Finding Relevant Information via a Discrete Fourier Expansion: Supplementary Material

Mohsen Heidari, Jithin K. Sreedharan, Gil I. Shamir, Wojciech Szpankowski

Contents

A. Proof of Proposition 1

Recall from the discussion in Section 2.1 that $\psi_{\mathcal{S}_i}$'s are orthonormal. 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 (O'Donnell, 2014), the function g can be written as

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

where $\mathrm{g}_{\mathcal{S}} = \mathbb{E}_{P_{X^d}}[f(X^d)\chi_{\mathcal{S}}(X^d)]$ and the expectation is taken with respect to P_{X^d}. By performing the reverse of the orthogonalization process in (4), each parity $\chi_{\mathcal{S}_i}, i = 1, 2, .., 2^d$, can be written as

$$\chi_{\mathcal{S}_i}(x^d) = \sum_{j\leq i} \alpha_{i,j}\psi_{\mathcal{S}_j}(\mathbf{x}), \tag{S.1}$$

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

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

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.

B. Proof of Lemma 1

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 $h(\mathbf{x}) \triangleq \mathsf{sign}[\mathbb{E}[Y|\mathbf{x}^{\mathcal{J}}]]$. Hence, it remains to show that $h = f^{\subseteq\mathcal{J}}$ as in the statement of the Lemma.

Note that h 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, h has a Fourier expansion of the form

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

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

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

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, h admits the same Fourier expansion as $f^{\subseteq \mathcal{J}}$. With that the proof is complete.

C. Proof of Theorem 1

Fix a subset $\mathcal{J} \subseteq [d]$ with at most k elements. Let $g : \{-1,1\}^d \mapsto \{-1,1\}$ be a function whose output depends on only $x^{\mathcal{J}}$. Here, g represents a predictor of Y from $X^{\mathcal{J}}$. Since Y and $g(\mathbf{X})$ take values from $\{-1,1\}$, then,

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

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

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

where (a) follows from the *law of total probability*, (b) holds because $g(\mathbf{X})$ is a function of $X^{\mathcal{J}}$, equality (c) follows by replacing g with $\mathsf{sign}\big[\mathbb{E}[Y|x^{\mathcal{J}}]\big]$, and lastly, (d) holds because $f^{\subseteq \mathcal{J}}(\mathbf{X}) = \mathbb{E}[Y|X^{\mathcal{J}}]$. This equality is shown in Lemma 1.

As a result of the above argument, 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 following and completes the proof

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

D. Proof of Theorem 2

From the proof of Theorem 1 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,

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

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

$$\begin{aligned}
\|f^{\subseteq \mathcal{J}^*}\|_1 - \|f^{\subseteq \hat{\mathcal{J}}_n}\|_1 &= \Big(\|f^{\subseteq \mathcal{J}^*}\|_1 - M_n(\mathcal{J}^*)\Big) + \Big(M_n(\mathcal{J}^*) - M_n(\hat{\mathcal{J}}_n)\Big) \\
&\quad + \Big(M_n(\hat{\mathcal{J}}_n) - \|f^{\subseteq \hat{\mathcal{J}}_n}\|_1\Big) \\
&\leq \Big(\|f^{\subseteq \mathcal{J}^*}\|_1 - M_n(\mathcal{J}^*)\Big) + \Big(M_n(\hat{\mathcal{J}}_n) - \|f^{\subseteq \hat{\mathcal{J}}_n}\|_1\Big),
\end{aligned} \tag{S.3}$$

where the last inequality follows as $M_n(\mathcal{J}^*) \leq M_n(\hat{\mathcal{J}}_n)$. Next, we provide upper bounds on the right-hand side of the above inequality. We first assume that there is no error in the estimation of the parities $\psi_{\mathcal{S}}$'s for all subsets with $|\mathcal{S}| \leq k$. 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 $|\mathcal{J}| \leq k$ and perform the orthogonalization process w.r.t $\mathcal{J}$. We proceed with the following lemma which is proved in Appendix D.1.

