Finding Relevant Information via a Discrete Fourier Expansion

Mohsen Heidari

PURDUE UNIVERSITY

Jithin Sreedharan

WADHWANI AI
AI FOR SOCIAL GOOD

Gil Shamir

Google

Wojciech Szpankowski

PURDUE UNIVERSITY

ICML 2021

Motivation

- High-dimensional datasets with many “redundant” or “irrelevant” features.
- Linear relations are easy to identify
  - Linearly redundant features

$$X_i = a_0 + a_1 X_{j_1} + a_2 X_{j_2}, + \cdots + a_k X_{j_k}$$

- Non-linear structures:
  - $X_i = g\left(X_{j_1}, X_{j_2}, \ldots, X_{j_k}\right)$
- How to capture multi-variate and non-linear relations?

  - Kernel-based approach:
    - Computationally expensive.
  - Information-theoretic measures:
    - High sample complexity.

This work:

- ✓ Fourie-based approach
- ✓ Feature Selection

Key Ideas

Standard Fourier on the Boolean Cube:

Any bounded $g:\{+1,-1\}^d \to \mathbb{R}$ is uniquely written as:

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

Fourier Coefficients:

$$g_{\mathcal{S}} = \frac{1}{2^d}\sum_{x} g(\boldsymbol{x})\, \chi_{\mathcal{S}}(\boldsymbol{x})$$

Monomials:

$$\chi_{\mathcal{S}} = \prod_{j\in\mathcal{S}} x_j$$

Restrictions:

- Uniform or product probability distributions.
- →Independent features.

This Work: Correlated Fourier Expansion

- Arbitrary distribution $(\boldsymbol{X}^d, Y) \sim D$

$$g(\boldsymbol{x}) = \sum_{\mathcal{S}\in\mathcal{T}} g_{\mathcal{S}}\, \psi_{\mathcal{S}}(\boldsymbol{x})$$

non-redundant features

Orthogonalized Parities

$$g_{\mathcal{S}} = \mathbb{E}_D[g(\boldsymbol{X})\psi_{\mathcal{S}}(\boldsymbol{X})]$$

Unsupervised FS
Orthogonalization
Algorithm

Supervised FS
Estimate $\hat{g}_{\mathcal{S}}$
SFFS Algorithm

Orthogonalization

Gram-Schmidt-Type Orthogonalization:

- Ordering the subsets of $\{1,2,\dots,d\}$

  $\emptyset, \{1\}, \{2\}, \{1,2\}, \{3\}, \{1,3\}, \{2,3\}, \{1,2,3\}, \dots, \{1,2,\dots,d\}$

- Orthogonalization w.r.t D

$$\tilde{\psi}_{S_i} \equiv \chi_{S_i} - \sum_{j=1}^{i-1} \langle \psi_{S_j}, \chi_{S_i} \rangle_D \, \psi_{S_j},$$

$$\psi_{S_i} \equiv \begin{cases} \frac{\tilde{\psi}_{S_i}}{\|\tilde{\psi}_{S_i}\|_{2,D}} & \text{if } \|\tilde{\psi}_{S_i}\|_{2,D} > 0 \\ 0 & \text{otherwise.} \end{cases}$$

- Nonlinear redundancy measure:

$$\|\tilde{\psi}_{\{j\}}\|_2 \leq \epsilon$$

Implementation:

Step 1: Fixed-depth Search

Only feature subsets of size at most t (say $t = 2$):

$\emptyset, \{1\}, \{2\}, \{1,2\}, \{3\}, \{1,3\}, \{2,3\}, \quad , \{4\}, \{1,4\}, \dots$

Step 2: Empirical Orthogonalization

- Matrix of empirical correlation coeffects:

$$B = \left[\frac{\mathrm{cov}(X_i, X_j)}{\sigma_i \sigma_j}\right]_{i,j \in [d]}$$

- Recursive formula:

$$\|\tilde{\psi}_{\{i\}}\|_2^2 \approx b_{i,i} + \sum_{j<i} a_{j,i}^2$$

$$a_{j,i} = \frac{1}{\sqrt{b_{j,j} - \sum_{r<j} a_{r,j}^2}} \left(b_{j,i} - \sum_{\ell<j} a_{\ell,j} a_{\ell,i} \right)$$

Representation in the Fourier Domain

- Binary classification with 0-1 loss:

Theorem (Fourier Characterization)

- The Bayes predictor of Y from a feature subset $\mathcal{J}$ is $\hat{y} = \text{sign}[f^{\subseteq \mathcal{J}}(\boldsymbol{x})]$, where

$$f^{\subseteq \mathcal{J}}(\boldsymbol{x}) = \sum_{\mathcal{S} \subseteq \mathcal{J}} \alpha_{\mathcal{S}}\, \psi_{\mathcal{S}}(\boldsymbol{x}), \qquad \alpha_{\mathcal{S}} = \mathbb{E}_D[Y \psi_{\mathcal{S}}(X)]$$

