complete. ■

Theorem 2. *The minimum attainable misclassification probability equals to*

$$L_D^*(k) = \frac{1}{2}\left[1 - \max_{\mathcal{J}\subseteq[d],\,|\mathcal{J}|=k} \|f^{\subseteq\mathcal{J}}\|_{1,D}\right]. \tag{9}$$

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

Proof: Fix a subset $\mathcal{J} \subseteq [d]$ with k elements. From the definition of $L_D^*(k)$ in (4), any predictor is a function $g : \{-1,1\}^k \mapsto \{-1,1\}$. Here, g represents a predictor of Y from $\mathbf{X}^{\mathcal{J}}$. Since Y and $g(\mathbf{X}^{\mathcal{J}})$ take values from $\{-1,1\}$, then,

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

Note that given $\mathcal{J}$, the above probability is minimized by the Bayes estimator. Further, such an estimator is given by $\mathsf{sign}\big[\mathbb{E}[Y|\mathbf{x}^{\mathcal{J}}]\big]$, for all $\mathbf{x}^{\mathcal{J}} \in \{-1,1\}^k$. Hence, it suffices to calculate the above misclassification probability for $g \equiv \mathsf{sign}\big[\mathbb{E}[Y|\mathbf{x}^{\mathcal{J}}]\big]$. For that, in the following, we calculate the expectation $\mathbb{E}[Y g(\mathbf{X}^{\mathcal{J}})]$ for $g = \mathsf{sign}\big[\mathbb{E}[Y|\mathbf{x}^{\mathcal{J}}]\big]$.

$$\begin{aligned}\mathbb{E}[Y g(\mathbf{X}^{\mathcal{J}})] &\overset{(a)}{=} \mathbb{E}\Big[\mathbb{E}[Y g(\mathbf{X}^{\mathcal{J}})|\mathbf{X}^{\mathcal{J}}]\Big] \overset{(b)}{=} \mathbb{E}\Big[\mathbb{E}[Y|X^{\mathcal{J}}]g(\mathbf{X}^{\mathcal{J}})\Big] \\ &\overset{(c)}{=} \mathbb{E}\Big[\big|\mathbb{E}[Y|\mathbf{X}^{\mathcal{J}}]\big|\Big] \overset{(d)}{=} \mathbb{E}\Big[\big|f^{\subseteq\mathcal{J}}(\mathbf{X})\big|\Big] \\ &= \|f^{\subseteq\mathcal{J}}\|_1,\end{aligned}$$

where (a) follows from the *law of total probability*, (b) holds because g is a function of $\mathbf{X}^{\mathcal{J}}$, equality (c) follows by replacing g with $\mathsf{sign}\big[\mathbb{E}[Y|\mathbf{x}^{\mathcal{J}}]\big]$, and lastly, (d) holds because $f^{\subseteq\mathcal{J}}(\mathbf{X}) = \mathbb{E}[Y|\mathbf{X}^{\mathcal{J}}]$. This equality is shown in Lemma 1. As a result, the minimum misclassification probability for a fixed subset $\mathcal{J}$ is equal to $\frac{1}{2} - \frac{1}{2}\|f^{\subseteq\mathcal{J}}\|_1$. Hence, optimizing over all k-element subsets $\mathcal{J}$ gives the desired expression and completes the proof. ■

D. A Measure for Feature Selection

The previous section provides the characterization in the ideal setting where D is known. We leverage this characterization to the agnostic setting, where only the training set $\mathcal{S}_n$ is available. For that, we derive a measure for selecting the features.

Based on Theorem 2, we define $M_n(\mathcal{J})$ to be an empirical estimate of $\|f^{\subseteq\mathcal{J}}\|_1$. Therefore, if the estimations are accurate enough, then maximizing M_n leads to a feature subset $\hat{\mathcal{J}}$ for which $L_D(\hat{\mathcal{J}})$ is close to the optimal loss $L_D(\mathcal{J}^*)$ as in (4). In what follows, we describe the derivation of M_n in three steps:

Step 1: We perform an empirical orthogonalization. Let $\hat{D}_n$ be the empirical distribution of the training set $\mathcal{S}_n$, that is $\hat{D}_n(\mathbf{x}, y) = \frac{1}{n}$ if $(\mathbf{x}, y) \in \mathcal{S}_n$, and zero otherwise. We get the empirical version of our results by replacing D with $\hat{D}_n$. In particular, Proposition 1, and the orthogonalization in (2). Let $\widehat{\psi}_{\mathcal{S}}$ denote the parities resulted from the orthogonalization with respect to $\hat{D}_n$. By construction, these functions are orthonormal with respect to $\hat{D}_n$.

Step 2: Next, we construct the estimate of the function $f^{\subseteq\mathcal{J}}$ as in (8). Let

$$\hat{f}_{\mathcal{S}} \triangleq \mathbb{E}_{\hat{D}_n}[Y\widehat{\psi}_{\mathcal{S}}(\mathbf{X})] = \frac{1}{n}\sum_i y_i \widehat{\psi}_{\mathcal{S}}(\mathbf{x}_i).$$

Once the empirical parities and the Fourier coefficients $\hat{f}_{\mathcal{S}}$ are calculated, the estimation of the projection function $f^{\subseteq\mathcal{J}}$ is obtained from the equation

$$\hat{f}^{\subseteq\mathcal{J}}(\mathbf{x}) \triangleq \sum_{\mathcal{S}\subseteq\mathcal{J}} \hat{f}_{\mathcal{S}}\ \widehat{\psi}_{\mathcal{S}}(\mathbf{x}).$$

Step 3: The next step is to approximate $\|\hat{f}^{\subseteq\mathcal{J}}\|_1$. By definition, $\|\hat{f}^{\subseteq\mathcal{J}}\|_1 \triangleq \mathbb{E}_{\mathbf{X}}[|\hat{f}^{\subseteq\mathcal{J}}(\mathbf{X})|]$. Hence, naturally, the estimation of this quantity is obtained by the empirical averaging

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

However, 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. That said, we make a correction and define our measure M_n as in the following

$$M_n(\mathcal{J}) = \widehat{\|f^{\subseteq\mathcal{J}}\|_1} \triangleq \frac{1}{n-1}\sum_{i=1}^{n}\ \Big|\ \sum_{\mathcal{S}\subseteq\mathcal{J}} \hat{f}_{\mathcal{S}}\widehat{\psi}_{\mathcal{S}}(\mathbf{x}(i)) - \frac{1}{n}y(i)\big(\widehat{\psi}_{\mathcal{S}}(\mathbf{x}(i))\big)^2\ \Big|. \tag{10}$$

This correction is done by subtracting the quantity $\frac{1}{n}y(i)\big(\widehat{\psi}_{\mathcal{S}}(\mathbf{x}(i))\big)^2$ in the above.

We use $M_n(\mathcal{J})$ as an estimate of $\|f^{\subseteq\mathcal{J}}\|_1$. We shown that this estimator is asymptotically unbiased, that is $\Big|\ \mathbb{E}\big[M_n(\mathcal{J})\ \big] - \|f^{\subseteq\mathcal{J}}\|_1\Big| \to 0$ as $n\to\infty$.

We conclude this section by presenting our analysis for the proposed measure. We note here that in our problem the function $f^{\subseteq\mathcal{J}}$ is not necessarily bounded. Hence, the standard concentration inequalities such as Rademacher complexity do not apply. We address this issue and prove the following theorem in Appendix B.

