Learning Global Probabilistic Explanations: A Spectral Approach with Black-Box Queries

Frédéric Koriche [1] **Louenas Bounia** [2]

[1] CRIL, UMR CNRS 8188 – Université d'Artois
[2] LIPN, UMR CNRS 7030 – Université Sorbonne Paris Nord

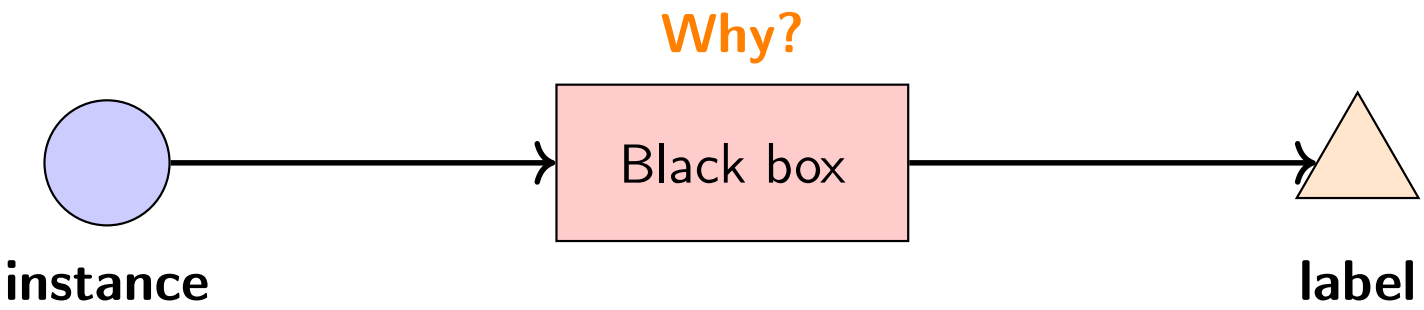
NeurIPS 2026







Why Explainable AI?



- High-performing models are **black boxes**, yet they decide in **critical applications** (credit, medicine, justice).
- Users, auditors and regulators ask **why**: which **features** does the model really use?
- Exact *sufficient reasons* are often too large: humans handle only **4 or 5** interacting variables [Halford et al. 1998].
- **Probabilistic explanations**: fix a size limit k, and find the **most relevant** subset of at most k features.

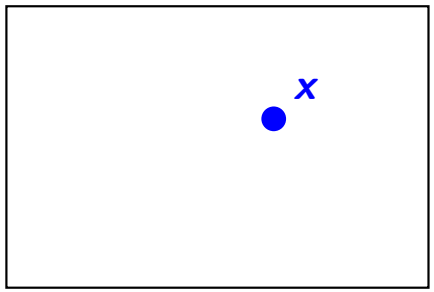
Local versus Global Explanations

Local explanation

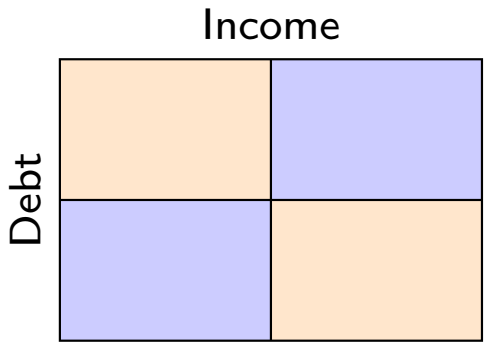
- Explains **one** prediction $f(\boldsymbol{x})$.
- "For this applicant, `Income` is enough."

Global explanation

- Explains the model on **all** inputs.
- "`Income` and `Debt` split all applicants into groups with the same decision."



local: one point



global: a partition by K

- The subset K acts as a **junta**: f should be (almost) constant on each region.

The Problem

Classifier $f : \mathbf{\Omega} \to \{-1, +1\}$, with $\mathbf{\Omega} = \Omega_1 \times \cdots \times \Omega_d$ and $|\Omega_i| \leq c$. Product distribution $\mathcal{D}$ with full support, and **black-box** access to f (membership queries).

Local relevance error of $K \subseteq [d]$ for an instance $\boldsymbol{x}$

$$\mathrm{err}_{f,\mathcal{D}}(K, \boldsymbol{x}) = \mathbb{P}_{\boldsymbol{z} \sim \mathcal{D}(\cdot | \boldsymbol{z}_K = \boldsymbol{x}_K)} \left[f(\boldsymbol{x}) \neq f(\boldsymbol{z}) \right].$$

Global relevance error: the average of the local error over $\boldsymbol{x} \sim \mathcal{D}$

$$\mathrm{err}_{f,\mathcal{D}}(K) = \mathbb{P}_{\boldsymbol{x} \sim \mathcal{D},\ \boldsymbol{z} \sim \mathcal{D}(\cdot | \boldsymbol{z}_K = \boldsymbol{x}_K)} \left[f(\boldsymbol{x}) \neq f(\boldsymbol{z}) \right].$$

Goal: find K with $|K| \leq k$ minimizing $\mathrm{err}_{f,\mathcal{D}}(K)$.