Fourier for Stochastic label Y
$(\boldsymbol{X}^d, Y) \sim D$,

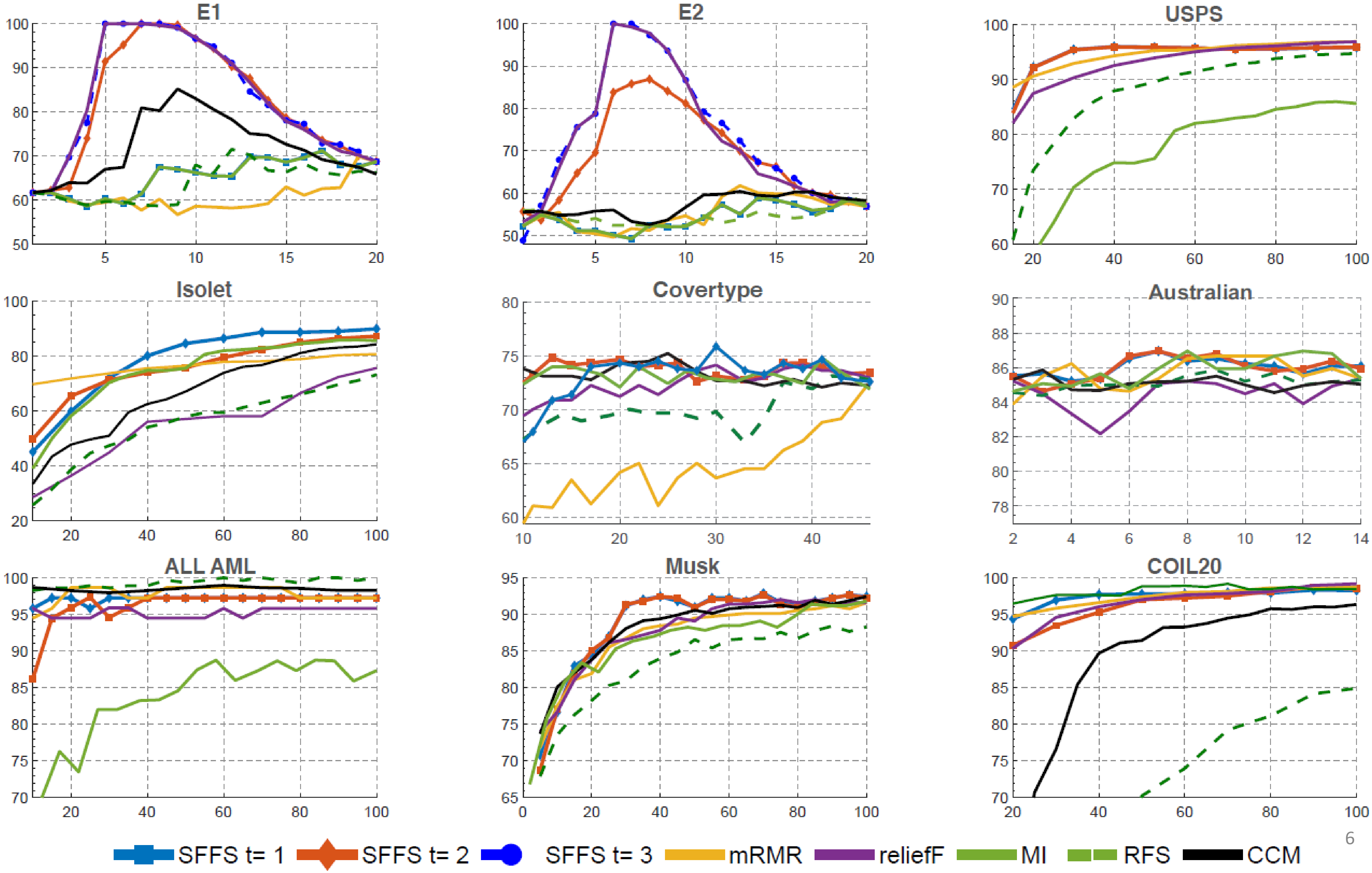
- Minimum loss when selecting k features:

$$L_{\text{opt}}(k) = \frac{1}{2} - \frac{1}{2} \max_{\mathcal{J}:|\mathcal{J}|=k} \left\| f^{\subseteq \mathcal{J}} \right\|_{1,D}$$

- Optimal feature subset:

$$\mathcal{J}^* = \underset{\mathcal{J}:|\mathcal{J}|=k}{argmax} \left\| f^{\subseteq \mathcal{J}} \right\|_{1,D}$$

measure for nonlinear *relevancy*
➔Estimation from samples.


E1
E2
USPS
Isolet
Covertype
Australian
ALL AML
Musk
COIL20
SFFS t= 1
SFFS t= 2
SFFS t= 3
mRMR
reliefF
MI
RFS
CCM

Measure for Feature Selection

- Empirical Fourier Expansion to estimate $\left\| f^{\subseteq \mathcal{J}} \right\|_{1,D}$
  - Orthogonalization Process → $\hat{\psi}_{\mathcal{S}}(\boldsymbol{x})$
  - Fourier Coefficients: $\hat{\alpha}_{\mathcal{S}} = \frac{1}{n} \sum_i y(i) \hat{\psi}_{\mathcal{S}}\big(\boldsymbol{x}(\boldsymbol{i})\big)$
- Relevancy Measure:

$$M_n(\mathcal{J}) = \frac{1}{n-1} \sum_{i=1}^{n} \left| \sum_{\mathcal{S} \subseteq \mathcal{J}} \hat{\alpha}_{\mathcal{S}}\, \hat{\psi}_{\mathcal{S}}(\boldsymbol{x}_i) - \frac{1}{n} y_i \left(\hat{\psi}_{\mathcal{S}}(\boldsymbol{x}_i) \right)^2 \right|$$

Theorem (Consistency of the measure)

If $\hat{\mathcal{J}}_n = \mathrm{argmin}_{\mathcal{J}} \mathrm{M_n}(\mathcal{J})$, then

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

with probability $(1-\delta)$, *where* $\lambda(k) = O\big(k 2^{2k}\big), \gamma \approx 1/2.$

Numerical Experiments:

Table I: Orthogonalization Output.

	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

Table II: Running Times (in sec).

	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