Theorem 3. *Let $\hat{\mathcal{J}}_n$ be the feature subset maximizing M_n over all binary feature subsets with k elements. Let $\mathcal{J}^*$ be the optimum feature subset as in (9). Then, with probability at least $(1-\delta)$, the following bound holds*

$$L_D(\hat{\mathcal{J}}_n) \le L_D(\mathcal{J}^*) + \sqrt{\frac{\lambda(k)}{n-1}\log(\frac{d}{\delta})} + O(n^{-\gamma}),$$

where $\gamma\in(0,1/2)$ and $\lambda(k) = 8\ k2^{2k}c_k^2$, with $c_k \triangleq \max_{\mathcal{S}\subseteq[d],|\mathcal{S}|\le k}\|\psi_{\mathcal{S}}\|_\infty^2$.

The exhaustive search over all k-element feature subsets is computationally expensive. Hence, in the next section, we present a few approximation methods and propose our algorithm.

IV. Proposed Algorithms

We build upon our Fourier expansion and propose our Supervised Fourier Feature Selection (SFFS) and Unsupervised Fourier Feature Selection (UFFS) algorithms. To reduce the computational complexity, we propose a few approximations.

A. Unsupervised Feature Selection

As for the algorithm, two issues need to be addressed: 1) estimation of $\|\tilde{\psi}_{\{j\}}\|_2$ from the training instances and 2) the orthogonalization is NP-hard, as there are 2^d feature subsets.

To address the first issue, we propose a recursive formula to perform the orthogonalization. Let $b_{j,i} = \langle\phi_{\mathcal{S}_j},\phi_{\mathcal{S}_i}\rangle$, and define $a_{j,i} = \langle\psi_{\mathcal{S}_j},\phi_{\mathcal{S}_i}\rangle$. With this notation, (2) can be written as

$$\tilde{\psi}_{\mathcal{S}_i} = \phi_{\mathcal{S}_i} - \sum_{j<i} a_{j,i}\psi_{\mathcal{S}_j}.$$

Hence, we only need to compute $a_{j,i}$'s. Note that since $\psi_{\mathcal{S}_i}$'s are orthonormal, then we obtain that

$$\|\tilde{\psi}_{\mathcal{S}_i}\|_2^2 = b_{i,i} - \sum_{j<i} a_{j,i}^2.$$

Further, the coefficients $a_{j,i}$ can be calculated recursively as

$$a_{j,i} = \frac{1}{\sqrt{b_{j,j} - \sum_{r<j} a_{r,j}^2}} \Big(b_{j,i} - \sum_{\ell<j} a_{\ell,j} a_{\ell,i} \Big). \tag{11}$$

With this formulas, we first compute an empirical estimate of $b_{j,i}$'s, denoted by $\hat{b}_{j,i}$. Hence, given the training samples, we compute

$$\hat{b}_{j,i} = \frac{1}{n} \sum_{\ell} \phi_{\mathcal{S}_j}(\mathbf{x}_\ell) \phi_{\mathcal{S}_i}(\mathbf{x}_\ell).$$

Then, we compute an estimation of $a_{j,i}$'s (denoted by $\hat{a}_{j,i}$) by calculating (11) with $b_{j,i}$ and $a_{j,i}$ replaced by $\hat{b}_{j,i}$ and $\hat{a}_{j,i}$, receptively.

Note that we further approximate (2) by declaring $\widehat{\psi}_{\mathcal{S}}$ as trivial, if $\|\tilde{\psi}_{\mathcal{S}}\|_2 \leq \epsilon$, where $\epsilon \in (0,1)$ is a parameter. As a result, we declare a feature j to be redundant if $\|\tilde{\psi}_{\{j\}}\|_2 \leq \epsilon$.

Fixed-depth search: To address the second issue, we propose to limit the size of the subsets involved in the orthogonalization. Given a parameter $t \leq d$, the orthogonalization is performed only on feature subsets of size at most t. For that we use the standard ordering as in (1), but restricted to subsets of size at most t. For most practical purposes $t \leq 3$. With that the search space is reduced to $\binom{d}{t}$. Further, this limitation is sufficient when the dependencies across the features are bounded to at most t features.

Algorithm 1: Unsupervised Fourier Feature Selection (UFFS)

Input: n training samples $\mathbf{x}_i \in \mathbb{R}^d$, depth parameter $t \leq d$, and redundancy threshold $\epsilon \in (0,1)$
Output: Features' measures $\mathsf{norm}(j), j = 1, 2, \ldots d$

Procedure `FourierOrth(`t`)`**:**
- Compute the empirical mean $\hat{\mu}_j$ and standard deviation $\hat{\sigma}_j$ of each feature.
- Generate all subsets $\mathcal{S}_i \subseteq [d]$ with size at most t and with the standard ordering as in (1). Compute the matrix $\hat{\mathbf{B}}$ with elements:

$$\hat{b}_{j,i} \leftarrow \frac{1}{n} \sum_{l=1}^{n} \Big[\prod_{u \in \mathcal{S}_j} \frac{x_{lu} - \hat{\mu}_u}{\hat{\sigma}_u} \Big] \Big[\prod_{v \in \mathcal{S}_i} \frac{x_{lv} - \hat{\mu}_v}{\hat{\sigma}_v} \Big].$$

- Set $\hat{\mathbf{A}} \leftarrow \hat{\mathbf{B}}$
- **for** *row* j *of* $\hat{\mathbf{A}}$ **do**
  - update the jth row: $\hat{\mathbf{A}}_{j,*} \leftarrow \hat{\mathbf{A}}_{j,*} - \sum_{\ell<j} \hat{a}_{\ell,j} \hat{\mathbf{A}}_{\ell,*}$
  - Compute $\mathsf{norm}(\mathcal{S}_j) \leftarrow \sqrt{[\hat{b}_{j,j} - \sum_{r<j} \hat{a}_{r,j}^2]^+}$
  - **if** $\mathsf{norm}(\mathcal{S}_j) \leq \epsilon$ **then**
    - Set the jth row of $\hat{\mathbf{A}}$ zero: $\hat{\mathbf{A}}_{j,*} \leftarrow \mathbf{0}$
    - **else**
      - Normalize the jth row: $\hat{\mathbf{A}}_{j,*} \leftarrow \frac{\hat{\mathbf{A}}_{j,*}}{\mathsf{norm}(\mathcal{S}_j)}$
- Declare all $j \in [d]$ with $\mathsf{norm}(j) \geq \epsilon$ as non-redundant.

Remark 2. *One can provide a consistency analysis on asymptotic performance of* `FourierOrth`*. More precisely, let* $\hat{\mathcal{J}}_n$ *be the set of features declared* `FourierOrth(`t`)`*. Then, there exists a function* λ *such that* $\hat{\mathcal{J}}_n$ *is a* $\lambda(\epsilon, n)$*-sufficiently informative and* $\lim_{n \mapsto \infty} \lim_{\epsilon \mapsto 0} \lambda(\epsilon, n) = 0$.

The main idea to show the above statement is as follows: Let $\hat{D}_n$ denote the empirical distribution of the training samples. That is $\hat{D}_n(\mathbf{x}) = \frac{1}{n}$ if $\mathbf{x}$ is in the training set; otherwise $\hat{D}_n(\mathbf{x}) = 0$. We apply Theorem 1 with $\hat{D}_n$ as the distribution of the random variables in the statement. As a result, $\hat{H}_n(\mathbf{X}|\mathbf{X}^{\hat{\mathcal{J}}_n}) = O(\epsilon)$, where $\hat{H}_n$ is the conditional entropy calculated over $\hat{D}_n$ as the distribution. The rest of the argument is a concentration analysis. In particular,

from McDiarmid's inequality and the continuity of entropy, we can show that $\hat{H}_n$ converges to the true conditional entropy.

