# Finding Relevant Information via a Discrete Fourier Expansion

**Mohsen Heidari** [1] **Jithin K. Sreedharan** [2] **Gil I. Shamir** [3] **Wojciech Szpankowski** [1]

## Abstract

A fundamental obstacle in learning information from data is the presence of nonlinear redundancies and dependencies in it. To address this, we propose a Fourier-based approach to extract relevant information in the supervised setting. We first develop a novel Fourier expansion for functions of correlated binary random variables. This expansion is a generalization of the standard Fourier analysis on the Boolean cube beyond product probability spaces. We further extend our Fourier analysis to stochastic mappings. As an important application of this analysis, we investigate learning with feature subset selection. We reformulate this problem in the Fourier domain and introduce a computationally efficient measure for selecting features. Bridging the Bayesian error rate with the Fourier coefficients, we demonstrate that the Fourier expansion provides a powerful tool to characterize nonlinear dependencies in the features-label relation. Via theoretical analysis, we show that our proposed measure finds provably *asymptotically optimal* feature subsets. Lastly, we present an algorithm based on our measure and verify our findings via numerical experiments on various datasets.

## 1. 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 (Guyon & Elisseeff, 2003; Li et al., 2018; Solorio-Fernández et al., 2020). Kernel-based methods are on the other hand able to capture non-linear relations (Gretton et al., 2005; Chen et al., 2017; Wei et al., 2016). However, they are prohibitive in large datasets as the computational complexity of computing a kernel grows super linearly with the number of the samples (Cesa-Bianchi et al., 2015). Alternatively, information-theoretic metrics are powerful candidates in quantifying nonlinear dependencies among the random variables (Vergara & Estévez, 2014; Koller & Sahami, 1996; Yu & Liu, 2004; Battiti, 1994; Peng et al., 2005). However, estimating such quantities usually requires high sample complexity. Other approaches are *wrapper* and *embedded* methods where the feature subsets are evaluated directly by an induction algorithm (Yamada et al., 2020b). Such approaches are usually computationally expensive and, hence, prohibitive in large datasets.

In this work, we take an alternative approach and adapt discrete Fourier analysis to capture 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 *parities* (O'Donnell, 2014; Wolf, 2008). Highly "nonlinear" functions tend to have Fourier expansion with large coefficients for high-degree parities — potentially making the Fourier expansion a powerful tool in learning problems. However, this expansion has a few limitations that need to be addressed. 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 intend to address these challenges by proposing a Fourier-based feature selection algorithm.

### 1.1. Main Contributions

We demonstrate, via theoretical analysis and numerical experiments, that the Fourier expansion provides a powerful tool to characterize nonlinear *redundancies* and *dependencies* in the data. To the best of our knowledge, this is the first instance of using the Fourier expansion as a measure in supervised feature selection. In what follows, we summarize the main contributions of this paper.

[1] NSF Center for Science of Information, Purdue University, West Lafayette, USA [2] Wadhwani AI, Mumbai, India [3] Google Inc., Pittsburgh, USA. Correspondence to: Mohsen Heidari <mheidari@purdue.edu>, Jithin K. Sreedharan <jithin@wadhwaniai.org>.

*Proceedings of the $38^{th}$ International Conference on Machine Learning*, PMLR 139, 2021. Copyright 2021 by the author(s).

**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 design 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). Although this Fourier expansion is defined on the Boolean cube, our analysis applies to non-binary features as well. We view the binary Fourier as a framework that captures a particular class of nonlinearities — those characterized via the *parities*. Our numerical results in Section 6 verifies that this Fourier expansion is sufficient to capture the nonlinearities. 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.

**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 (Guyon & Elisseeff, 2003).

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 an exhaustive search based on this measure finds an asymptotically optimal feature subset when the features are binary. 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 2).

**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 $d$. Our numerical results in Section 6 show that typically $\tilde{d}$ is much smaller than $d$ (See Table 2). Hence the overall computational complexity of SFFS is dominated by $O(nd)$ which is linear in the size of the data.

Through our numerical experiments, we show that SFFS, even with $t = 1$ or $t = 2$, performs consistently better on a variety of datasets as compared to several well-known feature selection algorithms such as mRMR (Peng et al., 2005), Mutual Information (Kraskov et al., 2011), RFS (Nie et al., 2010), CCM (Chen et al., 2017), and ReliefF (Kira & Rendell, 1992) (See Section 6). We thus overcome two well-known demerits of wrapper methods for feature selection that limit their usage in practice – heavy dependency on the predictive performance of the learning algorithm and huge search space.