Complexity: Local versus Global

Local (full support)

- NP$^{\text{PP}}$-hard for neural networks.
- NP-hard for decision trees.
- W[2]-hard in k.
- Not approximable within $d^{1-\alpha}$.

Global (this work)

- **PAC**: ϵ-optimal with probability $1-\delta$.
- **FPT** in k, c and $1/\epsilon$.
- Any classifier, with **membership queries only**.

- **Previous global work** [Bassan et al. 2026]: greedy selection on an *empirical* product distribution, which covers a tiny part of the space in high dimension.
- Our main tool: **Fourier analysis on product spaces**.

Fourier Analysis on Product Spaces

- For each feature i, an orthonormal basis of $\mathcal{L}^2(\Omega_i, \mathcal{D}_i)$, computed by the Stieltjes procedure in $O(c^2)$:

$$\{\phi_{i,0} = 1, \phi_{i,1}, \ldots, \phi_{i,c_i-1}\}.$$

- Their tensor product gives a unique expansion of any f:

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{s}} \hat{f}(\boldsymbol{s})\, \phi_{\boldsymbol{s}}(\boldsymbol{x}), \qquad \phi_{\boldsymbol{s}}(\boldsymbol{x}) = \prod_{i=1}^{d} \phi_{i,s_i}(x_i).$$

- **Projection on K**: keep only the coefficients supported on K,

$$f_K(\boldsymbol{x}) = \mathbb{E}\left[f(\boldsymbol{z}) \mid \boldsymbol{z}_K = \boldsymbol{x}_K\right] = \sum_{\mathrm{supp}(\boldsymbol{s}) \subseteq K} \hat{f}(\boldsymbol{s})\, \phi_{\boldsymbol{s}}(\boldsymbol{x}).$$

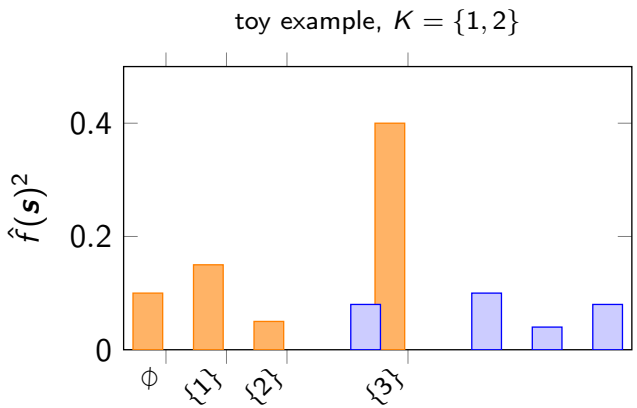
Spectral Characterization

Lemma

For any classifier f, any full-support product distribution $\mathcal{D}$ and any K:

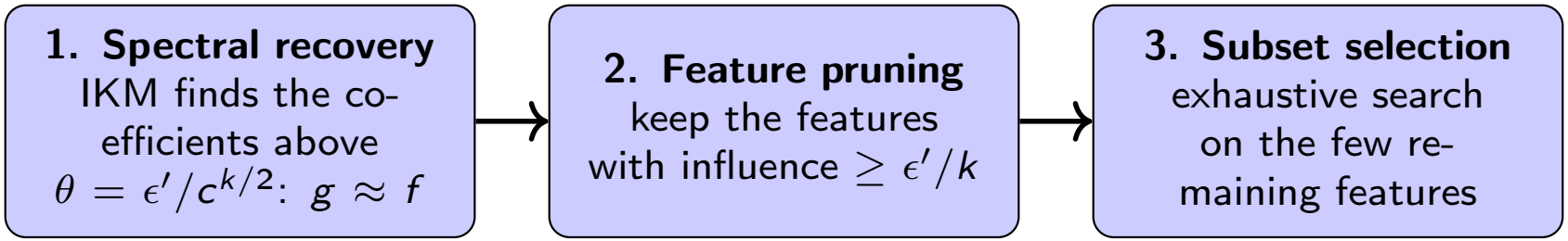
$$\mathrm{err}_{f,\mathcal{D}}(K) = \frac{1}{2}\left(1 - W(f_K)\right), \qquad W(f_K) = \sum_{\mathrm{supp}(\boldsymbol{s}) \subseteq K} \hat{f}(\boldsymbol{s})^2.$$

- The best explanation is the K that **captures the most spectral energy**.
- $W(f_K)$ does not depend on the chosen bases.
- An interaction shows up on a **joint** coefficient, which no single feature reveals.