Feature partitioning: For large dimensional data sets, we can further reduce the complexity by partitioning the features. We randomly partition the features into multiple groups of approximately equal size (say m features each). Then, we perform `FourierOrth` on each group, and remove the redundant features within it. With this approach, the computational complexity with depth parameter t and group size m is $O(n\frac{d}{m}m^{2t})$. The parameters m and t are chosen depending on the limitations on running time. These parameters are typically chosen independent of the size of the data set. For instance, we choose $t \leq 3$ and $m = 40$ for our numerical results. As a result, we obtain a complexity linear in the size of the data set. We present our experimental results in Section V.

B. Supervised Feature Selection

The measure M_n captures the joint effect of the candidate feature subsets. However, to further reduce the running time, we adopt the fixed-depth search here. Hence, instead of searching over all k-element feature subsets, we choose to search over all t element subsets (say $t = 3$). For that we calculate $M_n(\mathcal{T})$ for all t element feature subsets. Next, we rank these subsets in the descending order base on M_n. Then, starting from the top, we take the union of $\mathcal{T}$'s to obtain a k-element feature subset. With this approach, we present Algorithm 2. Note that with $t = 1$, our search algorithm reduces to feature-ranking method.

Algorithm 2: Supervised Fourier Feature Selection (SFFS)

Input: n training samples $(\mathbf{x}_i, y_i)$, desired number of features k, and the depth parameter $t \leq k$
Output: Feature subset $\hat{\mathcal{J}}_n$

Run `FourierOrth(`t`)` to get the non-trivial parities and non-redundant features.
Procedure `SFFS(`t, k`)`**:**
  Rank all t-variable features subsets $\mathcal{T}$ according to M_n as in (10).
  If $\mathcal{T}_i$ are the subsets in the descending order, set $\hat{\mathcal{J}}_n = \bigcup_{i=1}^{r} \mathcal{T}_i$, where r chosen such that the union has k different elements.
  Return $\hat{\mathcal{J}}_n$

With that, the computational complexity of our SFFS algorithm for a fixed parameter t is $O(n\tilde{d}^t)$, where $\tilde{d}$ is the number of non-trivial features declared from the UFFS algorithm. Our numerical results verifies that usually $\tilde{d}$ is much smaller than d, see Table II. As a result, the overall computational complexity of the algorithm $O(nd + n\tilde{d}^t)$ which is $O(nd)$ for large data sets.

V. Numerical Experiments

We now compare the performance of our UFFS and SFFS algorithms (Algorithm 1 and 2) with several well-known methods for unsupervised and supervised feature selection. Our numerical results are presented in two parts: unsupervised and supervised setting. We tested the algorithms on several real-world data sets as given in Table I. These data sets are benchmarks and taken from [2] and the UCI repository [35]. In addition, we generated five synthetic data sets: S1, S2, S3 are for unsupervised feature selection, and E1, E2 are for the supervised variant of the problem. These data sets are described below:

A. Synthetic data sets

For unsupervised feature selection, we generated three data sets, denoted by S1, S2, and S3. Each data set has 30 features: 10 informative denoted by $(X_1, X_2, ..., X_{10})$, 10 nonlinear redundant $(X_{11}, X_{12}, ..., X_{20})$, and 10 linearly redundant $(X_{21}, X_{12}, ..., X_{30})$. The informative features are generated according to three distributions, one for each data set. The distribution for S1 is $N(0, \mathbf{I}_{10})$, for S2 is uniform distribution over $[-1, 1]^{10}$, and for S3 is uniform distribution over $\{-1, 1\}^{10}$. Each nonlinear redundant feature is generated from $X_j = 3X_{i_1}X_{i_2}X_{i_3}$, where $j = 11, 12, ..., 20$, and i_1, i_2, i_3 are randomly and uniformly selected from $\{1, 2, ..., 10\}$. The linearly redundant features are generated from $X_j = \sum_{l=1}^{5} a_{j,l}X_{i_l}$, where i_l's are selected randomly from $\{1, 2, ..., 10\}$ and $a_{j,l} \sim \text{Unif}(0, 1)$. We

TABLE I: Properties of the tested data sets.

Data set	S1	S2	S3	E1	E2	USPS	Isolet	COIL20	Covertype	Australian	Musk	ALL	AML
Features	30	30	30	20	20	256	617	1024	46	14	166	7128	
Samples	1000	1000	1000	1000	1000	9298	1560	1440	581	690	467	72	

TABLE II: Comparison of unsupervised algorithms and number of non-redundant features declared by UFFS ($\tilde{d}$).

	S1	S2	S3	USPS	Isolet	COIL20	Covertype	Australian	Musk	ALL	AML
No FS	77.9	75.0	87.0	97.3	92.8	98.8	75.6	84.9	92.2	94.3	
UFFS $\tilde{d}$	11	12	11	93	309	331	34	12	35	39	
UFFS	**80.3**	**76.8**	**86.2**	**97.0**	**91.7**	**98.8**	**76.9**	**85.1**	**85.7**	**97.1**	
LS	55.1	61.2	71.0	95.6	88.6	98.9	72.8	85.4	84.5	97.2	
MCFC	56.6	59.0	65.8	93.9	90.1	94.0	72.3	84.8	84.2	95.9	
UDFS	64.0	60.6	64.3	80.8	90.2	98.0	72.0	84.9	80	86.2	

use the above redundancy model for each data set. For the sake of performance comparison, we add a labeling to the above data sets. However, the labels are not revealed to the algorithms. We generate a fixed but randomly generated labeling function $f(\mathbf{X})$ on $\mathbb{R}^{10}$. This function is the sign of the following randomly generated polynomial in $\mathbb{R}^{10}$:

$$f(\mathbf{x}) = \mathsf{sign}\Big[\prod_{1\le j\le 3}\big(b_{0,j} + \sum_{1\le i\le 10} b_{i,j}x_i\big)\Big],$$

where $b_{i,j} \sim \text{Unif}(0,1)$ and mutually independent.

For supervised feature selection, we generate two data sets denoted by E1 and E2. Each data set consists of 1000 samples each having 20 features distributed according to uniform distribution over $\{-1,1\}^{20}$ for E1 and $N(0,\mathbf{I}_{20})$ for E2. The label is a function of only $(X_1, X_2, ..., X_6)$ and is of the form $f(\mathbf{x}) = \mathsf{sign}[p(\mathbf{x})]$, where p is a polynomial generated randomly according to an Erlang distribution (See Appendix D for more details).

B. Numerical Results for Unsupervised Algorithms

We compare the performance of UFFS with Laplacian Score (LS) [17], MCFC [36], and UDFS [14] on the real and the synthetic data sets. The labels are not revealed to the algorithms but are used for measuring the performances. Features are randomly ordered so that the initial ordering would not affect the experiments' outcomes. Contrary to other algorithms, UFFS does not rank the feature; instead, it outputs a set of indices as the non-redundant features. We run UFFS three times: first with $t = 1, m = d$, second, with $t = 2, m = 50$ but on the selected features from the first run, and third, with $t = 3, m = 30$ but on the selected features from the second run. For each experiment, let $\tilde{d}$ denote the number of the selected features by UFFS at the third run. For comparing the performance to the ranking algorithms, we select only the $\tilde{d}$ features with the highest rank. Once the features are selected by each unsupervised algorithm, we reveal the samples of the selected features with the labels to a classifier and compute its prediction accuracy. A support vector machine (SVM) classifier with radial basis function as kernel is employed for all the studies. We perform a 5-fold cross-validation using this classifier and on the entire data set.

Table II shows the average of the resulted classification accuracies for each algorithm. The second row is the resulted accuracy without any feature selection. The third row is $\tilde{d}$ which is the number of non-redundant features declared by the UFFS. Observe that, in synthetic data sets, $\tilde{d}$ is very close to 10 which is the actual number of non-redundant features. The resulted accuracy by the UFFS is very close or greater than the accuracy without feature selection which verifies that the removed features were redundant. Further, it significantly outperforms other algorithms in the synthetic and many real data sets. This result shows that the UFFS performs well on data sets with nonlinear redundancies.