**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 $\langle g_1, g_2 \rangle_D = \mathbb{E}_D[g_1(\mathbf{X})g_2(\mathbf{X})]$, where $D$ is the distribution of the input variables.

## 2. Optimal Feature Selection: A Fourier Perspective

We consider the learning problem with $d$ real-valued features and with labels taking values from $\{-1, 1\}$. We restrict ourselves to binary classification with $0 - 1$ loss function for convenience in presenting the theoretical results. In this case, the expected loss is the *misclassification* probability.

The features $\mathbf{X} \in \mathbb{R}^d$ and the label $Y \in \{-1, 1\}$ are generated according to an unknown 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 a fixed, but unknown, distribution $D$.

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

$$L_D(\mathcal{J}) = \min_{g \in \mathcal{G}_k} \mathbb{P}_{(\mathbf{X},Y)\sim D}\big\{Y \neq g(X^{\mathcal{J}})\big\}, \qquad (1)$$

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

$$\mathcal{J}^* = \operatorname*{arg\,min}_{\mathcal{J} \subseteq [d],\, |\mathcal{J}|=k} L_D(\mathcal{J}), \qquad L^*_D(k) = L_D(\mathcal{J}^*). \quad (2)$$

In agnostic settings, where only a training dataset is available, the above optimization is not feasible 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 $\mathsf{T}_k$ 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 $L_D(\mathcal{J}^*)$ with the optimal feature subset $\mathcal{J}^*$. For that, we first represent the problem in the Fourier domain.

### 2.1. Fourier Expansion for Functions of Correlated Boolean Random Variables

The main ingredient for our learning approach is a Fourier expansion that incorporates correlated binary random variables. We first present an overview of the standard Fourier expansion on the Boolean cube (O'Donnell, 2014; Wolf, 2008). It states that any bounded function $g : \{-1, 1\}^d \rightarrow \mathbb{R}$ can be written as a linear combination of *monomials*, as in the following

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

where $\mathbf{x}^{\mathcal{S}} = \prod_{j\in\mathcal{S}} x_j$, and $\mathrm{g}_{\mathcal{S}} \in \mathbb{R}$ are called the *Fourier coefficients*. Further, such coefficients are calculated as

$$\mathrm{g}_{\mathcal{S}} = \frac{1}{2^d} \sum_{\mathbf{x}\in\{-1,1\}^d} g(\mathbf{x})\, \mathbf{x}^{\mathcal{S}}.$$

This expansion is suitable when the probability distribution of the features is uniform over $\{-1, 1\}^d$. Hence, it finds its applications in many computational learning problems such as (Linial et al., 1993; Mossel et al., 2003; Heidari et al., 2019). However, we need a more sophisticated Fourier expansion incorporating non-uniform distributions for other learning problems such as feature selection. For that, we construct a set of orthogonal parity functions. Based on that, we establish our Fourier expansion for functions of correlated binary random variables.

**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 $g : \{-1, 1\}^d \rightarrow \mathbb{R}$ 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 $g_{\mathcal{S}} = \mathbb{E}_{D_X}[g(\mathbf{X})\psi_{\mathcal{S}}(\mathbf{X})]$.*

**Proof idea:** We start by centralizing and normalizing the input random variables. Let $\mu_j$ and $\sigma_j$ 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 $j \in [d]$. For any subset $\mathcal{S} \subseteq [d]$ defined

$$\chi_{\mathcal{S}}(\mathbf{x}) \triangleq \prod_{j\in\mathcal{S}} \frac{x_j - \mu_j}{\sigma_j}, \qquad \text{for all } \mathbf{x} \in \{-1,1\}^d.$$

Note that $\chi_{\mathcal{S}}$'s are not orthogonal because $X_j$'s are correlated. That said, we construct our Fourier expansion by designing a Gram-Schmidt-type procedure to make these parities orthogonal. Then, we use this basis to develop our Fourier expansion for functions of correlated random variables. The orthogonalization process is explained in the following.

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

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

We apply the orthogonalization process on $\chi_{\mathcal{S}_i}$ with the above ordering. The first orthogonalized parity is given by $\psi_\emptyset(\mathbf{x}) = 1$ for all $\mathbf{x} \in \{-1, 1\}^d$. Then, the orthogonalized parity corresponding to the $i$th subset is obtained from the following operation:

$$\begin{aligned} \tilde{\psi}_{\mathcal{S}_i} &\equiv \chi_{\mathcal{S}_i} - \sum_{j=1}^{i-1} \langle \psi_{\mathcal{S}_j}, \chi_{\mathcal{S}_i}\rangle_D\, \psi_{\mathcal{S}_j}, \\ \psi_{\mathcal{S}_i} &\equiv \begin{cases} \frac{\tilde{\psi}_{\mathcal{S}_i}}{\|\tilde{\psi}_{\mathcal{S}_i}\|_{2,D}} & \text{if } \|\tilde{\psi}_{\mathcal{S}_i}\|_{2,D} > 0 \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{4}$$