Orange: energy captured by K, here $W(f_K) = 0.7$, so $\mathrm{err}_{f,\mathcal{D}}(K) = 0.15$.

The Spectral Explainer (SE)



- **IKM** (Integer Kushilevitz–Mansour) extends KM to product distributions over discrete domains; it explores a prefix tree of coefficients and keeps only the heavy branches.
- Pairs of inputs sharing the same suffix require **membership queries**.
- **Median-of-means** estimates: basis functions can take large values, but variances stay ≤ 1.
- After pruning, at most $63k^2/(8\epsilon)$ features remain: the search **does not depend on** d.

Main Result

Theorem

For any classifier f, any full-support product distribution $\mathcal{D}$, any $k \leq d$ and $\epsilon, \delta \in (0,1)$, SE returns K with $|K| \leq k$ such that, with probability at least $1-\delta$,

$$\mathrm{err}_{f,\mathcal{D}}(K) \leq \min_{|K'| \leq k} \mathrm{err}_{f,\mathcal{D}}(K') + \epsilon,$$

in time polynomial in d, $(ck/\epsilon)^k$ and $\ln(1/\delta)$.

Sketch of proof ($\epsilon' = 2\epsilon/7$, K^* optimal):

- Recovery: passing from f to g moves $W(\cdot_{K^*})$ and $W(\cdot_K)$ by at most $3\epsilon'$ each.
- Pruning: restricting K^* to the influential features costs at most ϵ'.
- Selection: K maximizes $W(g_{K'})$ on the pruned set, so there is no loss.
- Total: $7\epsilon' = 2\epsilon$ on the energy, hence ϵ on the error.

Experiments: Experimental Protocol

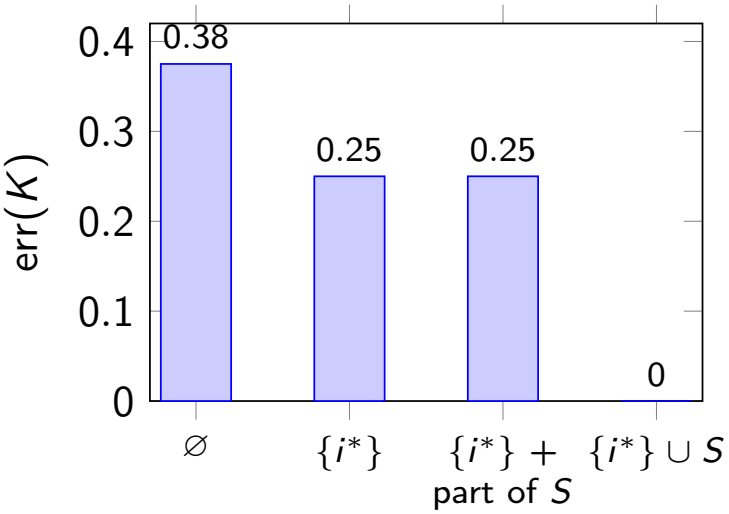
- **SE**: $\epsilon = 0.35$, $\delta = 0.1$, with **100 times fewer samples** than the theory requires.
- **GE**: greedy forward selection inspired by Bassan et al., with 5 000 fresh instances per candidate.
- **Evaluation**: exact $\mathrm{err}_{f,\mathcal{D}}(K)$ by enumerating $\mathbf{\Omega}$, and opt by brute force; a run succeeds if $\mathrm{err}_{f,\mathcal{D}}(K) \leq \mathrm{opt} + \epsilon$.
- **Synthetic targets** ($d = 16$, uniform $\mathcal{D}$): hidden parity, noisy parity and Synergy Trap.
- **Real-world benchmarks**: 9 OpenML datasets, d from 8 to 20, binarized features; f is an MLP with 100 ReLU units, and $\mathcal{D}$ is the product of the empirical marginals.

Results: the Synergy Trap

A model f on $d = 16$ features:

- if $x_{i^*} = 1$, then $f(\boldsymbol{x}) = 1$;
- otherwise, $f(\boldsymbol{x})$ is the parity of a hidden set S.

Alone, i^* helps; alone, no feature of S helps; together, $\{i^*\} \cup S$ explains everything.



- **GE** takes i^* first, then stops at error $1/4$ in 28 of the 30 runs with $k \geq 3$.
- **SE** returns $\{i^*\} \cup S$ in **40 / 40** runs, with error 0.

Interactions are invisible one feature at a time, but not to the spectrum.