C. Numerical Results for Supervised Setting

In this part, we compare SFFS with ReliefF [21], mRMR [18], MI [37], RFS [19], and CCM [20]. We run SFFS with $t = 1$ and $t = 2$ for real data sets and with $t = 3$ for synthetic ones. As a performance measure, we perform a

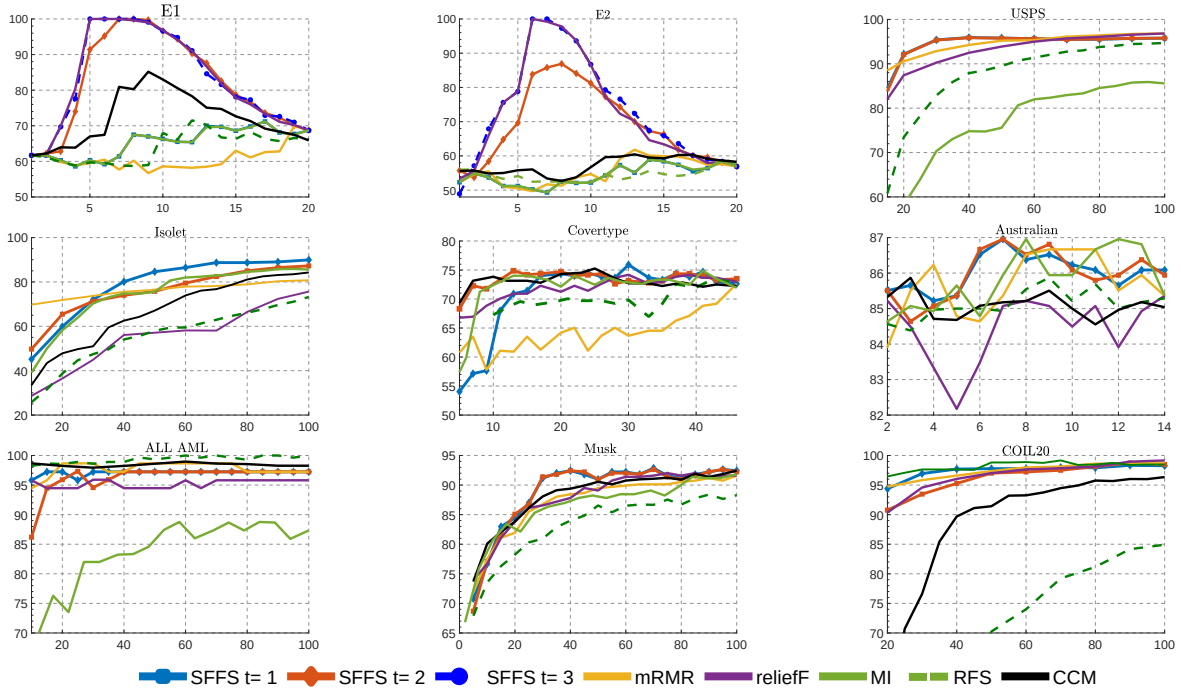

Fig. 1: Classification accuracy (y-axis) versus number of selected features k (x-axis). Our algorithm SFFS has consistent top performance, also leading in some ranges of k, in the above datasets. The kernel-based approach CCM has the computational complexity cubic in n, and thus we were unable to run CCM for USPS with its author's original implementation. Overlaps: SFFS $t = 3$ with ReliefF for E1 and E2, SFFS $t = 1$ with SFFS $t = 2$ for USPS and Musk.

5-fold cross-validation with feature selection and the SVM classifier described above in a pipeline. The parameters of the SVM classifier are the same as in our unsupervised experiments. The the implementation details are given in Appendix E. Figure 1 shows the average classification accuracy for various values of selected features (k). For real-world data sets, as the figure shows, we obtain consistently good results in all the data sets and leading in some ranges of k. The compared algorithms perform well only in some of the data sets while our algorithms have reliable steady performance in all the cases. For instance, in the Isolet data set, we observe a dominant performance by our SFFS for $k > 40$ as compared to other algorithms. In Musk, we observe a notable performance improvement for $k \in [25, 50]$. Note that SFFS with $t = 1$ and $t = 2$ are overlapping in these data sets and for many values of k.

As explained before, in E1 and E2, there are no redundant features and there are only 6 relevant features. This is verified in Figure 1, where the maximum accuracy (100%) is obtained at around $k = 6$. Further, we observe a significant performance gap between our approach and the other algorithms except for ReliefF. The low accuracy of these algorithms (below 60% in E2) suggests their failure to find the relevant features in these data sets. We believe this is due to the highly nonlinear feature-label relations in such data sets which are imposed by the Erlang distribution in our constructions. This observation calls for more sophisticated approaches in feature selection to address highly non-linear relations.

D. Classification accuracy vs t

Next, we analyze the effect of the depth parameter (t) on the algorithm's performance. Figure 2 presents classification accuracy of SFFS for various values of t when the number of selected features are $k = 5, 10$ and 20. We present the result for three of our datasets. As shown there, low values of t are sufficient to get a good performance. In some cases as t increases, the performance drops because the high value of t demands more number of samples.

E. Comparison of Running Times

Lastly, in Table III, we compare the running time of SFFS with other algorithms, and on the datasets we tested For the existing algorithms, the implementations are taken from [2] and correspond to the original implementations,

except for mRMR and CCM where we used the optimized implementations from the authors

TABLE III: Comparison of running times for 5-fold cross validation (in seconds). All the experiments were performed on 48-CPU workstation, with Intel(R) Xeon(R) CPU E7-8857 v2 @ 3.00GHz and 256GB RAM. The kernel-based algorithm CCM couldn't finish for long time in the USPS dataset due to cubic sample complexity

	Covertype	Australian	Musk	ALL_AML	USPS	Isolet	COIL20
SFFS (t=1)	2.7	3.5	3.3	303	298	74.26	41
SFFS (t=2)	3.1	3.9	4	378	378	74.35	65
RFS	6	4	2	447	1010	58	62
mRMR	1.41	0.89	56	300	510	3585	4238
relifF	1.33	1.88	1.3	4.35	550	36.5	41.42
MI	0.92	0.32	3.05	280	172	77	104
CCM	48	157	159	135	–	3276	3662

ACKNOWLEDGMENT

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

REFERENCES

[1] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[2] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Computing Surveys (CSUR)*, vol. 50, no. 6, p. 94, 2018.

[3] S. Solorio-Fernández, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, "A review of unsupervised feature selection methods," *Artificial Intelligence Review*, vol. 53, no. 2, pp. 907–948, 2020.

[4] N. Cesa-Bianchi, Y. Mansour, and O. Shamir, "On the complexity of learning with kernels," in *Conference on Learning Theory*, 2015, pp. 297–325.

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

[6] R. d. Wolf, *A Brief Introduction to Fourier Analysis on the Boolean Cube*, ser. Graduate Surveys. Theory of Computing Library, 2008, no. 1. [Online]. Available: http://www.theoryofcomputing.org/library.html

[7] W. Hoeffding, "A class of statistics with asymptotically normal distribution," *The Annals of Mathematical Statistics*, vol. 19, no. 3, pp. 293–325, 1948. [Online]. Available: http://www.jstor.org/stable/2235637

[8] I. M. Sobol, "Sensitivity estimates for nonlinear mathematical models," *Mathematical modelling and computational experiments*, vol. 1, no. 4, pp. 407–414, 1993.