where $\|\tilde{\psi}_{\mathcal{S}_i}\|_{2,D} = \sqrt{\langle \tilde{\psi}_{\mathcal{S}_i}, \tilde{\psi}_{\mathcal{S}_i}\rangle_D}$. By construction, the resulted nontrivial parities $\psi_{\mathcal{S}_i}$'s are orthonormal, that is $\langle \psi_{\mathcal{S}_i}, \psi_{\mathcal{S}_j}\rangle_D = 0$ for $i \neq j$ and $\langle \psi_{\mathcal{S}_i}, \psi_{\mathcal{S}_i}\rangle_D = 1$ if $\psi_{\mathcal{S}_i}$ is not trivial.

The rest of the argument, given in the Supplementary Material, follows by showing that $\psi_{\mathcal{S}}$'s span the space of all bounded functions. □

Different orderings of subsets of $[d]$ result in different orthogonalized parities. The standard ordering in (3) is beneficial to our problem. Depending on the statistics of the features, the number of non-trivial parities ranges from 1 to $2^d$. On one extreme, if the features are mutually independent, then $\psi_{\mathcal{S}_i} = \chi_{\mathcal{S}_i}$. 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.

Among the trivial parities, there might be sets of single-element ones $\psi_{\{j\}}$. The features $j \in [d]$ for which $\psi_{\{j\}}$ is trivial do not appear in the Fourier expansion. Hence, they can be removed as they are statistically redundant. Based on this argument, for the feature selection problem in (2), the optimal feature subset $\mathcal{J}^*$ does not contain any of the redundant features. Hence, we search over feature subsets corresponding to non-trivial parities only. This is done automatically with the orthogonalization process. We use this argument when proposing Procedure 1 (FOURIER-ORTH).

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 function 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 subsets, 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 (4), there are only four non-trivial orthogonalized parities as*

$$\psi_\emptyset = 1,\ \ \psi_{\{1\}} = x_1,\ \ \psi_{\{2\}} = x_2,\ \ \psi_{\{1,2\}} = x_1x_2.$$

*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 particular class of nonlinearities characterized via orthogonalized parities. Our numerical experiments confirm that such an approximation is sufficient in many real-world datasets.

**Remark 1.** *We note that the orthogonalization process detailed in this section is different from conventional polynomial kernel methods for dimension reduction. Our orthogonalization method incorporates the underlying distribution of the features and outputs the basis for Fourier expansion, thus characterizing the optimal Bayes error rate (given later in Theorem 1). Further, this orthogonalization is suitable for feature selection as it does not mix the features.*

### 2.2. Fourier Model

Next, we use the Fourier expansion in the previous section and develop a representation of $L^*_D(k)$ in the Fourier domain. The Fourier expansion in the previous section is defined for deterministic functions. The labeling in the feature selection problem is not necessarily a function of the features. Rather, it is a stochastic mapping $Y$. That said, we extend the Fourier expansion to such mappings.

We proceed by characterizing the Bayes predictor via the Fourier expansion. Finally, we present the following results with the proofs provided in the Supplementary Material.

**Lemma 1.** *Let $\mathcal{J}$ be the subset of the selected features. Then, the Bayes predictor of the label 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 real-valued function on $\{-1,1\}^{|\mathcal{J}|}$ admitting the Fourier expansion*

$$f^{\subseteq\mathcal{J}}(\mathbf{x}^\mathcal{J}) \triangleq \sum_{\mathcal{S}\subseteq\mathcal{J}} f_\mathcal{S}\psi_\mathcal{S}(\mathbf{x}^\mathcal{J}), \tag{5}$$

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

Based on this result, we characterize $L^*_D(k)$ in the Fourier domain and find the optimal Feature subset.

**Theorem 1.** *The minimum attainable misclassification probability equals to*

$$L^*_D(k) = \frac{1}{2}\left[1 - \max_{\mathcal{J}\subseteq[d],\,|\mathcal{J}|=k} \|f^{\subseteq\mathcal{J}}\|_{1,D}\right], \tag{6}$$

*where $D$ is the feature-label distribution over $\{-1,1\}^{d+1}$. 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.*

The proof of Lemma 1 and Theorem 1 are provided in Section B and C of Supplementary Material.

## 3. A Measure for Feature Selection

The previous section provides the characterization in the ideal setting where $D$, the statistics of the features and labels, is known. Next, we leverage this characterization to the agnostic setting and present a measure for feature selection.