Results: Noisy Parity and Cost

- **Noisy parity** ($d = 16$, $|S| = 3$, 5% of labels flipped, $\mathrm{opt} \approx 0.093$):
  - the gap of **SE** to the optimum is **0** for all $\epsilon \in [0.20, 0.45]$;
  - the gap of **GE** is about 0.4.
- **Confidence**: **0 failures in 600 runs** of SE, for δ from 0.01 to 0.5; GE succeeds in only 2% of these runs.
- **Cost**: the number of queries of SE is exactly $(2d+1)\, b\, n$, as predicted:
  - linear in d up to a logarithmic factor;
  - growing as c^{2k} in the parameters k and c.

Results: Real-World Benchmarks

		opt reached		$\mathrm{err}_{f,\mathcal{D}}(K)$, $k = 5$		
Dataset	d	SE	GE	SE	GE	opt
Breast Cancer	8	**30/30**	25/30	**0.080**	**0.080**	0.080
Tic-Tac-Toe	9	**30/30**	21/30	**0.249**	0.254	0.249
COMPAS	13	**30/30**	26/30	**0.110**	0.112	0.110
Mushroom	20	**24/30**	17/30	**0.183**	0.185	0.183
Credit-G	20	**24/30**	21/30	**0.217**	0.221	0.217

Best value in bold; GE values are means over 6 runs.

- On the 9 datasets, SE reaches the optimum in **246 / 270** runs, against 211 for GE.
- SE returns the same explanation in every run for **45 / 45** pairs (dataset, k), against 24 for GE.
- The price: one run of SE takes 4 to 41 minutes at $k = 5$, and GE less than a second.

Examples of Explanations

Explanations returned by SE for $k = 3$, all optimal and identical across runs:

Dataset	Explanation K	$\text{err}_{f,\mathcal{D}}(K)$
Tic-Tac-Toe	top-left, middle-middle, bottom-left	0.326
COMPAS	age > 31, priors count > 2, age category < 25	0.138
Credit-G	no checking account, duration > 18, savings < 100	0.247
Breast Cancer	clump thickness > 4, cell shape unif. > 1, bland chromatin > 3	0.139

- **Tic-Tac-Toe**: the center and two corners, which lie on the largest number of winning lines. GE always takes the center first, but picks the wrong corner in 3 of 6 runs for $k = 2$.
- A global explanation is a **set of features** (8 regions here), not a conjunction of literals.

Conclusion

- **Take-home message**
  - The global relevance of K is the **Fourier energy** of f on K.
  - The Spectral Explainer is **PAC** and **FPT**, using only membership queries.
  - It finds feature interactions missed by greedy selection.
- **Limits**
  - The number of samples grows as c^{2k}/ϵ^4: several hours for $d \approx 60$.
  - Product distributions only.
- **Future work**
  - Tighter, adaptive spectral estimates: 300 times fewer samples sufficed in calibration.
  - Beyond product distributions: Bayesian networks and probabilistic circuits.

Thank you!

Questions?

bounia@lipn.univ-paris13.fr

Code: https://github.com/Lounesbo/neurips-2026

Backup: IKM in Detail

For a prefix $\boldsymbol{\sigma}$ of length p, the weight of its cone is

$$W_{\boldsymbol{\sigma}} = \mathbb{E}_{\boldsymbol{u},\boldsymbol{u}',\boldsymbol{v}} \left[f(\boldsymbol{u},\boldsymbol{v})\, \phi_{\boldsymbol{\sigma}}(\boldsymbol{u})\, f(\boldsymbol{u}',\boldsymbol{v})\, \phi_{\boldsymbol{\sigma}}(\boldsymbol{u}') \right].$$

- Keep $\boldsymbol{\sigma}$ if $\hat{W}_{\boldsymbol{\sigma}} \geq \theta^2/2$, then extend it with all values of the next index.
- At most $4/\theta^2$ prefixes are kept at each depth, **whatever** c (Parseval).

Theorem (IKM)

With probability at least $1-\delta$, IKM returns g with $L_\infty(f-g) \leq \theta$ and $L_0(g) \leq 4/\theta^2$, using $O\big(\frac{dc}{\theta^{\mathbf{6}}} \ln \frac{dc}{\theta\delta}\big)$ queries, and exactly $(2d+1)\, b\, n$ with shared batches.

Backup: Cost of SE

Growth of the number of queries $Q = (2d + 1)\, b\, n$ on hidden parities ($\epsilon = 0.35$, $\delta = 0.1$):

Sweep	$2d + 1$	$n \propto c^{2k}$	b	Q observed
d: $8 \to 128$	$\times 15.1$	$\times 1$	$\times 1.20$	$\times 18.1$
$k \to k + 1$	$\times 1$	$\times 4$	$\times 1.04$ to 1.06	$\times 4.2$
c: $2 \to 6$	$\times 1$	$\times 81$	$\times 1.24$	$\times 101$

- Linear in d up to a logarithmic factor, and exponential only in the parameters k and c.