[9] G. Chastaing, F. Gamboa, C. Prieur *et al.*, "Generalized hoeffding-sobol decomposition for dependent variables-application to sensitivity analysis," *Electronic Journal of Statistics*, vol. 6, pp. 2420–2448, 2012.

[10] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of machine learning research*, vol. 5, no. Oct, pp. 1205–1224, 2004.

[11] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[12] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial intelligence*, vol. 97, no. 1-2, pp. 245–271, 1997.

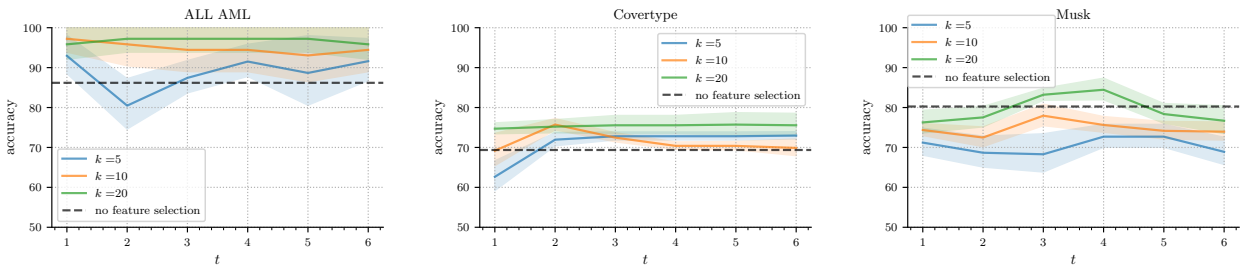

Fig. 2: Classification accuracy vs depth t for SFFS algorithm

[13] Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu, "Unsupervised feature selection using nonnegative spectral analysis," in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, ser. AAAI'12. AAAI Press, 2012, p. 1026–1032.

[14] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, "L2, 1-norm regularized discriminative feature selection for unsupervised," in *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[15] C. Feng, C. Qian, and K. Tang, "Unsupervised feature selection by pareto optimization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 3534–3541.

[16] H. Arai, C. Maung, K. Xu, and H. Schweitzer, "Unsupervised feature selection by heuristic search with provable bounds on suboptimality," in *Proceedings of the Thirtieth AAAI conference on Artificial Intelligence*, 2016, pp. 666–672.

[17] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Advances in neural information processing systems*, 2006, pp. 507–514.

[18] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[19] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint l2,1-norms minimization," in *Advances in Neural Information Processing Systems 23*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds. Curran Associates, Inc., 2010, pp. 1813–1821.

[20] J. Chen, M. Stern, M. J. Wainwright, and M. I. Jordan, "Kernel feature selection via conditional covariance minimization," in *Advances in Neural Information Processing Systems*, 2017, pp. 6946–6955.

[21] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *Proceedings of the 10th National Conference on Artificial Intelligence, San Jose, CA, USA, July 12-16, 1992*, W. R. Swartout, Ed. AAAI Press / The MIT Press, 1992, pp. 129–134. [Online]. Available: http://www.aaai.org/Library/AAAI/1992/aaai92-020.php

[22] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural computing and applications*, vol. 24, no. 1, pp. 175–186, 2014.

[23] D. Koller and M. Sahami, "Toward optimal feature selection," Stanford InfoLab, Tech. Rep., 1996.

[24] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537–550, Jul. 1994.

[25] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, "Measuring statistical dependence with hilbert-schmidt norms," in *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2005, pp. 63–77.

[26] X. Wei, B. Cao, and P. S. Yu, "Nonlinear joint unsupervised feature selection," in *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 2016, pp. 414–422.

[27] M. Feder and N. Merhav, "Relations between entropy and error probability," *IEEE Transactions on Information Theory*, vol. 40, no. 1, pp. 259–266, 1994.

[28] T. M. Cover and J. A. Thomas, *Elements of information theory*. Wiley-Interscience, 2006.

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

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

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

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

[33] M. Heidari, S. S. Pradhan, and R. Venkataramanan, "Boolean functions with biased inputs: Approximation and noise sensitivity," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, Jul. 2019, pp. 1192–1196.

[34] M. Heidari, G. I. Shamir, and W. Szpankowski, "Fourier-based universal learning," *JMLR (under review)*, 2021.

[35] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[36] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 333–342.

[37] A. Kraskov, H. Stögbauer, and P. Grassberger, "Erratum: Estimating mutual information [phys. rev. e 69, 066138 (2004)]," *Physical Review E*, vol. 83, no. 1, p. 019903, 2011.

APPENDIX A
PROOF OF PROPOSITION 1

Recall from the discussion in Section II that $\psi_{\mathcal{S}_i}$'s are orthogonormal. We complete the proof of the Proposition by showing that any function g can be written as a linear combination of these parities.

Let $D_{X_j}, j \in [d]$, be the marginals of $D_{\mathbf{X}}$ and let P_{X^d} be the product probability distribution with the same marginals D_{X_j}. Without loss of generality, assume that X_j's are non-trivial random variables. Then, from the Fourier analysis on the Boolean cube [5], the function g can be written as

$$g(\mathbf{x}) = \sum_{\mathcal{S}\subseteq[d]} g_{\mathcal{S}}\ \phi_{\mathcal{S}}(\mathbf{x}), \qquad \forall \mathbf{x} \in \{-1,1\}^d,$$

where $g_{\mathcal{S}} = \mathbb{E}_{P_{X^d}}[g(X^d)\phi_{\mathcal{S}}(X^d)]$ and the expectation is taken with respect to P_{X^d}. By performing the reverse of the orthogonalization process in (2), each parity $\phi_{\mathcal{S}_i}, i = 1, 2, .., 2^d$, can be written as

$$\phi_{\mathcal{S}_i}(\mathbf{x}) = \sum_{j\le i} \alpha_{i,j}\psi_{\mathcal{S}_j}(\mathbf{x}), \tag{12}$$

where $\alpha_{i,j} = \langle \phi_{\mathcal{S}_j}, \psi_{\mathcal{S}_j}\rangle$ and the above equality holds for all $\mathbf{x} \in \{-1,1\}^d$ except a measure-zero subset. Hence, replacing $\phi_{\mathcal{S}_i}$ with the right-hand side of (12), we can write

$$\begin{aligned} g(\mathbf{x}) &= \sum_{i=1}^{2^d} g_{\mathcal{S}_i}\ \Big(\sum_{j:j\le i} \alpha_{i,j}\psi_{\mathcal{S}_j}(\mathbf{x})\Big) \\ &= \sum_{j=1}^{2^d} \Big(\sum_{i:i\ge j} g_{\mathcal{S}_i}\alpha_{i,j}\Big)\psi_{\mathcal{S}_j}(\mathbf{x}). \end{aligned}$$

Hence, we obtain a decomposition of g as a linear combination of $\psi_{\mathcal{S}_i}$'s. Since, $\psi_{\mathcal{S}_i}$'s are orthogonal, the coefficients in this linear combination are unique and calculated as in the statement of the proposition.