We only have access to $n$ i.i.d. training samples drawn from an unknown but fixed $D$. Based on Theorem 1, 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 (2). In what follows, we describe the derivation of $M_n$ in three steps:

**Step 1:** First, 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 (4). We elaborate on this step in Section 4.1. 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 (5). For that we calculate the following

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

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:** When $\hat{f}^{\subseteq \mathcal{J}}$ is obtained, 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

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

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

$$M_n(\mathcal{J}) = \widehat{\|f^{\subseteq \mathcal{J}}\|}_1 \triangleq \\ \frac{1}{n-1}\sum_{i=1}^{n}\Bigg|\sum_{\mathcal{S}\subseteq\mathcal{J}} \hat{f}_{\mathcal{S}}\widehat{\psi}_{\mathcal{S}}(\mathbf{x}(i)) - \frac{1}{n}y(i)\big(\widehat{\psi}_{\mathcal{S}}(\mathbf{x}(i))\big)^2\Bigg|. \tag{7}$$

This correction is done by subtracting the quantity

$$\frac{1}{n}y(i)\big(\widehat{\psi}_{\mathcal{S}}(\mathbf{x}(i))\big)^2.$$

We use $M_n(\mathcal{J})$ as an estimate of $\|f^{\subseteq \mathcal{J}}\|_1$. It can be shown that this estimator is asymptotically unbiased (See Lemma D.1 in Supplementary Material), that is

$$\lim_{n\to\infty}\Big|\mathbb{E}_{\mathcal{S}_n\sim D^n}[M_n(\mathcal{J})] - \|f^{\subseteq \mathcal{J}}\|_1\Big| = 0.$$

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.

**Theorem 2.** *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 (6). Then, given any $\delta \in (0,1)$, with probability at least $(1-\delta)$, the following bound holds*

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

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

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.

## 4. Proposed Algorithm

We build upon our Fourier expansion and propose our Supervised Fourier Feature Selection (SFFS) algorithm. To reduce the computational complexity, we propose a few approximations. We start with approximating the orthogonalization process.

### 4.1. Implementing the Orthogonalization

We propose a recursive formula to perform the orthogonalization. Let $b_{j,i} = \langle \chi_{\mathcal{S}_j}, \chi_{\mathcal{S}_i}\rangle$, and define $a_{j,i} = \langle \psi_{\mathcal{S}_j}, \chi_{\mathcal{S}_i}\rangle$. With this notation, $\tilde{\psi}_{\mathcal{S}_i}$ in (4) can be written as

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

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

$$a_{j,i} = \frac{1}{\sqrt{b_{j,j} - \sum_{r<j} a_{r,j}^2}}\Big(b_{j,i} - \sum_{\ell<j} a_{\ell,j}a_{\ell,i}\Big). \tag{8}$$

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

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

Then, we compute an estimation of $a_{j,i}$'s (denoted by $\hat{a}_{j,i}$) by calculating (8) with $b_{j,i}$ and $a_{j,i}$ replaced by $\hat{b}_{j,i}$ and $\hat{a}_{j,i}$, receptively. This approach is presented in Procedure 1 with additional approximation techniques explained below:

First, we approximate (4) 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 $\|\tilde{\psi}_{\{j\}}\|_2 \leq \epsilon$. In our earlier work, we show in (Heidari et al., 2021) that the FOURIER-ORTH can be used as a standalone unsupervised feature selection algorithm.

Further, we apply a fixed-depth search and 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 (3), but restricted to subsets of size at most $t$. For most practical purposes, we set $t \leq 3$. 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.

**Procedure 1** FOURIER-ORTH

```
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 $\mathsf{norm}(j), j = 1, 2, ...d$
Generate all subsets $\mathcal{S}_i \subseteq [d]$ with size at most $t$ and
  with the standard ordering as in (3).
Compute the matrix $\hat{\mathbf{B}}$ with elements:
  $$\hat{b}_{j,i} \leftarrow \frac{1}{n}\sum_{l=1}^{n}\Big[\prod_{u\in\mathcal{S}_j} x_{lu} \prod_{v\in\mathcal{S}_i} x_{lv}\Big].$$
Set $\hat{\mathbf{A}} \leftarrow \hat{\mathbf{B}}$
for row $j$ of $\hat{\mathbf{A}}$ do
  update the $j$th row: $\hat{\mathbf{A}}_{j,*} \leftarrow \hat{\mathbf{A}}_{j,*} - \sum_{\ell<j} \hat{a}_{\ell,j}\hat{\mathbf{A}}_{\ell,*}$
  Compute $\mathsf{norm}(\mathcal{S}_j) \leftarrow \sqrt{[\hat{b}_{j,j} - \sum_{r<j} \hat{a}_{r,j}^2]^+}$
  if $\mathsf{norm}(\mathcal{S}_j) \leq \epsilon$ then
    Set the $j$th row of $\hat{\mathbf{A}}$ zero: $\hat{\mathbf{A}}_{j,*} \leftarrow \mathbf{0}$
  else
    Normalize the $j$th row: $\hat{\mathbf{A}}_{j,*} \leftarrow \frac{\hat{\mathbf{A}}_{j,*}}{\mathsf{norm}(\mathcal{S}_j)}$
  end if
end for
Declare all $j \in [d]$ with $\mathsf{norm}(j) \geq \epsilon$ as non-redundant.
```