Lemma D.1 *The measure $M_n(\mathcal{J})$ as in (7) 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 $|\mathcal{J}| \leq k$,*

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

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 independent and identically distributed (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 D.1 and the triangle inequality, we have, with probability at least $(1-\delta)$, that

$$\big|M_n(\mathcal{J}) - \|f^{\subseteq \mathcal{J}}\|_1\big| \leq \sqrt{\frac{\lambda(k)}{(n-1)} \log(\frac{d}{\delta})} + O(n^{-\gamma}), \quad \forall \mathcal{J} \subseteq [d],\ |\mathcal{J}| = k. \tag{S.4}$$

The proof completes by combining (S.2), (S.3), and (S.4).

D.1 Proof of Lemma D.1

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 $|\mathcal{S}_j| \leq k$. We start with rewriting M_n. Define, the function

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

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

$$\begin{aligned}
\mathbb{E}[M_n(\mathcal{J})] &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{X^d(1),\ldots,X^d(n)}\Big[\Big|\ \hat{f}_{(i)}^{\subseteq \mathcal{J}}(X^d(i))\ \Big|\Big] \\
&= \mathbb{E}_{X^d(1),\ldots,X^d(n)}\Big[\Big|\ \hat{f}_{(1)}^{\subseteq \mathcal{J}}(X^d(1))\ \Big|\Big] \\
&= \mathbb{E}_{X^d(2),\ldots,X^d(n)}\mathbb{E}_{X^d(1)}\Big[\Big|\ \hat{f}_{(1)}^{\subseteq \mathcal{J}}(X^d(1))\ \Big|\Big] \\
&= \mathbb{E}_{X^d(2),\ldots,X^d(n)}\big[\|\hat{f}_{(1)}^{\subseteq \mathcal{J}}\|_1\big],
\end{aligned}$$

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

Observe that

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

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

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

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

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

Hence,

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

Putting all together we get that

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

Next, we address the effect of mean and variance estimations. As a measure of accuracy of the estimations, we require the following event:

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

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 D.2 *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.*

Lemma D.3 *Conditioned on* (B)*, the inequality $\|\chi_{\mathcal{S}} - \widehat{\phi}_{\mathcal{S}}\|_\infty \leq \gamma(\epsilon_0)$ 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 $\epsilon_0 \to 0$.*

E. 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 $p(\mathbf{x}) = \sum_{i,j} \alpha_{i,j} \chi_{\mathcal{S}_{i,j}}(\mathbf{x})$. Note that by changing the parameters of the Erlang pdf, we get different randomized polynomials.

F. Implementation Details

In this section, we explain the details of our implementations of SFFS algorithm.

The following are some of the characteristics of our implementation:

- For benchmarking purposes, we use the original implementation of mRMR[1], scikit-feature[2] for MCFS, and ReliefF, and scikit-learn[3] 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 Procedure 1 and Fourier coefficient calculation in Algorithm 1) are converted to C++ using Cython.
- We have also parallelized some of the computations.
- All the experiments were performed on 48-CPU workstation, with Intel(R) Xeon(R) CPU E7-8857 v2 @ 3.00GHz and 256GB RAM.

FOURIER-ORTH with limited computational resources: To minimize the computational burden further, we follow a sequential approach for the FOURIER-ORTH procedure. 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 FOURIER-ORTH with $t = 1$. Let its count be d_1. If the actual number of features $d < a_1$, we directly run FOURIER-ORTH with $t = 2$ on the full set of features. Otherwise, if $d_1 < a_1$, FOURIER-ORTH ($t = 2$) is ran 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 FOURIER-ORTH ($t = 2$) 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 FOURIER-ORTH ($t = 2$): a) if $d < a_2$, run the FOURIER-ORTH directly; b) else if $d_2 < a_2$, run FOURIER-ORTH ($t = 3$) on d_2 features; c) in

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

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 hyper-parameters that needs to be chosen depending on the computational resources.

References

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