APPENDIX B
PROOF OF THEOREM 3

From the proof of Theorem 2 and the definition of $f^{\subseteq\mathcal{J}}$, we obtain that

$$L_D(\mathcal{J}) = \frac{1}{2} - \frac{1}{2}\|f^{\subseteq\mathcal{J}}\|_1.$$

As a result,

$$L_D(\hat{\mathcal{J}}_n) - L_D(\mathcal{J}^*) = \frac{1}{2}\Big(\|f^{\subseteq\mathcal{J}^*}\|_1 - \|f^{\subseteq\hat{\mathcal{J}}_n}\|_1\Big). \tag{13}$$

By adding and subtracting $M_n(\hat{\mathcal{J}}_n)$ and $M_n(\mathcal{J}^*)$, we obtain that

$$\begin{aligned} \|f^{\subseteq\mathcal{J}^*}\|_1 - \|f^{\subseteq\hat{\mathcal{J}}_n}\|_1 &= \Big(\|f^{\subseteq\mathcal{J}^*}\|_1 - M_n(\mathcal{J}^*)\Big) + \Big(M_n(\mathcal{J}^*) - M_n(\hat{\mathcal{J}}_n)\Big) \\ &\quad + \Big(M_n(\hat{\mathcal{J}}_n) - \|f^{\subseteq\hat{\mathcal{J}}_n}\|_1\Big) \\ &\le \Big(\|f^{\subseteq\mathcal{J}^*}\|_1 - M_n(\mathcal{J}^*)\Big) + \Big(M_n(\hat{\mathcal{J}}_n) - \|f^{\subseteq\hat{\mathcal{J}}_n}\|_1\Big), \end{aligned} \tag{14}$$

where the last inequality follows as $M_n(\mathcal{J}^*) \le M_n(\hat{\mathcal{J}}_n)$. Next, we provide upper bounds on the right-hand side of the above inequality. Let $\hat{\mu}_j$ and $\hat{\sigma}_j, j = 1, 2, ..., d$, denote the empirical estimate of the mean and standard deviation of the features. For any subset $\mathcal{S}$ with at most k elements, let $\widehat{\phi}_{\mathcal{S}}(x^d) = \prod_{j\in\mathcal{S}} \frac{x_j-\hat{\mu}_j}{\hat{\sigma}_j}$. Now, fix a subset $\mathcal{J}$ with $|\mathcal{J}| \le k$ and perform the orthogonalization process w.r.t $\mathcal{J}$. We proceed with the following lemma which is proved in Appendix C.

Lemma 2. *The measure $M_n(\mathcal{J})$ as in (10) is an asymptotically unbiased estimate of $\|f^{\subseteq\mathcal{J}}\|_1$. More precisely, given any $\gamma \in (0, \frac{1}{2})$ and for any feature subset $\mathcal{J}$ with $|\mathcal{J}| \leq k$,*

$$\Big| \mathbb{E}_{\mathcal{S}_n \sim D^n} \big[M_n(\mathcal{J}) \big] - \|f^{\subseteq\mathcal{J}}\|_1 \Big| \leq O(n^{-\gamma}),$$

where the expectation is taken with respect to the training samples.

Next, we apply McDiarmid inequality on $M_n(\mathcal{J})$ and show that $M_n(\mathcal{J})$ is an accurate estimate of $\|f^{\subseteq\mathcal{J}}\|_1$ with high probability. Note that M_n is a function of the random training samples $(\mathbf{x}_i, y_i)$. Suppose, for a fixed i, the training instant $(\mathbf{x}_i, y_i)$ is replaced with an i.i.d. copy $(\tilde{\mathbf{x}}_i, \tilde{y}_i)$. Let $\tilde{M}_n^{(1)}$ be the resulted measure with $(\tilde{\mathbf{x}}_i, \tilde{y}_i)$ replacing $(\mathbf{x}_i, y_i)$. Then, we can show that for any $\mathcal{J}$ with $|\mathcal{J}| \leq k$, the inequality holds almost surely

$$|M_n(\mathcal{J}) - \tilde{M}_n^{(1)}(\mathcal{J})| \leq \frac{4}{n-1} 2^k \max_{\mathcal{S}\subseteq[d], |\mathcal{S}|\leq k} \|\psi_{\mathcal{S}}\|_\infty^2 \triangleq \frac{4\ 2^k c_k}{n-1}.$$

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

$$\mathbb{P}\Big\{ \big| M_n(\mathcal{J}) - \mathbb{E}[M_n(\mathcal{J})] \big| \leq \epsilon' \Big\} \leq 2 \exp\Big\{ - \frac{(n-1)\epsilon'^2}{8\ 2^{2k} c_k^2} \Big\},$$

where the expectation is taken with respect to the training samples. Using the union bound, we obtain that

$$\mathbb{P}\Big\{ \bigcup_{\mathcal{J}:|\mathcal{J}|=k} \Big\{ \big| M_n(\mathcal{J}) - \mathbb{E}[M_n(\mathcal{J})] \big| \leq \epsilon' \Big\} \Big\} \leq 2 \binom{d}{k} \exp\Big\{ - \frac{(n-1)\epsilon'^2}{8\ 2^{2k} c_k^2} \Big\}.$$

Thus, with probability $(1-\delta)$, the inequality

$$\big| M_n(\mathcal{J}) - \mathbb{E}[M_n(\mathcal{J})] \big| \leq \sqrt{\frac{\lambda(k)}{(n-1)} \log(\frac{d}{\delta})},$$

holds for all $\mathcal{J} \subseteq [d]$ with $|\mathcal{J}| = k$, where $\lambda(k) = 8\ k 2^{2k} c_k^2$. Next, from Lemma 2 and the triangle inequality, we have, with probability at least $(1-\delta)$, that

$$\big| M_n(\mathcal{J}) - \|f^{\subseteq\mathcal{J}}\|_1 \big| \leq \sqrt{\frac{\lambda(k)}{(n-1)} \log(\frac{d}{\delta})} + O(n^{-\gamma}), \quad \forall \mathcal{J} \subseteq [d],\ |\mathcal{J}| = k. \tag{15}$$

The proof completes by combining (13), (14), and (15).

Appendix C
Proof of Lemma 2

We first assume that there is no estimation error for mean and standard deviation of the features; that is $\hat{\mu}_j = \mu_j$ and $\hat{\sigma}_j = \sigma_j$ for all $j \in [d]$. Further, $\hat{b}_{ij} = b_{ij}$ for all i, j for which their corresponding feature subsets satisfy $|\mathcal{S}_i| \leq k$ and $|\mathcal{S}_j| \leq k$. Let $\overline{M}_n$ be the M_n under this condition. Define, the function

$$\bar{f}_{(i)}^{\subseteq\mathcal{J}}(\mathbf{x}) \triangleq \frac{n}{n-1} \sum_{\mathcal{S}\subseteq\mathcal{J}} \Big(\bar{f}_S - \frac{1}{n} Y(i) \psi_{\mathcal{S}}(\mathbf{X}(i)) \Big) \psi_{\mathcal{S}}(\mathbf{x}), \tag{16}$$

for all $\mathbf{x} \in \mathcal{X}^d$, where $\bar{f}_S = \frac{1}{n} \sum_i Y(i) \psi_{\mathcal{S}}(\mathbf{X}(i))$. With this definition, given any $\mathbf{x}$, the quantity $\bar{f}_{(i)}^{\subseteq\mathcal{J}}(\mathbf{x})$ is independent of $(\mathbf{X}(i), Y(i))$. Further, we can write $\overline{M}_n$ as the summation $\overline{M}_n(\mathcal{J}) = \frac{1}{n} \sum_i |\bar{f}_{(i)}^{\subseteq\mathcal{J}}(\mathbf{X}(i))|$. Hence, the expectation of $\overline{M}_n$ taken over the training samples gives

$$\begin{aligned}
\mathbb{E}[\overline{M}_n(\mathcal{J})] &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\mathbf{X}(1),\ldots,\mathbf{X}(n)} \Big[\Big| \bar{f}_{(i)}^{\subseteq\mathcal{J}}(\mathbf{X}(i)) \Big| \Big] \\
&= \mathbb{E}_{\mathbf{X}(1),\ldots,\mathbf{X}(n)} \Big[\Big| \bar{f}_{(1)}^{\subseteq\mathcal{J}}(\mathbf{X}(1)) \Big| \Big] \\
&= \mathbb{E}_{\mathbf{X}(2),\ldots,\mathbf{X}(n)} \mathbb{E}_{\mathbf{X}(1)} \Big[\Big| \bar{f}_{(1)}^{\subseteq\mathcal{J}}(\mathbf{X}(1)) \Big| \Big] \\
&= \mathbb{E}_{\mathbf{X}(2),\ldots,\mathbf{X}(n)} \big[\|\bar{f}_{(1)}^{\subseteq\mathcal{J}}\|_1 \big],
\end{aligned} \tag{17}$$