For large dimensional datasets, 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 Procedure 1 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 independently of the size of the dataset. For instance, we choose $t \leq 3$ and $m = 40$ for our numerical results. Therefore, we obtain a complexity linear in the size of the dataset.

As for interpretability, FOURIER-ORTH not only removes redundant features but also extracts the equations describing such redundancies through the matrix $\hat{\mathbf{A}}$. This property gives important information about the redundancy structure of the dataset.

### 4.2. Feature Selection Algorithm

In this part of the paper, we combine FOURIER-ORTH (Procedure 1 with the feature subset measure $M_n$ as in (7) and present our feature selection algorithm. We first perform FOURIER-ORTH to remove the redundant features and then apply $M_n$ on the subsets of the remaining $\tilde{d}$ features to select one with the highest score.

The measure $M_n$ captures the joint effect of the candidate feature subsets. However, to further reduce the running time, we use a fixed-depth search. 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 descending order based 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 1. Note that with $t = 1$, our search algorithm reduces to a feature-ranking method. On the other hand, with $t = k$, we get the exhaustive search over feature subsets of size $k$.

**Algorithm 1** 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 FOURIER-ORTH($t$) to get the non-trivial parities
  and non-redundant features.
Construct all $t$-element subsets $\mathcal{T}$ of the non-redundant
  features.
Rank all subsets $\mathcal{T}$ according to $M_n$ as in (7).
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 $\hat{\mathcal{J}}_n$
```

Therefore, the computational complexity of SFFS algorithm without FOURIER-ORTH and for a fixed parameter $t$ is $O(n\tilde{d}^t)$, where $\tilde{d}$ is the number of non-trivial features declared from the orthogonalization (Procedure 1). Our numerical results verifies that usually $\tilde{d}$ is much smaller than $d$, see Table 2. As a result, with FOURIER-ORTH the overall computational complexity of SFFS is $O(nd + n\tilde{d}^t)$, that is dominated by $O(nd)$ for large datasets.

## 5. Related Works

The literature in this area is extensive. Thus, we only can point out some of the best known and most relevant results.

The standard Fourier expansion on the Boolean cube has been central in a wide range of applications such as computational learning theory (Linial et al., 1993; Mossel et al., 2003; 2004; Blais et al., 2010; Heidari et al., 2019), noise sensitivity (O'Donnell, 2014; Kalai, 2005), information-theoretic problems (Courtade & Kumar, 2014; Heidari et al., 2021), and other applications (Aghazadeh et al., 2020). We note that there are other forms of orthogonal decomposition, including the Hoeffding-Sobel decomposition (Hoeffding, 1948; Sobol, 1993; Chastaing et al., 2012) and its generalization (Chastaing et al., 2012). 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.

Feature selection methods are usually classified into three main groups: wrappers, filter, and embedded (Guyon & Elisseeff, 2003; Yamada et al., 2020a). 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 (see (Yamada et al., 2020b) and references therein). However, such approaches are usually computationally expensive and, hence, prohibitive in large datasets. 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 has 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 (Vergara & Estévez, 2014; Koller & Sahami, 1996; Yu & Liu, 2004; Battiti, 1994; Peng et al., 2005), and Kernel-based measures (Gretton et al., 2005; Chen et al., 2017; Wei et al., 2016). 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 the nonlinear dependencies. However, the computational complexity of computing a kernel grows super linearly, if not quadratic, with the number of the samples (Cesa-Bianchi et al., 2015). Mutual Information (MI) criteria, on the other hand, can detect nonlinear dependencies with lower computational complexity (Battiti, 1994). In addition, mutual information can be used to bound the Bayes misclassification rate (Feder & Merhav, 1994; Cover & Thomas, 2006). However, estimating multi-variate mutual information is known to be a difficult task with high sample complexity.