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 $\bar{f}_{(1)}^{\subseteq\mathcal{J}}$ is independent of $(\mathbf{X}(1), Y(1))$. Note that $\bar{f}_{(1)}^{\subseteq\mathcal{J}}$ is as an estimation of the projection $f^{\subseteq\mathcal{J}}$ using the $(n-1)$ training samples $(\mathbf{X}(i), Y(i)), i = 2, 3, ..., n$. Next, we bound the difference $\Big|\mathbb{E}\|\bar{f}_{(1)}^{\subseteq\mathcal{J}}\|_1 - \|f^{\subseteq\mathcal{J}}\|_1\Big|$. Observe that

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

where the first inequality is obtained by applying the triangle inequality twice, one for $\|\bar{f}_{(1)}^{\subseteq\mathcal{J}}\|_1$ and once for $\|f^{\subseteq\mathcal{J}}\|_1$. The second inequality is from the identity $\|\cdot\|_1 \leq \|\cdot\|_2$. The third inequality is due to the Jensen's inequality. Next, by Parseval's identity we have

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

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

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

Hence,

$$\begin{aligned}\mathbb{E}\big[\|f^{\subseteq\mathcal{J}} - \bar{f}_{(1)}^{\subseteq\mathcal{J}}\|_2^2\big] = \frac{1}{n-1}\sum_{S\subseteq J}(1 - {f_{\mathcal{S}}}^2) &= \frac{1}{n-1}(2^{|\mathcal{J}|} - \|f^{\subseteq\mathcal{J}}\|_2^2) \\ &\leq \frac{1}{n-1}2^k.\end{aligned}$$

Putting all together we get that

$$\Big|\mathbb{E}\big[\overline{M}_n\big] - \|f^{\subseteq\mathcal{J}}\|_1\Big| = \Big|\mathbb{E}\big[\|\bar{f}_{(1)}^{\subseteq\mathcal{J}}\|_1\big] - \|f^{\subseteq\mathcal{J}}\|_1\Big| \leq \frac{2^{k/2}}{\sqrt{n-1}}. \tag{18}$$

Next, we address the effect of 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 following event

$$\text{(B)}: \quad \big|\hat{\mu}_j - \mu_j\big| \leq \epsilon_0, \qquad \text{and} \qquad \big|1 - \frac{\sigma_j}{\hat{\sigma}_j}\big| \leq \frac{2\epsilon_0}{\sigma_j^2}, \quad \forall j \in [d], \tag{19}$$

to happen 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 3. *Given $\epsilon_0, \delta_0 \in (0, 1)$, the event* (B) *happens with probability at least $(1 - \delta_0)$, provided that atleast $n_0(\epsilon_0, \delta_0) = \frac{2}{\epsilon_0^2}\log\frac{2d}{\delta_0}$ samples are available.*

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

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

Therefore, applying the union bound gives

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

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

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

As a result,

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

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

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

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

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

As a result,

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

which completes the proof of the lemma. ■

By conditioning on (B) and from triangle inequality we obtain that

$$\begin{aligned}\Big|\mathbb{E}\big[M_n(\mathcal{J})|B\big]-\|f^{\subseteq\mathcal{J}}\|_1\Big|&\leq\Big|\mathbb{E}\big[M_n(\mathcal{J})|B\big]-\mathbb{E}\big[\overline{M}_n(\mathcal{J})\big]\Big|+\Big|\mathbb{E}\big[\overline{M}_n(\mathcal{J})\big]-\|f^{\subseteq\mathcal{J}}\|_1\Big|\\&\overset{(a)}{\leq}\Big|\mathbb{E}\big[M_n(\mathcal{J})|B\big]-\mathbb{E}\big[\overline{M}_n(\mathcal{J})\big]\Big|+\frac{2^{k/2}}{\sqrt{n-1}}\end{aligned}\tag{20}$$

where (a) follows from (18). We proceed with the following lemma which is proved in Appendix C-A.

Lemma 4. *Conditioned on B the inequalities $\big|\,\overline{M}_n(\mathcal{J})-M_n(\mathcal{J})\,\big|\leq\lambda(\epsilon_0)$ hold, almost surely, for all k-element subsets $\mathcal{J}$, where λ is a function satisfying $\lambda(\epsilon_0)=O(k2^kc_k\epsilon_0)$ as $\epsilon\to 0$.*

Consequently, from (20) we have

$$\Big|\mathbb{E}\big[M_n(\mathcal{J})|B\big]-\|f^{\subseteq\mathcal{J}}\|_1\Big|\leq\lambda(\epsilon_0)+\frac{2^{k/2}}{\sqrt{n-1}}$$

A. *Proof of Lemma 4*

Recall, from (16) and the argument afterward, that $\overline{M}_n$ can be written as $\overline{M}_n(\mathcal{J})=\frac{1}{n}\sum_i\Big|\bar{f}_{(i)}^{\subseteq\mathcal{J}}(\mathbf{X}(i))\Big|$, where $\bar{f}_{(i)}^{\subseteq\mathcal{J}}$ is defined in (16) and is repeated here

$$\bar{f}_{(i)}^{\subseteq\mathcal{J}}(\mathbf{x})\triangleq\frac{n}{n-1}\sum_{\mathcal{S}\subseteq\mathcal{J}}\Big(\bar{f}_S-\frac{1}{n}Y(i)\psi_{\mathcal{S}}(\mathbf{X}(i))\Big)\psi_{\mathcal{S}}(\mathbf{x}),$$

with $\bar{f}_S=\frac{1}{n}\sum_iY(i)\psi_{\mathcal{S}}(\mathbf{X}(i))$. Similarly, M_n can be written as

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

where $\hat{f}_{(i)}^{\subseteq\mathcal{J}}$ is defined as

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

Note that $\hat{f}_{(i)}^{\subseteq\mathcal{J}}$ is similar to $\bar{f}_{(i)}^{\subseteq\mathcal{J}}$ as in (16), but with μ_j, σ_j and $\psi_{\mathcal{S}}$ replaced with their estimations $\hat{\mu}_j, \hat{\sigma}_j$ and $\widehat{\psi}_{\mathcal{S}}$, respectively.

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

$$\big|\overline{M}_n(\mathcal{J}) - M_n(\mathcal{J})\big| \le \frac{1}{n}\sum_i \big|\bar{f}_{(i)}^{\subseteq\mathcal{J}}(\mathbf{X}(i)) - \hat{f}_{(i)}^{\subseteq\mathcal{J}}(\mathbf{X}(i))\big| \le \|\bar{f}_{(1)}^{\subseteq\mathcal{J}} - \hat{f}_{(1)}^{\subseteq\mathcal{J}}\|_\infty,$$

where the last inequality follows by maximizing over all realizations of $\mathbf{X}(i)$ and the symmetricity with respect to i. Define the function $\bar{f}^{\subseteq\mathcal{J}}$ as

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

Note that, $\bar{f}_{(1)}^{\subseteq\mathcal{J}}$ and $\hat{f}_{(1)}^{\subseteq\mathcal{J}}$ are, respectively, equal to $\bar{f}^{\subseteq\mathcal{J}}$ and $\hat{f}^{\subseteq\mathcal{J}}$ when the first sample $(\mathbf{X}(1), Y(1))$ is removed from the training samples. Hence, to bound $\|\bar{f}_{(1)}^{\subseteq\mathcal{J}} - \hat{f}_{(1)}^{\subseteq\mathcal{J}}\|_\infty$, it suffices bound$\|\hat{f}^{\subseteq\mathcal{J}} - \bar{f}^{\subseteq\mathcal{J}}\|_\infty$. That said, it remains to showing that

$$\|\hat{f}^{\subseteq\mathcal{J}} - \bar{f}^{\subseteq\mathcal{J}}\|_\infty \le \lambda(\epsilon_0), \tag{21}$$

where $\lambda(\epsilon_0) = O(k2^k c_k\epsilon_0)$ as $\epsilon_0 \to 0$. The argument is as follows.