In the unsupervised settings, some recent approaches worth mentioning are pseudo-label based and spectral/manifold based. Methods in the first approach attempt to generate pseudo-labels via clustering (Li et al., 2012; Yang et al., 2011). The second approach usually assumes linear dependencies among the feature (Feng et al., 2019; Arai et al., 2016; Derezinski et al., 2020).

# 6. Numerical Experiments

In this section, we present our numerical results and compare the performance of our SFFS algorithm with several well-known methods for supervised feature selection[1].

We test the algorithms on synthetic and real-world datasets as given in Table 1. The real-world datasets are benchmarks and taken from (Li et al., 2018) and the UCI repository (Dua & Graff, 2017). The synthetic datasets are described in the following.

**Synthetic datasets:** We generate two synthetic datasets (called E1 and E2) to test the ability of feature selection algorithms on capturing nonlinear feature-label dependencies. Each dataset 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 $(X_1, X_2, ..., X_6)$. We generate the labeling function randomly and independently of the algorithms. To control the level of the nonlinearity of this function, we generate it according to an Erlang distribution described in Section E of Supplementary Material.

## 6.1. Performance of FOURIER-ORTH Procedure

We start with the numerical results for the FOURIER-ORTH procedure. Table 2 shows the number of non-redundant features ($\tilde{d}$) declared by Procedure 1 and compares it with $d$, the original number of features in the datasets. This table confirms that $\tilde{d}$ is usually much smaller than $d$ for large datasets, implying that many of the features are redundant according to the Fourier expansion.

## 6.2. Comparisons of the Feature Selection Algorithms

Next, we present our comparison of the feature selection algorithms in terms of the classification accuracy. We use SFFS algorithm (with $t = 1, 2$) and compare it with the bench-marking algorithms such as ReliefF (Kira & Rendell, 1992), mRMR (Peng et al., 2005), MI (Kraskov et al., 2011), RFS (Nie et al., 2010), and CCM (Chen et al., 2017)[2].

Figure 1 shows the average classification accuracy for various numbers of selected features ($k$). The experiments employ $5$-fold cross-validation with feature selection and the support vector machine (SVM) classifier with radial basis function as a kernel. The implementation details are given in the Supplementary Material.

For real-world datasets, as Figure 2 shows, we obtain consistently good results in all the datasets and leading in some ranges of $k$. The compared algorithms perform well only in some of the datasets, while our algorithms have reliable, steady performance in all the cases. For instance, we observe a dominant performance by our SFFS in the

[1] The source codes are available at https://github.com/jithin-k-sreedharan/Fourier_feature_selection.

[2] We were unable to run CCM for USPS with its author's original implementation.

Table 1: Properties of the tested datasets.

| dataset | E1 | E2 | USPS | Isolet | COIL20 | Covertype | Australian | Musk | ALL AML |
|---|---|---|---|---|---|---|---|---|---|
| Features | 20 | 20 | 256 | 617 | 1024 | 54 | 14 | 166 | 7128 |
| Samples | 1000 | 1000 | 9298 | 1560 | 1440 | 581 | 690 | 467 | 72 |

Table 2: Number of non-trivial features ($\tilde{d}$).

| | E1 | E2 | USPS | Isolet | COIL20 | Covertype | Australian | Musk | ALL AML |
|---|---|---|---|---|---|---|---|---|---|
| $d$ | 20 | 20 | 256 | 617 | 1024 | 54 | 14 | 166 | 7128 |
| $\tilde{d}$ | 20 | 20 | 93 | 309 | 331 | 34 | 12 | 35 | 39 |
| $\tilde{d}/d$ | 1 | 1 | 0.36 | 309 | 0.50 | 0.63 | 0.86 | 0.21 | 0.005 |

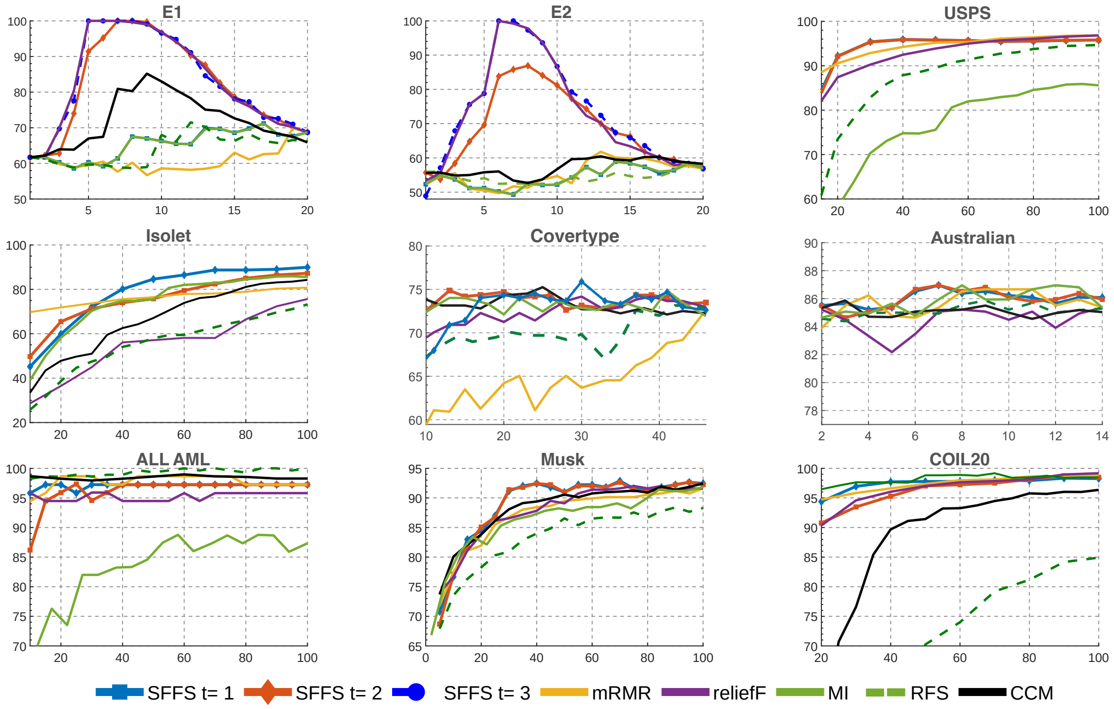

Figure 1: Classification accuracy (vertical axis) versus the number of selected features $k$ (horizontal axis). There are overlaps between SFFS($t = 3$) with ReliefF and SFFS($t = 1$) with MI for the E1 and E2 datasets. Also SFFS($t = 1$) overlaps with SFFS($t = 2$) for USPS and Musk datasets.

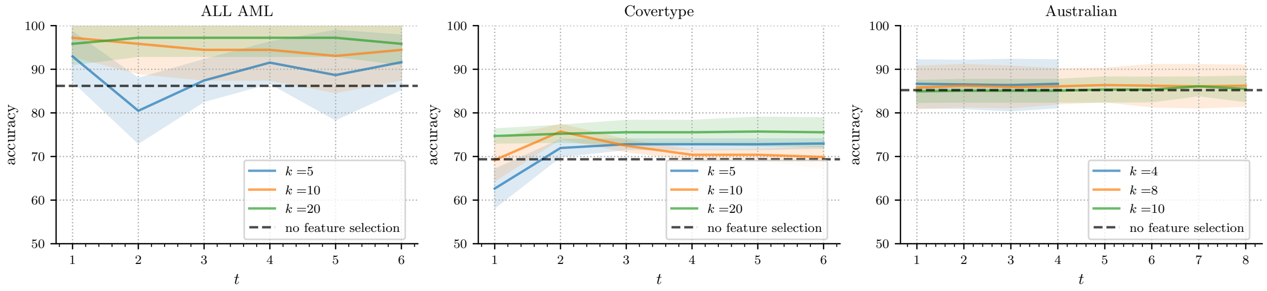

Figure 2: The effect of the depth parameter $t$ in the performance of SFFS algorithm. The figure shows classification accuracy (vertical axis) versus different values of $t$ for the ALL AML, Covertype, and Australian datasets.

Table 3: Comparison of running times for the tested feature selection algorithms. The numbers are in seconds.

| | 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 |

Isolet dataset for $k > 40$ and in the USPS dataset for $k < 50$. Moreover, in the Musk dataset, we observe a notable performance improvement for $k \in [25, 50]$. Note that SFFS($t = 1$) and SFFS($t = 2$) are overlapping in the USPS and Musk datasets for many values of $k$. Also, there are overlaps between SFFS($t = 3$) with ReliefF and and SFFS($t = 1$) with MI for the E1 and E2 datasets.

We also run SFFS with $t = 3$ on the synthetic datasets E1 and E2. As explained before, in E1 and E2, there are no redundant features, and there are only six relevant features. This is verified in Figure 1, where the maximum accuracy (100%) is obtained using SFFS at around $k = 6$. This observation also implies that SFFS detects all irrelevant features in these datasets. Further, we observe a significant performance gap between our approach and the other algorithms except for ReliefF. The low accuracy of other algorithms (below 60% in E2) suggests their failure to find the relevant features in these datasets. We believe this is due to the highly nonlinear feature-label relations in such datasets imposed by the Erlang distribution in our construction. In general, we do not see any reason for not having high nonlinearity in the real data sets. This observation calls for more sophisticated approaches in feature selection to address highly nonlinear relations.

### 6.3. Effect of the Depth Parameter $t$

Next, we analyze the effect of the depth parameter ($t$) on the algorithm's performance. For that, we fix $k$ and run the SFFS algorithm with different values of $t$ on the ALL AML, Covertype, and Australian datasets.

Figure 2 presents the resulted classification accuracy versus $t$ for various values of the number of selected features. It is observed that the performance of SFFS is relatively unchanged for large values of $t$. This observation suggests that low values of $t$ are sufficient to get a satisfactory performance for real datasets. However, it is expected that one could find/design a dataset for which larger values of $t$ are required to get a higher accuracy for SFFS.

We also note that in some cases as $t$ increases, the performance drops because the high value of $t$ demands more number of samples. The reason is that there are more Fourier coefficients to estimate.

### 6.4. Comparison of Running Times

Lastly, in Table 3, 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 (Li et al., 2018) and correspond to the original implementations, except for mRMR and CCM, where we used the optimized implementations from the authors.

## 7. Conclusion

In this work, we proposed a Fourier-based approach for feature selection. First, we presented a Gram-Schmidt orthogonalization using which we developed a Fourier expansion for functions of correlated binary random variables. We characterized the optimal feature subset and the minimum misclassification accuracy in the Fourier domain. Then, we proposed a measure for selecting subsets of features. The measure is an empirical estimate of the minimum misclassification probability in the Fourier domain. We proved that this measure finds asymptotically optimal feature subsets in binary settings. Further, we propose an algorithm (SFFS) for feature selection based on this measure and the orthogonalization. Lastly, we numerically analyzed the SFFS algorithm and showed performance improvements compared to several well-known feature selection algorithms.