From triangle inequality for ∞-norm and the definition of $\hat{f}^{\subseteq\mathcal{J}}$ and $\bar{f}^{\subseteq\mathcal{J}}$ we obtain

$$\|\hat{f}^{\subseteq\mathcal{J}} - \bar{f}^{\subseteq\mathcal{J}}\|_\infty \le \sum_{S\subseteq J}\|\hat{f}_{\mathcal{S}}\ \widehat{\psi}_{\mathcal{S}} - \bar{f}_S\ \psi_{\mathcal{S}}\|_\infty. \tag{22}$$

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

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

Next, note that from triangle inequality

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

Therefore,

$$\|\hat{f}_{\mathcal{S}}\ \widehat{\psi}_{\mathcal{S}} - \bar{f}_S\ \psi_{\mathcal{S}}\|_\infty \le \big(\|\widehat{\psi}_{\mathcal{S}}\|_\infty + |\bar{f}_S|\big)\|\widehat{\psi}_{\mathcal{S}} - \psi_{\mathcal{S}}\|_\infty. \tag{23}$$

We proceed by bounding each term on the right-hand side of (23). As for the first term, we have that $\|\widehat{\psi}_{\mathcal{S}}\|_\infty \le \|\psi_{\mathcal{S}}\|_\infty + \|\widehat{\psi}_{\mathcal{S}} - \psi_{\mathcal{S}}\|_\infty$.

As for the second term,

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

Lastly, the third term is bounded using the following lemma which is proved in Appendix C-B.

Lemma 5. *Conditioned on* (B)*, the inequality* $\|\psi_{\mathcal{S}} - \widehat{\psi}_{\mathcal{S}}\|_\infty \le \gamma(\epsilon_0)$ *holds for all* k*-element subsets* $\mathcal{S}$*, almost surely, where* γ *is a function satisfying* $\gamma(\epsilon_0) = O(k\epsilon_0\sqrt{c_k})$ *as* $\epsilon_0 \to 0$*.*

Recall that c_k is defined as $c_k = \max_{\mathcal{S}:|\mathcal{S}|\le k}\|\psi_{\mathcal{S}}\|_\infty^2$. Therefore, combining these bounds for the terms in (23) gives the following bound

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

Plugging the above inequality into (22) gives the desired bound

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

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

B. *Proof of Lemma 5*

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

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

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

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

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

$$\|\psi_{\mathcal{S}} - \widehat{\psi}_{\mathcal{S}}\|_\infty \leq |1-b_{\mathcal{S}}|\|\psi_{\mathcal{S}}\|_\infty + \frac{k\epsilon}{\sigma_{\min}}b_{\mathcal{S}}(1+\epsilon)^k\|\psi_{\mathcal{S}}\|_\infty. \tag{24}$$

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

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

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

$$1-e^{k\epsilon} \leq 1-b_{\mathcal{S}} \leq k\epsilon \leq e^{k\epsilon}-1. \tag{25}$$

Lastly, combining (24) and (25) gives the following inequality

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

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

APPENDIX D
GENERATING RANDOM LABELING FUNCTIONS VIA ERLANG DISTRIBUTION

We generate randomly a labeling function which is the sign of a polynomial of the form

$$p(\mathbf{x}) \triangleq \sum_{\mathcal{S}} \alpha_{\mathcal{S}} \mathbf{x}^{\mathcal{S}},$$

where $\mathbf{x}^{\mathcal{S}} = \prod_{j \in \mathcal{S}} x_j$ and the coefficients $\alpha_{\mathcal{S}} \in [0, 1]$ are generated randomly according to the following process:

Let $f_E(x)$ where f_E is the pdf of the Erlang random variable with *shape* and *rate* parameters equal to 8 and 1, respectively. Let $w_i = f_E(i), i = 1, 2, ..., m$. For each w_i, we select 10 subsets randomly from the collection of all subsets $\mathcal{S} \subseteq [d]$ that have i-elements. The selected subsets for each i are denoted as $\mathcal{S}_{i,j}, j = 1, 2, ..., 10$. Let $V_{i,j} \sim \mathsf{Unif}([0, 1])$, $i \in [m]$ and $j \in [10]$ be i.i.d. random variables. Then, the Fourier coefficient corresponding to $\mathcal{S}_{i,j}$ is determined as $\alpha_{i,j} = W_i \times V_{i,j}$. With that the polynomial p can be written as $p(\mathbf{x}) = \sum_{i,j} \alpha_{i,j} \prod_{\ell \in \mathcal{S}_{i,j}} x_\ell$. Note that by changing the parameters of the Erlang pdf, we get different randomized polynomials.

APPENDIX E
IMPLEMENTATION DETAILS

In this section, we explain the details of our implementations of UFFS and SFFS algorithms.

The following are some of the characteristics of our implementation:

- For benchmarking purposes, we use the original implementation of mRMR[2], scikit-feature[3] for UDFS, MCFS, and ReliefF, and scikit-learn[4] for mutual information (MI)-based algorithm.
- Though most parts are written in Python, the code snippets that require heavy computations (B and A matrix computations in Algorithm 1 and Fourier coefficient calculation in Algorithm 2) are converted to C++ using Cython.
- We have also parallelized some of the computations.
- The code also contains a greedy implementation of the SFFS algorithm which we have not discussed in the paper, and can be employed for extremely large datasets.

UFFS with limited computational resources: To minimize the computational burden further, we follow a sequential approach for the UFFS algorithm. Let the target depth t be 3, and a_1, a_2, m_1, and m_2 be some positive integers. First we find the set of non-redundant features outputted by the UFFS with $t = 1$. Let its count be d_1. If the actual number of features $d < a_1$, we directly run the UFFS with $t = 2$ on the full set of features. Otherwise, if $d_1 < a_1$, UFFS ($t = 2$) is run on the selected features from $t = 1$ step. In case $d_1 \geq a_1$, we split the d_1 features from step $t = 1$ to multiple non-overlapping clusters of size m_1, and the UFFS ($t = 2$) is executed on these clusters and combine the selected features. Let the number of selected features from step $t = 2$ be d_2. For step $t = 3$, we pursue a similar approach as in the previous step with the selected features from the UFFS ($t = 2$): a) if $d < a_2$, run the UFFS directly; b) else if $d_2 < a_2$, run the UFFS ($t = 3$) on d_2 features; c) in case $d_2 \geq a_2$, divide d_2 features into non-overlapping clusters of size m_2 and run $t = 3$ step on each of them. Here a_1, a_2, and m_1, m_2 are hyperparameters that needs to be chosen depending on the computational resources.

Our supervised algorithm (SFFS) does not require the above approximation since it mainly depends on the computation of the Fourier coefficients and it usually has fast execution with our C++ implementation.

[2] http://home.penglab.com/proj/mRMR/
[3] http://featureselection.asu.edu/
[4] https://scikit-learn.org