## Acknowledgements

We thank all the reviewers for a thorough reading of our paper and many valuable comments. 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

Aghazadeh, A., Ocal, O., and Ramchandran, K. CRISPRL and: Interpretable large-scale inference of DNA repair landscape based on a spectral approach. *Bioinformatics*, 36(Supplement_1):i560–i568, jul 2020. doi: 10.1093/bioinformatics/btaa505.

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

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

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

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

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

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

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

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

Derezinski, M., Khanna, R., and Mahoney, M. W. Improved guarantees and a multiple-descent curve for column subset selection and the nystrom method. *Advances in Neural Information Processing Systems*, 33, 2020.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Feder, M. and Merhav, N. Relations between entropy and error probability. *IEEE Transactions on Information Theory*, 40(1):259–266, 1994. doi: 10.1109/18.272494.

Feng, C., Qian, C., and Tang, K. Unsupervised feature selection by pareto optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3534–3541, 2019.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. Measuring statistical dependence with hilbert-schmidt norms. In *Lecture Notes in Computer Science*, pp. 63–77. Springer Berlin Heidelberg, 2005. doi: 10.1007/11564089_7.

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

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

Heidari, M., Sreedharan, J. K., Shamir, G., and Szpankowski, W. Information sufficiency via fourier expansion. In *Proc. IEEE Int. Symp. Information Theory (ISIT)*, 2021.

Hoeffding, W. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. ISSN 00034851. URL http://www.jstor.org/stable/2235637.

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

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

Koller, D. and Sahami, M. Toward optimal feature selection. Technical report, Stanford InfoLab, 1996.

Kraskov, A., Stögbauer, H., and Grassberger, P. Erratum: Estimating mutual information [phys. rev. e 69, 066138 (2004)]. *Physical Review E*, 83(1):019903, 2011.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., and Liu, H. Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, 50(6):94, 2018.

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

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

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

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

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

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

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

Sobol, I. M. Sensitivity estimates for nonlinear mathematical models. *Mathematical modelling and computational experiments*, 1(4):407–414, 1993.

Solorio-Fernández, S., Carrasco-Ochoa, J. A., and Martínez-Trinidad, J. F. A review of unsupervised feature selection methods. *Artificial Intelligence Review*, 53(2):907–948, 2020.

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

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

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

Yamada, Y., Lindenbaum, O., Negahban, S., and Kluger, Y. Feature selection using stochastic gates. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10648–10659. PMLR, 13–18 Jul 2020a. URL http://proceedings.mlr.press/v119/yamada20a.html.

Yamada, Y., Lindenbaum, O., Negahban, S., and Kluger, Y. Feature selection using stochastic gates. In *International Conference on Machine Learning*, pp. 10648–10659. PMLR, 2020b.

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

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